Computational Methods for Learning Population History from Large Scale Genetic Variation Datasets

Ming-Chi Tsai

CMU-CB-13-102

July 2, 2013

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Russell Schwartz (Department of Biological Science, Chair)
Guy Blelloch (Department of Computer Science)
R. Ravi (Tepper School of Business)
Eleanor Feingold (Department of Human Genetics)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy.

Keywords: Population Genetics, Minimum Description Length, Population History, Markov chain Monte Carlos, Coalescent Theory, Genome Wide Association Study

Abstract

Understanding how species have arisen, dispersed, and intermixed over time is a fundamental question in population genetics with numerous implications for basic and applied research. It is also only by studying the diversity in human and different species that we can understand what makes us different and what differentiates us from other species. More importantly, such analysis could give us insights into applied biomedical questions such as why some people are at a greater risk for diseases and why people respond differently to pharmaceutical treatments. While there are a number of methods available for the analysis of population history, most state-of-the-art algorithms only look at certain aspects of the whole population history. For example, phylogenetic approaches typically look only at non-admixed data in a small region of a chromosome while other alternatives examine only specific details of admixture events or their influence on the genome.

We first describe a basic model of learning population history under the assumption that there was no mixing of individuals from different populations. The work presents the first model that jointly identifies population substructures and the relationships between the substructures directly from genetic variation data. The model presents a novel approach to learning population trees from large genetic datasets that collectively converts the data into a set of small phylogenetic trees and learns the robust population features across the tree set to identify the population history.

We further develop a method to accurately infer quantitative parameters, such as the precise times of the evolutionary events of a population history from genetic data. We first propose a basic coalescent-based MCMC model specifically for learning time and admixture parameters from two-parental and one-admixed population scenarios. As a natural extension, we then expanded that method to identify population substructures and learn population models and the specific time and admixture parameters pertaining to the population history for three or more populations. Analysis on simulated and real data shows the effectiveness of the approach in working toward unifying the learning of different aspect of population history into single algorithm.

Finally, as a proof of concept, we propose a novel structured test statistic using the historic information learned from our prior method to improve demographic control in association testing. The success of the structured association test demonstrates the practical value of population histories learned from genetic data for applied biomedical research.

Acknowledgments

Pursuing a Ph.D. is a challenging and difficult journey one may face in their life. In my pursuit of this goal, I would like to thank, first and foremost, my advisor, Dr. Russell Schwartz, for introducing me to the field of phylogenetics and population genetics. It is his insightful vision, continuous encouragement, and great patience that have made this dissertation possible. Russell not only put an enormous amount of time and effort into directing this research but he also helped tremendously in editing the documents and confirming every aspect of research results. I could not thank him enough for teaching me how to be a researcher, starting from understanding the problem to formulating a hypothesis to determining appropriate methodologies for solving the problem. I will always be indebted to Russell for encouraging me to participate and present our research papers in national and international conferences as well as reviewing and critiquing conference papers.

I would like thank Dr. Guy Blelloch and Dr. R. Ravi for their guidance and encouragements throughout the years at Carnegie Mellon especially their introduction of information theory to me that served as the basis of the algorithms.

I would also like to thank Dr. Eleanor Feingold for serving on my thesis committee, and for the helpful suggestions, ideas, and comments on my thesis.

I would like to personally thank Dr. Srinath Sridhar for his support during my first year of graduate study that helped me jump start on my research. I would also like to thank Dr. Haijun Gong, Dr. Navodit Misra, Ayshwarya Subramanian, and Salim Chowdhury for their helpful discussions about my project.

Most importantly, I would like to thank my families: my mother, my father, and my sisters for their support and encouragement throughout my life and the course of my Ph.D. I would like to especially thank my beloved wife, Tina. Without her love, support, and sacrifices, I would not be here and achieve what I accomplished today. Thank you and I love you and the little one from the bottom of my heart.

Contents

1 Introduction 1

1.1 Genetic Variations . . . 2

1.2 Genetic Variation Datasets . . . 5

1.3 Inference of Population History . . . 6

1.3.1 Population Substructure . . . 6

1.3.2 Phylogenetic Analysis for Ancestry Inference . . . 7

1.4 Ancestry Inference in the Presence of Admixture . . . 10

1.5 Limitations of Existing Approaches for Learning Population History . . . 13

1.6 Contributions . . . 14

1.7 Thesis Organization . . . 15

2 Learning Population Histories From Large-Scale Datasets in the Absence of Admixture 17

2.1 Methods . . . 19

2.1.1 Consensus Tree Model . . . 19

2.1.2 Algorithms . . . 23

2.2 Validation Experiments . . . 26

2.2.1 Simulated Dataset . . . 26

2.2.2 Real Data . . . 27

2.2.3 Benchmarks . . . 27

2.2.4 Sensitivity Test . . . 30

2.3 Results . . . 31

2.4 Discussion . . . 34

2.5 Conclusion . . . 37

3 Coalescent-based Method for Learning Parameters of Admixture Events from Large-Scale Genetic Variation Datasets 45

3.1 Materials and Methods . . . 50

3.2 Validation Experiments . . . 58

3.3 Results . . . 61

3.4 Discussion . . . 69

4 Coalescent-based Method for Joint Estimation of Population History, Time, and Admixture 75

4.1 Material and Methods . . . 78

4.1.1 Learning Summary Descriptions . . . 79

4.1.2 Admixture Model . . . 80

4.1.3 MCMC Sampling . . . 83

4.1.4 Validation on Simulated Data . . . 84

4.1.5 Validation on Real Data . . . 89

4.2 Results . . . 91

4.2.1 Simulated Data . . . 91

4.2.2 Real Data . . . 94

4.3 Discussion . . . 97

5 Applying Structure Information to Association Mapping 111

5.1 Methods . . . 113

5.1.1 Localized Structured Association Test . . . 114

5.1.2 Weighted Localized Structured Association Test . . . 116

5.1.3 Validation . . . 118

5.2 Results . . . 120

5.3 Discussion . . . 122

6 Conclusions and Future Directions 125

6.1 Future Work . . . 127

Bibliography 131

List of Figures

1.1 The types of genetic variations 3

2.1 Example of a tree converted to bipartitions 21
2.2 Illustration of the DMST construction for determining model description length . 24
2.3 Inferred consensus tree from simulated and real world data 39
2.4 Inferred population structure from simulated data 40
2.5 Variation of information and inconsistency scores for trees learned from real datasets 41
2.6 Average runtime of the consensus tree algorithm 42
2.7 Consensus trees produced using varying numbers of input trees 43

3.1 Example of a history of two parental populations and an admixed population . . . 49
3.2 Mean estimated parameters on 3.5×10^6-base simulated sequences 62
3.3 The average absolute difference between the estimated and true parameter values 63
3.4 Estimated t_1/t_2 ratio against true t_1/t_2 64
3.5 Parameters probability density for bovine, Hapmap, and 1000 Genome 69

4.1 Example of a history of two parental populations and an admixed population . . . 80
4.2 Possible evolutionary models for 3 populations 85
4.3 Population assignment for 1000 Genome Project dataset 95
4.4 Top four evolutionary models estimated from 1000 Genome Project 96
4.5 Population assignment for HapMap Phase 3 Release 2 dataset 97
4.6 Top four evolutionary models estimated from HapMap Phase 3 dataset 98

5.1 Example of a Cochran-Mantel-Haenszel test applied using separation of chromsomes into two partition sets using the a model partition 115
5.2 Negative log p-values computed for all 1866 variant sites generated from coalescent simulator, MS, and 120 simulated loci 121
5.3 Negative log p-values computed for all 45487 variant sites obtained from HapMap phase 2 dataset and 120 simulated candidate loci 123

List of Tables

1.1 **List of Some Important Large-Scale Genetic Variation Datasets** 6

3.1 Three quartiles for 135 simulated data sets 70

4.1 Estimated evolutionary model distribution for $\alpha = 0.05$ on 3.5×10^6-base sequences 99
4.2 Estimated evolutionary model distribution for $\alpha = 0.2$ on 3.5×10^6-base sequences 100
4.3 Estimated evolutionary model distribution for $\alpha = 0.6$ on 3.5×10^6-base sequences 101
4.4 Estimated evolutionary model distribution for $\alpha = 0.05$ on 3.5×10^7-base sequences 102
4.5 Estimated evolutionary model distribution for $\alpha = 0.2$ on 3.5×10^7-base sequences 103
4.6 Estimated evolutionary model distribution for $\alpha = 0.6$ on 3.5×10^7-base sequences 104
4.7 Average parameter estimation error from three-population simulated datasets . . 105
4.8 Estimated evolutionary model distribution for none-admixed four-population scenarios 105
4.9 Estimated evolutionary model distribution for one admixture, one divergence four-population scenarios 106
4.10 Estimated evolutionary model distribution for one divergence, one admixture four-population scenarios 107
4.11 Estimated evolutionary model distribution for two-admixture four-population scenarios 108
4.12 Average parameter estimation error for four-population scenarios 109

Chapter 1

Introduction

For centuries, understanding how species have arisen, dispersed, and intermixed over time has been one of the most sought-after questions man has tried to address. Since the publication of *On the Origin of Species* in 1859, tremendous efforts have been made to characterize the relationship and significance of the diversity between and within species, as the problem seems to possess an irresistible aesthetic appeal to mankind. It is also only by studying the diversity in humans and other species that we can understand what makes us different and what differentiates us from other species. More importantly, such analysis could give us insights as to why some people are at a greater risk for diseases and why people respond differently to pharmaceutical treatments.

Before the discovery of genetic material, works on the inference of the phylogenetic relationships between organisms largely relied on morphological, physiological, and phenotypic differences [70]. By quantifying the similarity and dissimilarity between different organisms, one can infer the relationships among organisms. Analyses based on morphological, physiological, and phenotypic differences have worked particularly well for quantifying the relationships between species that long ago diverged and evolved into remotely related species with distinct physical features, but are limited in close-species or within-species differentiations where physical appearances may be highly similar. Advances in ancestry inference did not significantly progress until the development of tools for detecting genetic variations [97, 104, 107]. The large amount of genetic differences between organisms provided sufficient resolution to infer precise and accurate relations between closely related species. Since then, a large number of studies utilizing genetic data have been published [28, 36, 44, 79, 107, 129]. However, close-species and within-species analyses of genetic variations were not fully realized due to the difficulties in obtaining large quantities of genetic data until the development of high-throughput sequence techniques in the late 1990s [22, 123, 130]. With ongoing efforts of high-throughput sequencing jump started by the Human Genome Project [123], we are now at an unprecedented stage where genetic variations are gathering at an exponential rate. Such quantities of genetic data provide enormous opportunities for us to examine and understand the history of human population as well as the rise of diseases in unprecedented detail. However, with such enormous amounts of genetic data, we face the challenge of developing efficient and accurate algorithms for analyzing large-scale datasets. Therefore, one of the intents of this thesis is to develop a way to solve some of the problems in the inference of population history in the context of large genetic variation datasets.

1.1 Genetic Variations

Variations can occur within and among populations, within and between species, and in phenotypic features as well as in genetic materials. When variation occurs at the DNA level, we call such variation genetic variation. Genetic variation is important because it is what makes us different and it provides clues to a number of questions from how we arise as a species to how a disease may have arisen. Genetic variation is typically brought by different mutational forces that can be largely categorized into two groups: point mutations and structural variations (Figure 1.1). Point mutation occurs when a DNA base is substituted with another base. Structural variation occurs when a DNA sequence is inserted, deleted, duplicated, or inverted.

Before high-throughput technologies were developed, detection of point mutations was mainly achieved through restriction enzyme assays that identify restriction fragment length polymor-

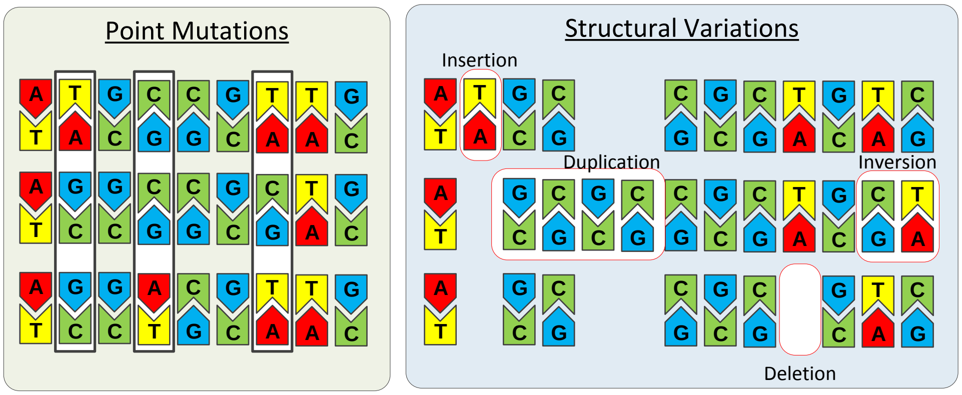


Figure 1.1: Genetic variations can largely be divided into two groups: Point mutations and structural variations. Point mutations are genetic variations caused by substitutions of bases while structural variations are genetic variations due to insertions, deletions, duplications, inversions, and translocations.

phisms (RFLPs) [36, 44, 107]. RFLP employs a technique for fragmenting a DNA sample by restriction enzymes that can recognize and cut DNA at specific locations. Once DNAs are fragmented by the restriction enzyme into different length fragments, gel electrophoresis then separates the fragments by their lengths. If a mutation occurs within one of the cleavage sites, the restriction enzyme would no longer able to cleave the site, resulting in longer fragments on samples with such a mutation. By comparing the lengths of DNA fragments resulting from restriction enzyme cleavage on gel electrophoresis between sample and control groups, one can identify if a particular point mutation occurs.

In addition to RFLP, traditional sequencing techniques through automated chain-termination DNA sequencing were also used to identify single nucleotide polymorphisms (SNPs). SNPs are single point mutations that occur throughout the genome where the bases are switched from one nucleotide to another. These variations can result in changes in protein sequence that may lead to certain diseases. After high-throughput techniques were introduced, detecting and typing SNPs through microarray chips became very popular. To detect SNPs, one would first sequence a small region or the entire genome from a small sample of individuals. By aligning the sequences, one can then identify the bases that are polymorphic. SNPs can then be typed by running samples on microarray chips with probes representing short sequences around each polymorphic site. While SNPs genotyping using microarrays known as tiling arrays is the most common approach today, efforts to sequence the entire genome for all samples are becoming more and more popular today as the cost of whole-genome sequencing becomes affordable.

A second group of genetic variations is known as structural variations. Although structural variation was initially believed to be of lesser importance, researchers have begun to recognize its importance in disease association [32]. One way to detect a type of structural variation is through polymerase chain reaction (PCR) [97] that identifies microsatellite polymorphisms [88]. Microsatellite polymorphisms are short repeating sequences ranging between 2 and 6 base pairs that vary in the number of repeat copies. These polymorphisms can be detected and typed by amplifying the microsatellite region using PCR with specific primers outside the microsatellite region and then separating different lengths of the microsatellite using gel electrophoresis. Those individuals with heterozygous allele would have two different bands on the gel, while those with homozygous major or minor alleles would have just a single band on the gel.

In addition to microsatellite polymorphism, detection of other structural variations can be achieved through high-throughput techniques via sequencing or tiling arrays. Although there are fewer structural variations compared to SNPs, researchers have shown that structural variations can also result in disease phenotypes [108, 128]. Detections of larger structural variations are commonly conducted through array comparative genome hybridization (aCHG) [82] by measuring a sample's florescent intensity compared to a reference sample, but a recent advances in sequencing technology have led to a newer approach known as paired-end mapping that not only enables detection of insertion/deletion polymorphisms but also translocations and inversions [56].

1.2 Genetic Variation Datasets

With different types of genetic variations data and different genotyping and sequencing efforts managed by different groups, locating specific genetic data can be difficult. Luckily, efforts to collect data from multiple studies into a centralized location have been initiated. A database known as dbSNP was initiated by National Center for Biotechnology Information (NCBI) in 1998 to enable researchers to submit newly identified genetic variations [99]. To search for existing genetic variations submitted to dbSNP, one could use the NCBI's Entrez SNP search tool to learn about a particular genetic variation [72], a set of SNPs within a particular gene, or even the set of SNPs within an entire chromosome. Alternatively, one could also utilize a genome browser, such as the UCSC genome browser, to learn about genetic variations across different regions of the genome [55, 69].

While dbSNP and associated browsers allow one to search for genetic variations identified by various studies, data genotyped and sequenced for known SNPs on cohorts of samples needed for actual analyses are typically deposited on different websites and databases. For small and medium scale studies on collecting genetic variation data from different cohorts of individuals, one can often find the sample data in National Center for Biotechnology Center's (NCBI) database of genotype and phenotype (dbGaP) [64]. The database contains information on each genetic variation study listed, including the study documentation, phenotypic data, genetic data, and statistical results. While aggregated information such as statistical analysis and summary descriptions are available publicly, access to individual level information including genotypic data requires one to apply for access.

As an alternative to dbGaP, large-scale whole genome genetic variation data are also available from a number of resources as summarized in Table 1.1. Among the databases listed in Table 1.1, HapMap is perhaps the most well-known whole genome genetic variation dataset, consisting of over 1.6 million SNPs from 1,184 reference individuals from 11 global populations in its phase 3 release [2, 4, 5]. In addition to HapMap, a number of large scale datasets using genotyping technologies have emerged including Human Genome Diversity Project (HGDP) [50], Population Reference Sample (POPRES) [76], Japanese Single Nucleotide Polymorphism (JSNPA) [46], and Pan-Asian SNP (PASNP) [25]. While most large scale projects employ genotyping technologies, a newer project known as 1,000 Genome Project is the first large scale project to sequenced entire genomes on more than 1,000 individuals [6].

Table 1.1: **List of Some Important Large-Scale Genetic Variation Datasets**

Database	Data Types	Populations	Samples
HapMap[24]	SNP (1.6M), CNV	11	1184
HGDP[95]	SNP(500K), CNV (1k)	29	485
1000 Genome[23]	SNP (38M), CNV, Ins/Del/Inv	14	1092
POPRES[75]	SNP(500K), CNV	7	5886
PASNP[1]	SNP (56K), CNV	71	1982
JSNP[71]	SNP(500K)	1	934

1.3 Inference of Population History

Past work on population history inference has essentially involved two inference problems: identifying meaningful population groups and ancestry inference among them. In this section, we survey major methods for these separate inference problems.

1.3.1 Population Substructure

Population groups or substructures may be assumed in advance based on common conceptions of ethnic groupings, although the field increasingly depends on computational analysis to make such inferences automatic. Probably the most well-known system for identifying population substructure is STRUCTURE [85]. STRUCTURE infers population substructures from genetic variation data using a probabilistic model that assumes each population is characterized by a set of frequencies for each variant form, or allele, across variation sites, or loci, in the dataset. Assuming that the allele of each locus for each individual is dependent on the allele frequency of the subpopulations the individual belongs to, STRUCTURE tries to identify the probability distribution of the ancestral origin Z of each individual and the allele frequencies P of each subpopulation given the observed genetic variation data X. Namely, STRUCTURE aims to learn the distribution

$$Pr(Z, P|X) \propto Pr(X|Z, P)Pr(Z)Pr(P)$$

using a Markov Chain Monte Carlo (MCMC) method to group sequences into K ancestral population groups each with its own allele frequency profile.

Another well known program is EIGENSOFT [81], which uses principal components analysis (PCA) to identify a set of distinguishing vectors of alleles that allow one to spatially separate a set of individuals into subgroups. Recently, two additional algorithms known as Spectrum [105] and mStruct [101] have been proposed by Sohn and Xing and Shringarpure and Xing respectively. While both algorithms are similar in nature to STRUCTURE, Spectrum constructs a more realistic model by incorporating recombinations and mutations into their statistical model and avoids the specification of ancestral population number *a priori* by modeling genetic polymorphism based on the Dirichlet process. On the other hand, mStruct proposes a new admixture model to identify subgroups by representing each population as *mixtures of ancestral alleles* rather than a single ancestral allele profile.

1.3.2 Phylogenetic Analysis for Ancestry Inference

Traditionally, analysis of ancestry between individuals has largely been done through the use of classic phylogenetic algorithms. Defined as methods to infer evolutionary relationship between different taxa or individuals using a tree and, in some more complicated cases, a graph, phylogenetic algorithms can be largely divided into two general classes of algorithm: distance-based and character-based. Distance-based phylogenetic algorithms aim to piece together the relationships between taxa or individuals by using a measure of evolutionary distances between taxa or individuals. Pairwise distances are typically computed between every pair of taxon or individual and are then used to construct a tree in which the phylogenetic distances between taxa or individuals closely resemble the computed distances. While a number of distance-based methods exist, they can largely be grouped into non-objective based and objective based methods. Among non-objective based methods, two of the most well-known are the Unweighted Pair-Group Method Using Arithmetic Averages (UPGMA) [31] and Neighbor Joining (NJ) [31]. Both methods compute a tree progressively from the bottom-up by joining two closest taxa into a single tree node and updating the distance matrix at each step until all taxa are joined into a tree. While the two methods are similar, NJ differs from UPGMA in its updates of distance matrix in that NJ incorporates different mutation rate at different tree branches into distance calculations. This makes NJ a better choice of algorithm when the mutation rate is variable. While NJ and UPGMA are popular distance-based methods, a second group of distance-based methods using an objective function are also gaining popularity. Objective-based methods, such as minimum evolution, aim to optimize for the best tree using objective functions such as the sum of the edge weights. Although with higher computational cost, objective-based methods have the advantage of having theoretical guarantees of identifying the optimal tree by some precise criteria by searching through all possible trees rather than greedily looking at a subset of all possible trees in the case of non-objective based methods.

A second class of phylogenetic algorithms is the character-based approach. A character-based algorithm takes an aligned set of characters, such as DNA sequences, and constructs a tree describing the changes in individual characters needed to produce the observed set of characters. Each node in the tree would represent a unique string of characters and each edge connected to a node would describe the changes to the character that lead to a new string of character from another node. Character-based algorithms can largely be divided into three groups: maximum parsimony, maximum likelihood, and Bayesian. In maximum parsimony, the goal of the algorithms is to identify the tree that minimizes the total number of changes or mutations occurred along the edges of the tree. The intuition behind maximum parsimony is that repeated or recurrent mutations are typically rare. Thus, by optimizing for the minimum number of mutations that have occurred throughout history, maximum parsimony would give us a tree satisfying such assumption. Because maximum parsimony is one of the first class of methods introduced, a number of well-known software suites have utilized this approach, including Phylip, PAUP, and more recent mixed-ILP methods [30, 96, 109]. The advantage of the maximum parsimony is that the method utilizes a simple but informative model of molecular evolution that can provide correct evolutionary trees in some regions of the genome that may be under selective pressures that prevents frequent mutation or for short time scales where few mutations would be expected. However, the method is generally much more computational intensive than most distance-based methods and can produce incorrect tree when the assumption is violated.

Another group of character-based method is the maximum likelihood (ML) approach, where finding the optimal tree is proposed in a probabilistic framework. In a maximum likelihood approach, the method finds the optimal tree by maximizing the likelihood function, $P(D|M)$, where data (D) is the observed sequences and the model (M) is the set consisting of the tree topology, ancestral sequences, and other parameters. Such an approach can provide a finer and generally more accurate depiction of the evolutionary history than the maximum parsimony approach when the parsimony assumption no longer holds, but is generally more computational costly than the maximum parsimony approach.

In addition to maximum likelihood, a third group of character-based methods is the Bayesian approach [49]. Rather than maximizing the probability function $P(D|M)$, a Bayesian method tries to learn the posterior distribution $P(M)$ over possible trees, sequences, and parameters. While the Bayesian approach is generally harder computationally, it has the advantage of not requiring the users to specify parameters that can bias the tree inference.

When comparing the two main classes of phylogenetic tree reconstruction algorithm, there is a general consensus that character-based approach provide a more realistic and generally more accurate and detailed depiction of the evolutionary history but suffers from high computational cost that limits its usefulness on large genomic datasets. As a result, distance-based methods are still currently the only feasible choice in building evolutionary trees from large genome-scale datasets. Therefore, part of this thesis is to provide an efficient solution in learning population history using character-based methods.

1.4 Ancestry Inference in the Presence of Admixture

While ancestry inference through traditional phylogenetic algorithms generally works well when the populations rarely interact with one another, traditional phylogenetic methods can fail when there are interactions between individuals from different populations. When individuals from one population migrate and come into contact with another population that was long separated, incorporation of genetic materials from one distinct population into another can result. This process of mixing genetic material from different populations is known as 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. When one is interested in detecting and learning ancestral history in the presence of admixed individuals, traditional phylogenetic analyses may not necessarily produce correct results. Imagine if we have a group of admixed individuals that have a mixture of genetic materials from two different populations in the same dataset. In the best-case scenario, the traditional phylogenetic tree algorithm would simply attach the admixed individuals as a sub-branch to one of the parental populations. However, it is more likely that the algorithm would return an evolutionary tree that is far from the true evolutionary history, where the topology of the tree is reshuffled due to the mixing of genetic materials from admixed individuals. As a result, a different set of tools and algorithms are needed to learn about admixture.

One popular approach for analyzing admixture is principal component analysis (PCA) [81]. PCA is a type of techniques for taking high-dimensional data and transform them into a more tractable, lower-dimensional form, without losing too much information. Mathematically, PCA tries to minimize the projection residuals when transforming p-dimensional data into lower-dimension form:

$$\min \frac{1}{n} \sum_{i}^{n} \sum_{j}^{k} ||\vec{x}_i - (\vec{x}_i \cdot \vec{w_j})\vec{w_j}||^2$$

where $\vec{x}_i$ is the p-dimensional vector of ith data point and $\vec{w_j}$ is the jth orthonormal vector. The minimization can be achieved by finding the eigenvectors and eigenvalues of the data where each eigenvector is associated with an eigenvalue. The value of the eigenvalue indicates how large the variance of the data is when projecting onto the corresponding eigenvector. The idea behind PCA for ancestry analysis is that user would take the genetic variation data as a matrix, learn the eigenvalues and eigenvectors of the matrix, and project each individual onto the largest k eigenvectors to visualize individuals' genetic variance across populations. Since variances across populations are usually the largest, individuals from each population should nicely project into different population clusters using the first few eigenvectors with the largest eigenvalues. When applying PCA on a dataset with admixed individuals, the admixed individuals would generally be projected linearly between the centers of two or more parental populations. This approach is popular due to its low computational cost and its ability to easily visualize the separation and intermixing of populations. Nonetheless, the PCA-based approach generally does not have an easy and accurate way to quantify the separation or the intermixing of populations.

To quantify the amount of admixtures between populations or among individuals, one common approach is the admixture model-based methods that model individuals as probabilistic mixtures from k ancestral population. Such an approach can typically perform detailed estimations of the admixture proportions at the individual level or even at the loci level for each individual. While a number of likelihood-based methods exist [98], one common implementation is the hidden Markov model (HMH) [84, 111]. An HMM is a probabilistic graphical model that assumes a Markov process with hidden states. A general HMM framework for inferring admixture generally models the ancestry composition of an individual at each genetic variation site as the hidden state that must be inferred from the genotypes. Each hidden state is connected to its neighbors by a chain where the probability of the hidden state is conditionally dependent on the states of its neighbors. By using the observed genotypes and the correlations between the nearby markers, the HMM can then produce a probability map quantifying the ancestry of each individual. While there are a number of different HMM-based methods introduced in recent years, they are mainly based on the same framework with additional improvements such as inclusion of linkage disequilibrium (LD) or other hyperparameters. In addition to HMM, other likelihood methods such as LAMP [98], FRAPP [110], and ADMIXTURE [8] are also popular for quantifying admixture.

Despite success in learning admixture using PCA and admixture model-based approaches, neither approach provides a way to fully illustrate the complete evolutionary history, such as the relationships between the non-admixed populations or the precise time at which the admixture happened. To learn about the time of admixture and the possible relationships between populations, a third type of admixture inference algorithm known as the coalescent-based algorithms can be used. In coalescent-based algorithms, models of general population history with different time and admixture parameters are evaluated by enumerating all possible trees generated from a coalescent model consistent with the general population model and then computing the probability of observing the data given the generated coalescent trees [17, 77, 126]. Coalescent-based methods generally have the advantage that these methods can provide additional evolutionary information, such as the time of the admixture and time of divergence in which one may be interested in phylogenetic analysis. While coalescent-based methods can provide additional evolutionary information, existing methods suffer from expensive computational cost as well as the requirement to know the model of population history beforehand instead of learning it from the data directly. Despite the limitations of current coalescent-based methods, the ability to learn additional evolutionary information is desirable. Therefore, addressing the limitations of the coalescent-based methods will be a focus in this thesis.

1.5 Limitations of Existing Approaches for Learning Population History

Efforts at learning the history of populations from genetic data remain a problem solved in bits and pieces: from population assignments to evolutionary events inferences to parameter estimation. While there have been significant advances recently in subpopulation detection [84, 85, 101], in phylogenetic inference [31], and in parameter estimation [17, 126], there is no single method that learns all the information needed to give a detailed depiction of how different populations emerged over time and, perhaps more importantly, how long ago the populations emerged. Methods for identifying substructure in a dataset can provide highly accurate mapping of an ancestral origin for each region of the individual's chromosome [98, 111] but leave out information regarding the relationships between ancestral origins. On the other hand, classical phylogenetic methods [30] provide highly detailed evolutionary relationships between individuals but are mostly limited to tree-like structures. Furthermore, phylogenetic inferences frequently require large datasets to achieve statistical significance and confidence but become computational infeasible when given large datasets. Similarly, algorithms for estimating parameters of evolutionary events can be computational intensive [11, 17] and require a restrictive assumption that the history of the population is known or assumed beforehand. Some parameter estimators circumvented the computational issue but, in exchange, only estimate a subset of the parameters, such as admixture [84, 85].

1.6 Contributions

Despite different types of methods excelling in learning different aspects of the population history, no single method of which we are aware can provide a full picture of the population history, which not only can be informative and time-saving to researchers but also helpful in enhancing the accuracy of the estimations. For example, when using a divergence time estimator that did not take admixture into account, the time estimated could significantly deviate from the true divergence time if admixture events have actually occurred. As a result, given the potential advantages of joint learning of multiple aspects of population history, the goal of this thesis is to work toward unifying different aspects of the inference of population history into one algorithmic package. Since inference of population history can encompass a broad range of problems, we here specifically try to unify the problem of population substructure, the inference of evolutionary events involving divergence events and/or admixture events, and the exact times and admixture fractions describing the events given large datasets.

The key contribution of this thesis is the development of novel algorithms for automatically learning detailed descriptions of population history from large scale genetic variation datasets with and without the presence of admixture. The thesis first describes a model to learn population trees from large genomic datasets under the assumption that no admixtures occurred throughout the history of the populations. The method described here employs a character-based algorithm to take advantage of its better modeling of the evolutionary processes but avoids the high computational cost by generating small phylogenetic trees on fragments of the complete dataset and then infers robust tree branches across the tree set. In addition to solving the computational issue for learning evolutionary history from large datasets, another contribution of this work is to combine the inference of the population substructures along with the history of the populations as both problems depend on similar data sources and in principle can help inform the decisions of one another. Through a series of tests on both simulated and real datasets, this thesis demonstrates the feasibility of automatically learning of population substructures and their relationships in a reasonable time frame.

Analysis on the evolutionary history of human populations typically assumes a tree-like structure and ignores the migratory nature of human populations. While a tree assumption has long worked for evolutionary analysis on distance species, such an assumption may not always hold for closely-related species or intra-species analysis. Admixtures, a result of the migratory nature of human populations, have been proven to be a crucial factor in the analysis of human population history and an important step in understanding the etiology of diseases. Methods for detecting and quantifying admixtures are on the rise in recent years, but these methods usually look at a limited aspect of the whole admixture history or lack the capability to analyze large quantities of data to provide a fuller picture of the evolutionary history of human populations. To resolve these issues, the second contribution of this thesis is the development of a novel algorithm capable of running on large-scale datasets for learning the time and admixture parameters describing a population history involving two non-admixed populations and one admixed population.

As a natural extension to automatic learning of parameters of population history involving two non-admixed and one admixed populations, a third contribution of this thesis is to expand previous algorithm of learning parameters of population history for two non-admixed population and one admixed population to learn the precise parameters and population model for any arbitrary number of subpopulations.

Finally, to explore the possible applications of learning population history from large genomic datasets, one final contribution in this thesis is to propose and test a simple structured association test statistic that effectively removes the effect of population substructure learned from our prior algorithms.

1.7 Thesis Organization

Chapter 2 gives a detail description of the computational method for joint inference of population substructures and their evolutionary history from large scale genomic datasets under the assumption that there is no admixture. A series of validations done through simulated and real datasets are also conducted and detailed in the chapter. Chapter 3 describes a coalescent-based algorithm involving a two-step process for learning the parameters of a population history from large-scale genomic data involving two non-admixed populations and one admixed population. As a natural extension of the algorithm described in Chapter 3, Chapter 4 details a generalized algorithm for automatic identification of population substructures, their evolutionary histories, and the specific parameters pertaining to each evolutionary event from large-scale datasets with or without the presence of admixture for any arbitrary number of subpopulations in the dataset. In Chapter 5, we describe a simple structured test statistic to test the applicability of population history learned from genomic datasets. Finally, Chapter 6 summaries the findings of these studies and their conclusions and outlines possible directions for future work on this topic.

Chapter 2

Learning Population Histories From Large-Scale Datasets in the Absence of Admixtur[1]

The recent completion of the human genome [21, 124] and the subsequent discovery of millions of common genetic variations in the human genome [100] has created an exciting opportunity to examine and understand how modern human population arose from our common ancestor at unprecedented detail. Several major studies have recently been undertaken to assess genetic variation in human population groups, thus enabling the detailed reconstruction of the ancestry of human population groups [4, 10, 50, 76]. In addition to its importance as a basic research problem, human ancestry inference has great practical relevance to the discovery of genetic risk factors of disease due to the confounding effect of unrecognized substructure on genetic association tests [114].

As discussed in Chapter 1, past work on human ancestry inference has treated ancestral inference as two distinct inference problems: identifying meaningful population groups and inferring evolutionary trees among them. While most earlier works performed the task of identifying

[1] This chapter was developed from material published in [119]

meaningful population groups manually by assuming in advance the groups based on common conceptions of ethnic groupings, the field has increasingly rely on computational analysis to make such inferences automatically. Two popular approaches for learning population substructures are STRUCTURE [85] and EIGENSOFT [81] that uses probabilistic model and principal component anaylsis (PCA) to identify fine population structure from genetic dataset.

A separate literature has arisen on the inference of relationships between populations, typically based on phylogenetic reconstruction of limited sets of genetic markers — such as classic restriction fragment length polymorphisms [74], mtDNA genotypes [14, 52], short tandem repeats [52, 116], and Y chromosome polymorphism [41] — supplemented by extensive manual analysis informed by population genetics theory. While current phylogenetic reconstruction algorithms, such as maximum parsimony or maximum likelihood, work well on small datasets with little recombination, most do not work well when utilizing genome wide datasets. Furthermore, there has thus far been little cross-talk between the two problems of inferring population substructure and inferring phylogenetics of subgroups, despite the fact that both problems depend on similar data sources and in principle can help inform the decisions of one another.

To unify these two inference problems, this chapter introduces a novel approach for reconstructing a species history conceptually based on the idea of consensus trees [73], which represent inferences as to the robust features of a family of trees. The approach takes advantage of the fact that the availability of large-scale variation data sets, combined with new algorithms for fast phylogeny inference on these data sets [96], has made it possible to infer likely phylogenies on millions of small regions spanning the human genome. The intuition behind this method is that each such phylogeny will represent a distorted version of the global evolutionary history and population structure of the species, with many trees supporting the major splits or subdivisions between population groups while few support any particular splits independent of those groups. By detecting precisely the robust features of these trees, we can assemble a model of the true evolutionary history and population structure that can be made resistant to overfitting and to noise in the SNP data or tree inferences.

For the remainder of this chapter, Section 2.1 will present a detailed description of the mathematical model of the consensus tree problem and a set of algorithms for finding consensus trees from families of local phylogenies. Section 2.2 presents strategies for evaluating the method on a set of simulated data and two real datasets from the HapMap Phase II [4] and the Human Genome Diversity Project [50]. Section 2.3 then shows the results of the validation experiments. Finally, Section 2.5 considers some of the implications of the results and future prospects of the consensus tree approach for evolutionary history and substructure inference.

2.1 Methods

2.1.1 Consensus Tree Model

Assume we are given a set S of m taxa representing the paired haplotypes from each individual in a population sample. If we let $\mathcal{T}$ be the set of all possible labeled trees connecting the $s \in S$, where each node of any $t \in T$ may be labeled by any subset of zero or more $s \in S$ without repetition, then our input will consist of some set of n trees $\mathcal{D} = (T_1, \ldots, T_n) \subseteq \mathcal{T}$. Our desired output will also be some labeled tree $T_M \in \mathcal{T}$, intended to represent a consensus of $T_1, \ldots, T_n$.

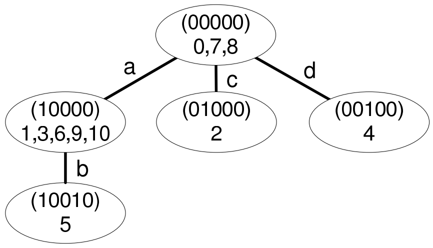
The objective function for choosing T_M is based on the task of finding a consensus tree [73] from a set of phylogenies each describing inferred ancestry of a small region of a genome. The consensus tree problem aims to identify tree structure that is persistent across a set of trees. The typical approach for finding the optimal consensus tree involves counting occurrences of each edge across the set of trees. If the frequency of the edge exceeds some threshold, the edge will be incorporated into the consensus tree. The present application is, however, fairly different from standard uses of consensus tree algorithms in that the phylogenies are derived from many variant markers, each only minimally informative, within a single species. Standard consensus tree approaches, such as majority consensus [65] or Adam consensus [7], would not be expected to be effective in this situation as it is likely there is no single subdivision of a population that is consistently preserved across more than a small fraction of the local intraspecies trees and that many similar but incompatible subdivisions are supported by different subsets of the trees. We therefore require an alternative representation of the consensus tree problem designed to be robust to large numbers of trees and high levels of noise and uncertainty in data.

Given such criterion, a model of the problem based on the principle of minimum description length (MDL)[38] was chosen. The principle of minimum discription length is a standard technique for avoiding overfitting when making inferences from noisy data sets. An MDL method models an observed data set by seeking to minimize the amount of information needed to encode the model and to encode the data set given knowledge of the model. Suppose we have some function $L : \mathcal{T} \rightarrow \mathcal{R}$ that computes a description length, $L(T_i)$, for any tree T_i. We will assume the existence of another function, which for notational convenience we will also call L, $L : \mathcal{T} \times \mathcal{T} \rightarrow \mathcal{R}$, which computes a description length, $L(T_i|T_j)$, of a tree T_i given that we have reference to a model tree T_j. Then, given a set of observed trees, $\mathcal{D} = \{T_1, T_2, ..., T_n\}$ for $T_i \in \mathcal{T}$, our objective function is

$$\mathcal{L}(T_M, T_1, \ldots, T_n) = \\ \arg\min_{T_M \in \mathcal{T}} \left(L(T_M) + \sum_{i=1}^{n} L(T_i|T_M) + f(T_M) \right)$$

The first term computes the description length of the model (consensus) tree T_M. The sum computes the cost of explaining the set of observed (input) trees $\mathcal{D}$. The function $f(T_M) = c|T_M| \log_2 m$ defines an additional penalty on model edges where c is a constant used to define a minimum confidence level on edge predictions. The higher the penalty term, the stronger the support for each edge must be for it to be incorporated into the consensus tree.

We next need to specify how to compute the description length of a tree. For this purpose, this method use the fact that a phylogeny can be encoded as a set of bipartitions (or *splits*) of the taxa with which it is labeled, each specifying the set of taxa lying on either side of a single edge of the tree. The algorithm represent the observed trees and candidate consensus trees as sets of



(a)

e_a:	1,3,5,6,9,10|0,2,4,7,8	e_a:	01010110011
e_b:	0,1,2,3,4,6,7,8,9,10|5	e_b:	00000100000
e_c:	0,1,3,4,5,6,7,8,9,10|2	e_c:	00100000000
e_d:	0,1,2,3,5,6,7,8,9,10|4	e_d:	00001000000

(b) (c)

Figure 2.1: (a) A maximum parsimony (MP) tree consisting of 11 labeled individuals or haplotypes. (b) The set of bipartitions induced by edges (e_a, e_b, e_c, e_d) in the tree. (c) 0-1 bit sequence representation for each bipartition.

bipartitions for the purpose of calculating description lengths. Once the method identified a set of bipartitions representing the desired consensus tree, the method then apply a tree reconstruction algorithm to convert those bipartitions into a tree.

A bipartition b can in turn be represented as a string of bits by arbitrarily assigning elements in one part of the bipartition the label "0" and the other part the label "1". As an example, in the tree of Fig. 2.1(a), the edge labeled a induces the bipartition $\{1,3,5,6,9,10\} : \{0,2,4,7,8\}$. This edge would have the bit representation "10101001100." Such a representation allows us to compute the encoding length of a bipartition b as the entropy [38] of its corresponding bit string. If we define $H(b)$ to be the entropy of the corresponding bit string, p_0 to be the fraction of bits of b that are zero and p_1 as the fraction that are one, then:

$$\begin{aligned} L(b) &= mH(b) \\ &= m\left(-p_0 \log_2 p_0 - p_1 \log_2 p_1\right) \end{aligned}$$

Similarly, we can encode the representation of one bipartition b_1 given another b_2 using the concept of conditional entropy. If we let $H(b_1|b_2)$ be the conditional entropy of bit string of b_1 given bit string of b_2, p_{00} be the fraction of bits for which both bipartitions have value "0," p_{01} be the fraction for which the first bipartition has value "0" and the second "1," and so forth, then:

$$\begin{aligned} L(b_1|b_2) &= mH(b_1|b_2) \\ &= m\left[H(b_1, b_2) - H(b_2)\right] \\ &= m\left[\sum_{s,t\in\{0,1\}} -p_{st}\log_2 p_{st} + \right. \\ &\qquad \left. \sum_{u\in\{0,1\}} (p_{0u} + p_{1u})\log_2(p_{0u} + p_{1u})\right] \end{aligned}$$

where the first term is the joint entropy of b_1 and b_2 and the second term is the entropy of b_2.

We can use these definitions to specify the minimum encoding cost of a tree $L(T_i)$ or of one tree given another $L(T_i|T_M)$. We first convert the tree into a set of bipartitions $b_1, \ldots, b_k$. We can then observe that each bipartition b_i can be encoded either as an entity to itself, with cost equal to its own entropy $L(b_i)$, or by reference to some other bipartition b_j with cost $L(b_i|b_j)$. In addition, we must add a cost for specifying whether each b_i is explained by reference to another bipartition and, if so, which one. The total minimum encoding costs, $L(T_M)$ and $L(T_i|T_M)$, can then be computed by summing the minimum encoding cost for each bipartition in the tree. Specifically, let $b_{t,i}$ and $b_{s,M}$ be elements from the bipartition set B_i of T_i and B_M of T_M, respectively. We can then compute $L(T_M)$ and $L(T_i|T_M)$ by optimizing for the following objectives over possible reference bipartitions, if any, for each bipartition in each tree:

$$L(T_M) = \underset{b_s\in B_M\cup\{\emptyset\}}{\arg\min} \sum_{s=1}^{|B_M|} \left[L(b_{s,M}|b_s) + \log_2\left(|B_M| + 1\right)\right]$$

$$L(T_i|T_M) = \underset{b_t\in B_M\cup B_i\cup\{\emptyset\}}{\arg\min} \sum_{t=1}^{|B_i|} \left[L(b_{t,i}|b_t) + \log_2\left(|B_M| + |B_i| + 1\right)\right]$$

2.1.2 Algorithms

Encoding Algorithm To optimize the objectives for computing $L(T_M)$ and $L(T_i|T_M)$, we can pose the problem as a weighted directed minimum spanning tree (DMST) problem by constructing a graph, illustrated in Fig. 2.2, such that finding a directed minimum spanning tree allows us to compute $L(T_M)$ and $L(T_i|T_M)$. We construct a graph $G = (V, E)$ in which each node represents either a bipartition or a single "empty" root node r explained below. Each directed edge (b_j, b_i) represents a possible reference relationship by which b_j explains b_i. If a bipartition b_i is to be encoded from another bipartition b_j, the weight of the edge e_{ji} would be given by $w_{ji} = L(b_i|b_j) + \log_2 |V|$ where the term $\log_2 |V|$ represents the bits we need to specify the reference bipartition (including no bipartition) from which b_i might be chosen. This term introduces a penalty to avoid overfitting. We add an additional edge directly from the empty node to each node to be encoded whose weight is the cost of encoding the edge with reference to no other edge, $w_{empty,j} = L(b_j) + \log_2 |V|$.

To compute $L(T_M)$, the bipartitions B_M of T_M and the single root node collectively specify the complete node set of the directed graph. One edge is then created from every node $B_M \cup \{r\}$ to every node of B_M. To compute $L(T_i|T_M)$, the node set will include the bipartitions B_i of T_i, the bipartitions B_M of T_M, and the root node r. The edge set will consist of two parts. Part one consists of one edge from each node of $B_i \cup B_M \cup \{r\}$ to each node of B_i, with weights corresponding to the cost of possible encodings of B_i. Part two will consist of a zero-cost edge from r to each node in B_M, representing the fact that the presumed cost of the model tree has already been computed. Fig. 2.2 illustrates the construction for a hypothetical model tree T_M and observed tree T_i (Fig. 2.2(a)), showing the graph of possible reference relationships (Fig. 2.2(b)), a possible solution corresponding to a specific explanation of T_i in terms of T_M (Fig. 2.2(c)), and the graph of possible reference relationships for T_M by itself (Fig. 2.2(d)).

Given the graph construction, the minimum encoding length for both constructions is found by solving for the DMST with the algorithm of Chiu and Liu [18] and summing the weights of

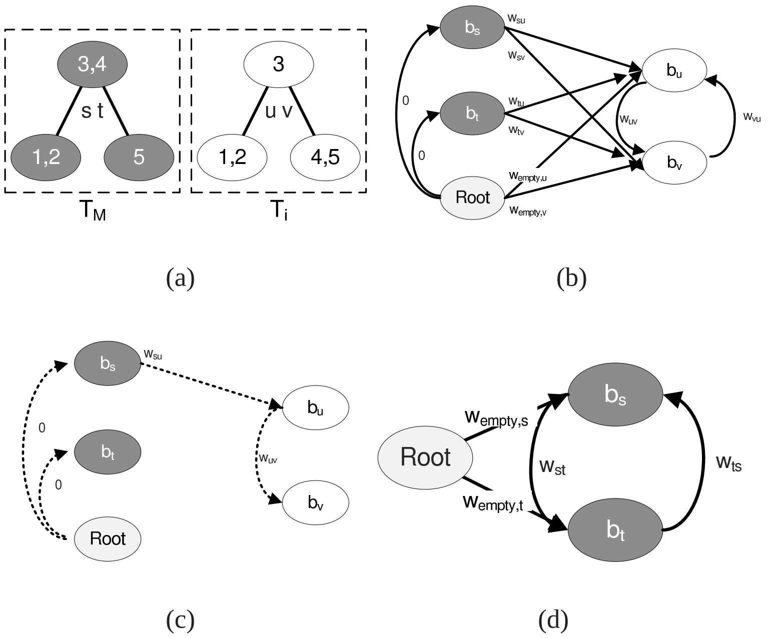


Figure 2.2: Illustration of the DMST construction for determining model description length. (a) Hypothetical model tree T_M (gray) and observed tree T_i (white). (b) Graph of possible reference relationships for explaining T_i (white nodes) by reference to T_M (gray nodes). (c) A possible resolution of the graph of (b). (d) Graph of possible reference relationships for explaining T_M by itself.

the edges. This cost is computed for a candidate model tree T_M and for each observed tree T_i, for $i = 1, ..., n$, to give the total cost $[\mathcal{L}(T_M, T_1, \ldots, T_n)]$.

Tree Search While the preceding algorithm gives us a way to evaluate $L(T_M)$, $L(T_i|T_M)$, and $\mathcal{L}(T_M, T_1, ..., T_n)$ for any possible consensus tree T_M, we still require a means of finding a high-quality (low-scoring) tree. The space of possible trees is too large to permit exhaustive search and we are unaware of an efficient algorithm for finding a global optimum of our objective function. We therefore employ a heuristic search strategy based on simulated annealing. The algorithm relies on the intuition that the bipartitions to be found in any high-quality consensus tree are likely to be the same as or similar to bipartitions frequently observed in the input trees. The algorithm runs for a total of t iterations and at each iteration i will either insert a new bipartition chosen uniformly at random from the observed (non-unique) bipartitions with probability $1 - i/t$

or delete an existing bipartition chosen uniformly at random from the current T_M with probability i/t to create a candidate model tree T'_M. This strategy is intended to encourage the addition of new bipartitions at the beginning of the search and the cleanup of redundant bipartitions at the end of search cycle.

If the algorithm chooses to insert a new bipartition b, it then performs an additional expectation-maximization-like (EM) local optimization to improve the fit, as many of the bipartitions in the observed trees will be similar but not exact matches to the global splits inferred for the populations. The EM-like local optimization repeatedly identifies the set B_o of observed bipartitions explained by b and then locally improves b by iteratively flipping any bits that lower the cost of explaining B_o, continuing until it converges on some locally optimal b. This final bipartition is then added to T_M to yield the new candidate tree T'_M. Once a new candidate tree T'_M has been established, the algorithm tests the difference in cost between T_M and T'_M. If T'_M has reduced cost then the move is accepted and T'_M becomes the new starting tree. Otherwise, the method accepts T'_M with probability $p = \exp \frac{\mathcal{L}(T_M, T_1, \ldots, T_n) - \mathcal{L}(T'_M, T_1, \ldots, T_n)}{T}$ where $T = 400/t$ is the simulated annealing temperature parameter.

Tree Reconstruction A final step in the algorithm is the reconstruction of the consensus tree from its bipartitions. Given the bipartitions found by the tree search heuristics, we first sort the model bipartitions $b_1 \prec b_2 \ldots \prec b_k$ in decreasing order of numbers of splits they explain (i.e., the number of out-edges from their corresponding nodes in the DMST). The method then initialize a tree T_0 with a single node containing all haplotype sequences in S and introduce the successive bipartitions in sorted order into this tree. The intuition is that bipartitions that explain a greater fraction of the observed variation should generally correspond to earlier divergence events. For each $b_i = 1$ to k, the method subdivide any node v_j that contains elements with label 0 in b_i (b_i^0) and elements labeled as 1 in b_i (b_i^1) into nodes v_{j1} and v_{j2} corresponding to the subpopulations of v_j in b_i^0 or b_i^1. The method also introduce a Steiner node s_j for each node v_j to represent the ancestral population from which v_{j1} and v_{j2} diverged. The method then replace the prior tree

T_{i-1} with $T_i = (V_i, E_i)$ where $V_i = V_{i-1} - \{v_j\} + \{v_{j1}, v_{j2}, s_j\}$ and $E_i = E_{i-1} - \{e = (t, v_j) | e \in E_{i-1}, t \in \text{parent}(v_j)\} + \{e = (t, s_j) | t \in \text{parent}(v_j)\} + \{(s_j, v_{j1}), (s_j, v_{j2})\}$. After introducing all k bipartitions, T_k is the final consensus tree.

2.2 Validation Experiments

2.2.1 Simulated Dataset

Evaluation of the method is initially performed on a simulated dataset consisting of three independent populations, each with 150 individuals (300 chromosomes). To generate the sequence data, we first generated the genealogies, or trees that trace back the possible lineages and history between observed individuals, for each population using the coalescent simulator MS [47] on sequence of length 10^7 base pair long with a mutation rate of 10^{-9}, a recombination rate of 10^{-8}, and an effective population size of 25,000. The resulting simulated branch length between the root node of each population and the leaves was 1,600 generations. In order to simulate the effect of three populations diverging from a common ancestor, we subsequently merged the genealogy trees from each population. We first defined a common ancestor for the root nodes of populations one and two as shown in Fig. 2.3(b) with branch length 1,000 generations between their most recent common ancestor (MRCA) and the root nodes of the two populations. We then defined a common ancestor between the MRCA of populations one and two and the root node of population three, with branch length 1,000 generations to the MRCA of populations one and two, and 2,000 generations to the root node of population three. The sum of branch lengths between any leaf and the MRCA of all of the populations was thus estimated at 3,600 generations. Given this defined tree structure, we generated sequence for each individual using Seq-Gen [89]. We used a mutation rate of 10^{-9} per site to generate a 10 million base pair sequence with 83,948 SNP sites in order to accommodate the branch lengths simulated from MS. Using the 83,948 SNP sites, we constructed 83,944 trees from 5 consecutive SNPs spanning across the sequences. Given the dataset, we ran the algorithms on 10,000 randomly selected trees or their corresponding 33,295 unique SNPs.

2.2.2 Real Data

We further evaluated the method by applying it to samples from two real SNP variation datasets. We first used the Phase II HapMap data set (phased, release 22) [4] which consists of over 3.1 million SNP sites genotyped for 270 individuals from four populations: 90 Utah residents with ancestry from Northern and Western Europe (CEU); 90 individuals with African ancestry from Ibadan, Nigeria (YRI); 45 Han Chinese from Beijing, China (CHB); and 45 Japanese in Tokyo, Japan (JPT). For the CEU and YRI groups, which consist of trio data (parents and a child), we used only the 60 unrelated parents with haplotypes as inferred by the HapMap consortium. For each run, we randomly sampled 10,000 trees each constructed from 5 consecutive SNPs uniformly at random from 45,092 trees generated from chromosome 21, which represented an average of 28,080 unique SNPs. For the purpose of comparison, we used 10,000 trees or the corresponding 28,080 SNPs as inputs to the method and the comparative algorithms. We next used phased data (version 1.3) from the Human Genome Diversity Project (HGDP) [50], which genotyped 525,910 SNP sites in 597 individuals from 29 populations categorized into seven region of origin: Central South Asia (50 individuals), Africa (159 individuals), Oceania (33 individuals), Middle East (146 individuals), America (31 individuals), East Asia (90 individuals), and Europe (88 individuals). For each test with the HGDP data, we sampled 10,000 trees from a set of 39,654 trees uniformly at random from chromosome 1. The 10,000 trees on average consisted of 30,419 unique SNPs.

2.2.3 Benchmarks

There are no known existing method that perform the joint inference of population substructure and the evolutionary tree, and therefore the method cannot be benchmarked directly against any competitor. Consequently, the method was assessed by two criteria. We first assessed the quality of the inferred population histories from the simulated data using the gold standard tree and assessed the quality of the inferred population histories from the real data by reference to a expert-curated model of human evolution derived from a review by Shriver and Kittles[102], which we treat as a "gold standard." Shriver and Kittles used a defined set of known human population groups rather than the coarser grouping inferred by the consensus-tree method. To allow comparison with either of the inferred trees, we therefore merged any subgroups that were joined in our tree but distinct in the Shriver tree and deleted any subgroups corresponding to populations not represented in the samples from which our trees were inferred. (For example, for the HapMap Phase II dataset, we removed Melanesian, Polynesian, Middle Eastern, American, and Central South Asian subgroups from the tree, as individuals from those populations were not typed in the Phase II HapMap). We also ignored inferred admixture events in the Shriver and Kittles tree. We then manually compared our tree to the resulting condensed version of the Shriver and Kittles "gold standard" tree.

As a secondary validation, we also assessed the quality of our inferred population subgroups relative to those inferred by two of the leading substructure algorithms: STRUCTURE (version 2.2) [85] and Spectrum [105]. We selected these programs because of they are well accepted as leading methods for the substructure problem and are able to handle comparable sizes of data set to the method. We chose to omit EIGENSOFT, despite its wide use in this field, as the program is mainly used to visualize substructure and does not lead to an unambiguous definition of substructure to which we can compare. STRUCTURE requires that the user specify a desired number of populations, for which we supplied the true number for each data set (three for simulated data, four for HapMap, and seven for HGDP). For each run of STRUCTURE, we performed 10,000 iterations of burn-in and 10,000 iterations of the STRUCTURE MCMC sampling. We did not make use of STRUCTURE's capacity to infer admixture or to use additional data on linkage disequilibrium between sites. Spectrum did not require any user inputs other than the dataset itself.

We first visualize the cluster assignments by plotting each individual in each population as a vertical line showing the population(s) to which he or she is assigned. Because the clusters assigned by the algorithms have arbitrarily labels, we assign colors to these labels so as to best capture their correspondence to the true population groups. To do so, we first arbitrarily assign a color to each population group in the gold standard. For the consensus tree method, all sequences found in a common node of the consensus tree are considered a single cluster; we assign to each such cluster the color of the gold standard group that has maximum overlap with that cluster. For STRUCTURE, which assigns each individual a probability of being in each cluster, we color each cluster according to the gold standard population that has maximum overlap with the most probable cluster assignments for all individuals. For Spectrum, which assigns each individual a fractional ancestry from a set of inferred founder haplotypes, we choose an arbitrary color for each founder haplotype and color each individual to reflect that individual's inferred fractional ancestries. If we were to use the same assignment protocol for Spectrum as for STRUCTURE, all individuals would be assigned to the same subgroup.

We quantify clustering quality using variation of information [67], a measure commonly used to assess accuracy of a clustering method relative to a pre-defined "ground truth." Variation of information is defined as

$$\mathrm{VI(X,Y)} = 2H(X,Y) - H(X) - H(Y)$$

where $H(X,Y)$ is the joint entropy of the two labels (inferred clustering and ground truth) and $H(X)$ and $H(Y)$ are their individual entropies. Given that most algorithms returns the fraction or probability that each individual belongs to population k, for the purpose of evaluation, we assigned each individual to the population group of the highest likelihood as determined by STRUCTURE. While Spectrum also provided a fraction or probability profile for each individual, the number specifies probability or fraction a person originated from a ancestral haplotype rather than the ancestral population. As a result, arbitrarily assigning each individual by the likelihood fraction will lead to poor clustering results. Consequently, we chose not to evaluate Spectrum by this criterion.

For the three comparative algorithms (STRUCTURE, Spectrum, and Consensus Tree), we also assessed robustness of the method to repeated subsamples. For each pair of individuals (i, j) across five independent samples, we computed the number of samples a_{ij} in which those individuals were grouped in the same cluster and the number b_{ij} in which they were grouped in different clusters. Each method was assigned an overall inconsistency score:

$$\text{Inconsistency} = \sum_{i,j} \frac{\min\left\{1 - \frac{2b_{ij}}{\lfloor(a_{ij}+b_{ij})\rfloor}, 1 - \frac{2a_{ij}}{\lfloor(a_{ij}+b_{ij})\rfloor}\right\}}{\binom{n}{2}}$$

The measure will be zero if clusters are perfectly consistent from run-to-run and approach one for completely inconsistent clustering. We defined the ground truth for HapMap as the four population groups. For the HGDP data, we treated the ground truth as the seven regions of origin rather than the 29 populations, because many population groups are genetically similar and cannot be distinguished with limited numbers of SNPs.

2.2.4 Sensitivity Test

To characterize the relationship between data quantity and accuracy of the inference, we further performed the analysis for a variable number of tree sizes. We ran the consensus-tree method, STRUCTURE, and Spectrum for 4 different data sizes — 10,000, 1,000, 100, and 10 trees (or the corresponding SNPs) — and computed the variation of information and the inconsistency score for each.

2.3 Results

Fig. 2.3 shows the trees inferred by the consensus-tree method on the simulated data and the two real datasets alongside their corresponding true simulated tree or the condensed Shriver and Kittles "gold standard" trees. Fig. 2.3(a) shows the inferred tree produced by the consensus-tree model on the simulated dataset. Based on the numbers of observed bipartitions explained by each model bipartition, the tree reconstruction correctly infers the key divergence events across the 3 populations when compared to Fig. 2.3(b). The method also picks up some additional splits below the division into three subgroups that represent substructure within the defined subgroups. The fractions of mutations assigned to each edge roughly correspond to the number of generations simulated on that edge, although with the edge from the MRCA of all populations to the MRCA of populations one and two assigned slightly fewer mutations and the two edges below that somewhat more mutations than would be proportional to their divergence times.

Fig. 2.3(c) shows the inferred tree from the HapMap dataset. The tree reconstruction infers there to be an initial separation of the YRI (African) sub-population from the others (CEU+JPT+CHB) followed by a subsequent separation of CEU (European) from JPT+CHB (East Asian). When collapsed to the same three populations (African, European, East Asian), the gold standard tree (Fig. 2.3(d)) shows an identical structure. Furthermore, these results are consistent with many independent lines of evidence for the out-of-Africa hypothesis of human origins [54, 102, 117]. The edge weights indicate that a comparable number of generations elapsed between the divergence of African and non-African subgroups and the divergence of Asian from European subgroups, consistent with a single migration of both groups out of Africa long before the two separated from one another.

For the HGDP dataset, the trees differ slightly from run to run, so we arbitrarily provide the first run, Fig. 2.3(e), as a representative. The tree infers the most ancient divergence to be that between Africans and the rest of the population groups, followed by a separation of Oceanian from other non-Africans, a separation of Asian+American from European+Middle Eastern (and a subset of Central South Asian), and then a more recent split of American from Asian. Finally, a small cluster of just two Middle Eastern individuals is inferred to have separated recently from the rest of the Middle Eastern, European, and subset of Central South Asian. The tree is nearly identical to the that derived from Shriver and Kittles for the same population groups (Fig. 2.3(f)). The only notable distinctions are that gold standard tree has no equivalent to the purely Middle Eastern node identified by consensus-tree method; that the gold standard does not distinguish between the divergence times of Oceanian and other non-African populations from the African, while the consensus-tree method predicts a divergence of Oceanian and European/Asian well after the African/non-African split; and that the gold standard groups Central South Asian with East Asians while the consensus-tree method splits Central South Asian groups between European and East Asian subgroups (an interpretation supported by more recent analyses [91]). The results are also consistent with the simpler picture provided by the HapMap data as well as with a general consensus in the field derived from many independent phylogenetic analyses [54, 118]. The relative edge weights provide a qualitatively similar picture to that of the HapMap data regarding relative divergence times of their common subpopulations, although the HGDP data suggests a proportionally longer gap between the divergence of African from non-African subgroups and further divergence between the non-African subgroups.

Fig. 2.4 visualizes the corresponding cluster assignments, as described in Methods, in order to provide a secondary assessment of our method's utility for the simpler sub-problem of subpopulation inference. Note that STRUCTURE and the consensus-tree method assign sequences to clusters while Spectrum assigns each sequence a distribution of ancestral haplotypes, accounting for the very different appearance of the Spectrum output.

The three methods produced essentially equivalent output for the simulated and HapMap data. For the simulated data (Fig. 2.4(a)), all of the methods were able to separate the three population groups. For HapMap (Fig. 2.4(b)), all three methods consistently identified YRI and CEU as distinct subpopulations but failed to separate CHB (Chinese) and JPT (Japanese).

Results were more ambiguous for HGDP (Fig. 2.4(c)). The consensus tree method reliably finds five of the seven populations, usually conflating Middle Eastern and European and failing to recognize Central South Asians, consistent with a similar outcome from He *et al.* [45]. STRUCTURE showed generally greater sensitivity but slightly worse consistency than our method, usually at least approximately finding six of the annotated seven population groups and having difficulty only in identifying Central South Asians as a distinct group. Spectrum showed a pattern similar to STRUCTURE but the individual ancestral profile seemed to be similar in several population subgroups. For example, the African subgroup seemed to have a similar ancestral profile to the European subgroup.

We further quantified the quality of the cluster inference from the consensus-tree method and STRUCTURE by converting the result to the most likely cluster assignment and computing VI scores and inconsistency scores. Fig. 2.5 shows the VI and inconsistency scores of the three algorithms using inputs with different number of trees and SNPs. When examining the variation of information across different data sets, we can see increased accuracy for both STRUCTURE and consensus tree as we increase the number of trees or SNPs. When we compare the inconsistency scores, neither of the algorithms showed a clear trend with increasing numbers of trees or SNPs. When the number of trees or SNPs is large, however, the consensus-tree method typically becomes more consistent than STRUCTURE.

We also measured the runtimes of the algorithms using 10, 100, 1,000, and 10,000 trees or the corresponding SNPs (Fig. 2.6). In all cases, the consensus-tree method consistently ran faster than both STRUCTURE and Spectrum, which both use similar Gibbs sampling approaches.

Fig. 2.7 shows the consensus trees constructed using different sizes of dataset subsampled from the simulated data. From the figure, we can see that the trees never infer substructure that cuts across the true groups, but that as the data set size increases, the method yields increasingly refined tree structures. This observation is what we would expect for the chosen MDL approach. The method identifies the separation of populations one and two with 100 trees but not with

10, and can discriminate substructure within the individual populations when provided 10,000 trees but not 1,000 or fewer. The number of mutations assigned to each edge increases as we increased the number of observed trees, but the fraction of all mutations assigned to each edge remains nearly constant with increasing data set size.

2.4 Discussion

While population substructure inference is only one facet of the problem solved by the consensus-tree method, it nonetheless provides for a convenient partial validation. Comparison with leading population substructure algorithms shows that the consensus-tree method provides very good performance on the substructure problem. The consensus-tree approach shows equal or slightly superior VI scores relative to STRUCTURE on both simulated and HapMap data while showing slightly worse VI scores in HGDP. The consensus-tree method is also quite competitive on run time with these alternatives, although other substructure methods that were not amenable to a direct comparison, such as mStruct [101], can yield substantially superior run times for closely related analyses. The consensus-tree method also shows an ability to automatically adjust to varying amounts of data while avoiding over-fitting, as demonstrated by the consistency scores, as would be expected for the chosen MDL approach.

One key advantage of the consensus-tree approach is that by treating substructure inference as a phylogenetic rather than a clustering problem, it can provide additional information about relationships between subgroups. Such information may be helpful in better completing our picture of how modern human populations arose and may provide information of use in correcting for population stratification during association testing. Because we are aware of no comparable methods for this problem, we must resort to validation on simulated data and by comparison to our best current models of true human population histories to evaluate its performance on the full population history inference problem. The consensus-tree method correctly infers tree structures from the simulated data using as few as 100 trees. Furthermore, application to HapMap and

HGDP data also shows that the method produces a portrait of human evolution consistent with our best current understanding. The basic qualitative model of human population history that emerges is further consistent between the two independent datasets, despite different individuals, populations represented, and markers selected.

The consensus-tree model also provides information about how many mutations one can attribute to each edge of a given tree. These edge lengths can be interpreted to approximately correspond to divergence times along different edges of the trees. In particular, provided one assumes that mutations accumulate at a constant rate across human lineages then one would expect that mutations would accumulate in any subpopulation at a rate proportional to the size of that subpopulation and to become fixed with a probability inversely proportional to the size of that subpopulation. To a first approximation, then, edge weight normalized by the total number of mutations used in the model should be approximately proportional to the time elapsed along a given edge independent of the size of the population represented or the number of input trees. The quantitative results do approximately fit this expectation for the simulated data. There is, however, some apparent bias towards lengthening the edges from the MRCA of subpopulations one and two to the MRCAs of the two individual subpopulations and shorting the edge from their MRCA to that of all three subpopulations. This observation may reflect imprecision in the rough approximation that edge length should be proportional to elapsed time. Alternatively, it may derive from misattribution of some SNPs formed within the subpopulations to the edges leading to those subpopulations. While the method can provide estimates of relative times elapsed along edges, it does not have sufficient information to convert these numbers of mutations into absolute elapsed time. In principle, one could make inferences of absolute elapsed time along tree edges given more detailed population genetics models and a complete, unbiased set of variant markers from which to construct phylogenies. Similarly, having some absolute time assigned to even a single edge would allow one to estimate absolute times along all other edges in a tree.

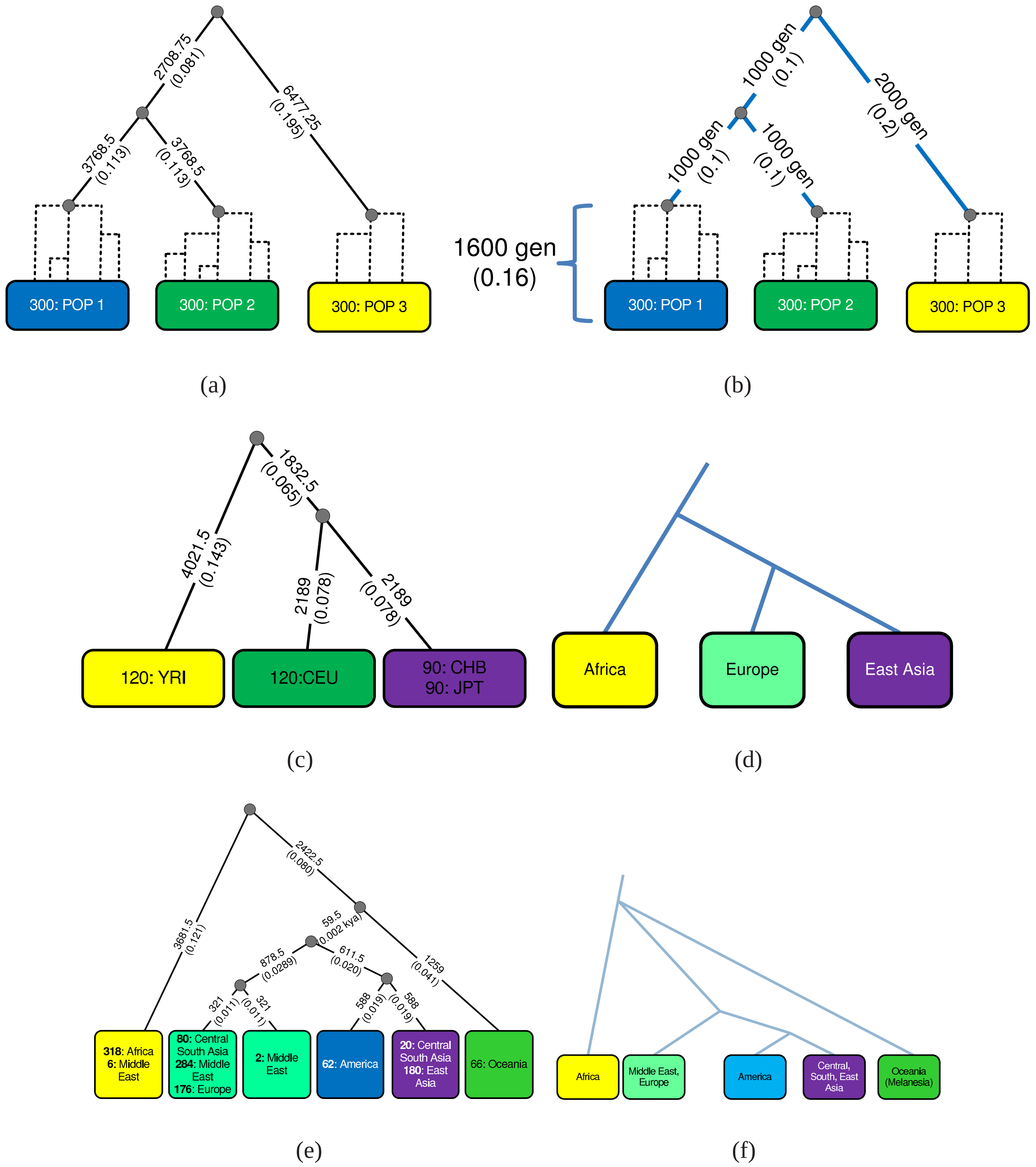
Given that edge weights can be expected to be approximately proportional to elapsed time, we can use those derived on the real data to gain some additional insight into how the inferred human subgroups may be related. The two data sets yield qualitatively similar models supporting a single emergence of an Asian/European ancestral group from Africa followed by divergence of that ancestral subgroup into Asian and European subgroups. There are, however, some notable quantitative differences between relative divergence times of various subgroups between the two data sets. In particular, the HGDP data suggest a proportionally longer gap between separation of African from non-African and separation of Asian from European. For example, if we assume that the African/non-African divergence occurred 60 thousand years ago (60 kya), around the middle of the range of recent estimates [117], then the HapMap data would place the Asian/European divergence at 32.7 kya while the HGDP would lead to an estimate of 19.5 kya. This observation could reflect an inherent bias in the edge length estimates, as noted for the simulated data, or biases intrinsic to the data sets. Several previous studies estimating divergence times have found that inferences can be sensitive to the choice of population groups, the specific genetic regions examined, or the particular individuals in those populations [51, 92, 132].

While the results show that the consensus-tree method is capable of making robust but sensitive inferences of population structure as well as tree structure, the consensus-tree method does nonetheless have some significant limitations. One such limitation is runtime; while the consensus-tree method is superior in this regard to STRUCTURE and Spectrum, its runtime is still considerable and far worse other algorithms such as mStruct and EIGENSOFT. Although this compute time is still a trivial cost compared to the time required to collect and genotype the data, it may nonetheless be an inconvenience to users. Furthermore, it prevents us from processing the full HapMap or HGDP data sets in a single run, as opposed to the subsamples done in the present work, likely preventing discovery of finer resolutions of population substructure. This high run-time is largely due to the many calls the consensus-tree method must make to the DMST algorithm to repeatedly evaluate the MDL objective function and may be addressed in future work by more sophisticated optimization methods to reduce the number of function evaluations or by introducing a more highly optimized subroutine for evaluating MDL costs. In addition, the computations should be easily amenable to parallelization.

Another limitation, noted above, is that the current version of the consensus tree method does not handle admixture in population groups as do competing methods. We would expect admixture to appear during inference of bipartitions as the discovery of sets of bipartitions that cannot be reconciled with a perfect phylogeny. In principal, then, the core MDL algorithm should function correctly on admixed data but our conversion of the bipartitions into a tree would need to be replaced with a method for inferring a phylogenetic network rather than a tree, similar to methods for inferring ancestral recombination graphs from haplotype data [39]. New methods will likewise be required to perform admixture mapping of individual admixed genomes to label them by population group. These additions are important goals for future work and will help to determine whether this novel approach, whatever its initial promise, proves a competitive method in practice for detailed substructure analysis.

2.5 Conclusion

We have presented a novel method for simultaneously inferring population ancestries and identifying population subgroups. The method builds on the general concept of a "consensus tree" summarizing the output of many independent sources of information, using a novel MDL realization of the consensus tree concept to allow it to make robust inferences across large numbers of measurements, each individually minimally informative. It incidentally provides a *de novo* inference of population subgroups comparable in quality to that provided by leading methods. The consensus-tree method also provides edge length estimates that can roughly be interpreted as relative times between divergence events, although there appears to be some biases in these estimates. As we will demonstrate in the next chapter, it will be possible to translate these relative times into estimates of absolute elapsed times using a coalescent-based population genetic models. The MDL approach also allows our method to automatically adapt to larger data sets, producing more detailed inferences as the data to support them becomes available.



(a)

(b)

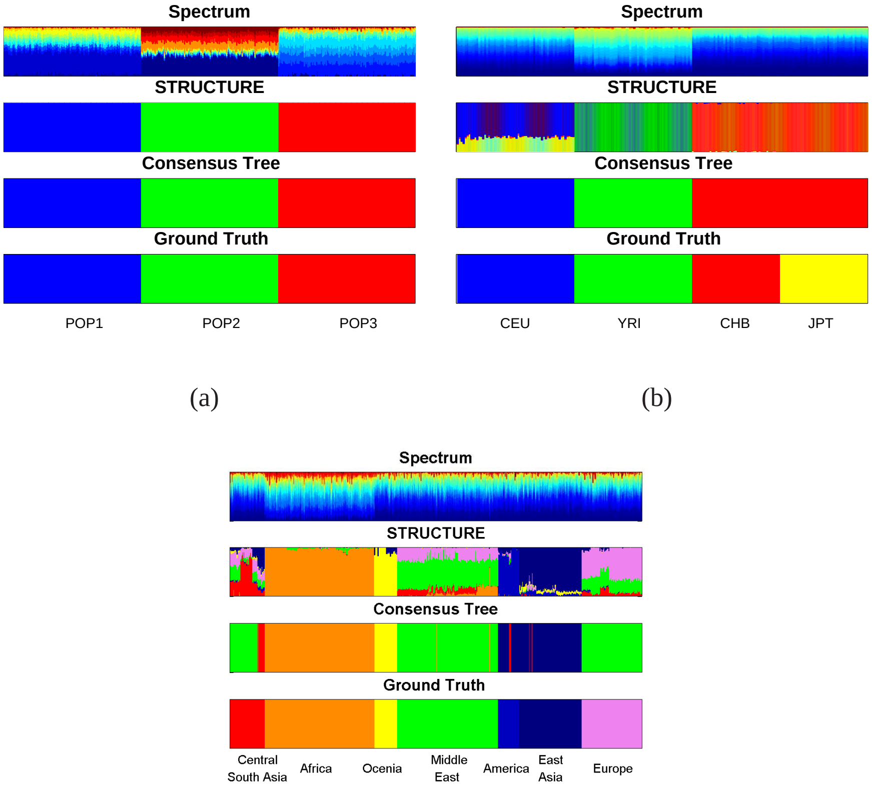
(c)

(d)

(e)

(f)

Figure 2.3: Inferred consensus trees. Node labels show numbers of haplotypes belonging to each known population. Edges in inferred trees are labeled by the number of splits assigned to each and, in parentheses, the fraction of all splits assigned to each. For the simulated gold standard tree, edges are labeled by a number of generations and, in parentheses, the expected number of substitutions per site occuring on the corresponding edge in generating the data. (a) Consensus tree obtained from simulated data. (b) Gold standard for the simulated data. (c) Consensus tree obtained from the HapMap dataset. (d) Trimmed and condensed tree from [102]. (e) Consensus tree obtained from the HGDP dataset. (f) Trimmed and condensed tree from [102].

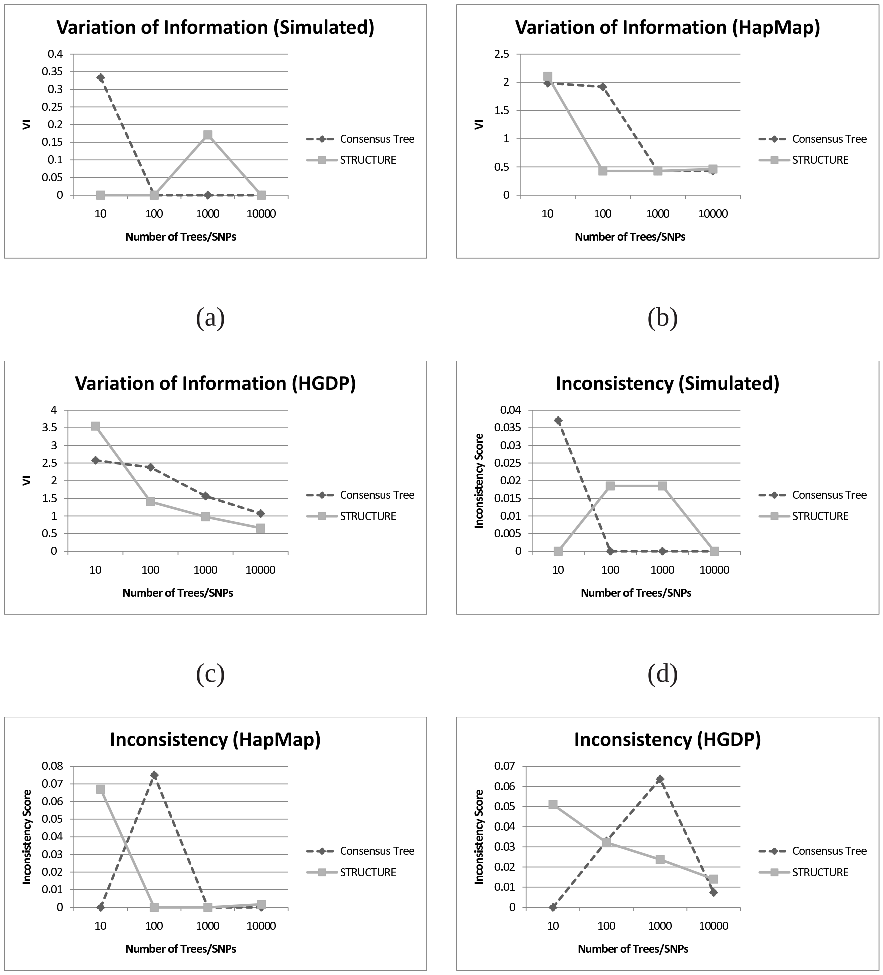


(a)

(b)

(c)

Figure 2.4: Inferred population structures. Each colored vertical line shows the assigned population(s) for a single sequence for one method. From top to bottom: Spectrum (with colors representing fractional assignments to distinct ancestral haplotypes), STRUCTURE (with colors representing probabilities of assignment to distinct clusters), consensus-tree (with colors showing assignments to single clusters), and ground truth (with colors representing assignments to true clusters). (a): Inferred population structure from a single trial of 10,000 trees from simulated data. (b): Inferred population structures from a single trial of 10,000 trees from HapMap Phase II dataset. (c): Inferred population structures from one trial of 10,000 trees.

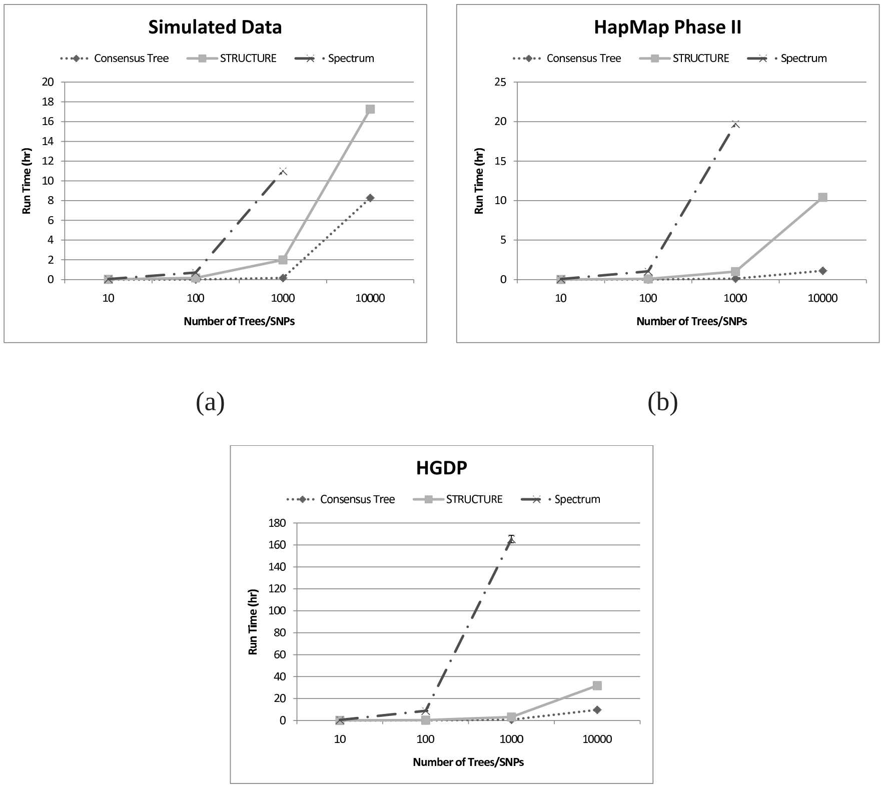


(a) (b)

(c) (d)

(e) (f)

Figure 2.5: Variation of information (VI) and inconsistency scores. Lower VI reflects higher accuracy in identifying known population structure. Lower inconsistency reflects greater reproducibility between independent samples.



(a)

(b)

(c)

Figure 2.6: Average runtime of the algorithms on different data sets and different data set sizes.

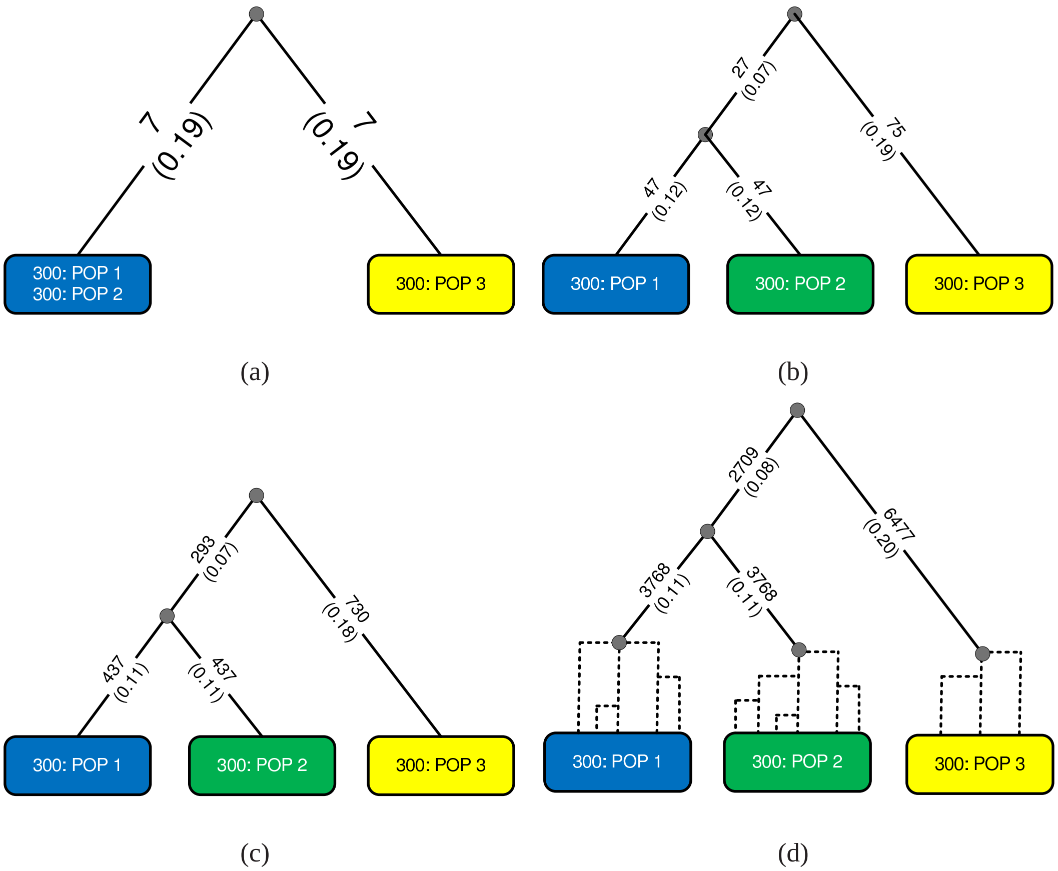


Figure 2.7: Consensus trees produced using varying numbers of input trees. Node labels show numbers of haplotypes belonging to each simulated population. Edges are labeled by the number of splits assigned to each and, in parentheses, the fraction of all splits assigned to each. From left to right: Consensus Tree from (a) 10, (b) 100, (c) 1000, and (d) 10,000 observed trees.

Chapter 3

Coalescent-based Method for Learning Parameters of Admixture Events from Large-Scale Genetic Variation Datasets[1]

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 [80]. 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[57]. In statistical genetics, studying ad-

[1] This chapter was developed from material published in [120] and under review in [121]

mixture and population structure can help in identifying and correcting for confounding effects of population structure in disease association tests [37]. Studying admixture can also help in understanding the acquisition of disease-resistance alleles [27].

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 [13, 34]. 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 [84, 85, 98], 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 [77], and others different parameters, such as effective population size [61]. 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 [16]. 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.

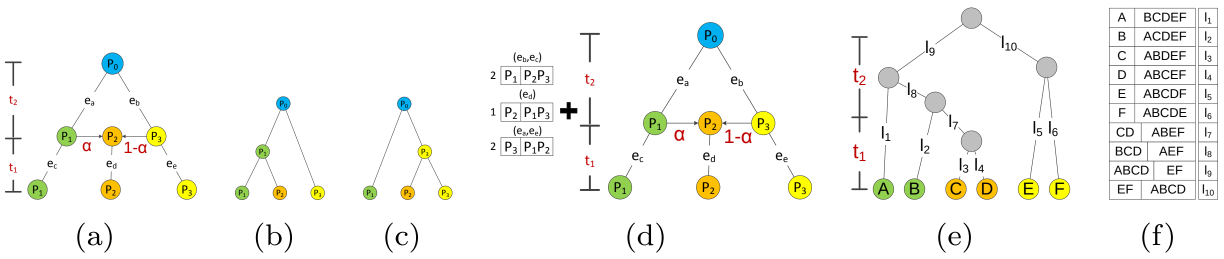
This issue has been previously addressed by methods utilizing coalescent theory, [17, 126]. a probabilistic model of ancestral relationships that can be used to efficiently sample among possible evolutionary histories of a set of individuals in a population. *MEAdmix* [126], for instance, uses coalescent 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 chapter, we describe a novel approach to reconstructing parameters of admixture events that addresses several limitations of the prior art. This 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, $t_1 \in \mathcal{R}^+$, $t_2 \in \mathcal{R}^+$, $\alpha \in [0, 1]$. 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 [17, 77, 126], 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.



A	BCDEF	l_1
B	ACDEF	l_2
C	ABDEF	l_3
D	ABCEF	l_4
E	ABCDF	l_5
F	ABCDE	l_6
CD	ABEF	l_7
BCD	AEF	l_8
ABCD	EF	l_9
EF	ABCD	l_{10}

Figure 3.1: 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 actually parameters. (e) Genealogy generated from a parameter of t_1, t_2, and α showing the specific relationship and branch length between every sample in the data. Here, AB is in P_1, CD is in P_2, and EF is in P_3. (f) The corresponding bipartition and associated branch length obtained from genealogy in (e).

3.1 Materials and Methods

To learn population history for a dataset, our approach first tries to determine a number of subpopulations (K), potential evolutionary models ($\tilde{G} = (G, V)$)between the subpopulations, and a summary description (H) 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 divided 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 [11, 16, 17, 126], we specifically address the problem of accurately reconstructing parameters of a single historical admixture event. As shown in Fig. 3.1(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 such assumption may not necessarily hold for all analyses, it is a reasonable assumption in some cases such as human population since non-African populations share about the same effective population size [68, 113]. Furthermore, as our analysis has shown, such assumption can still give accurate results when effective population size does not vary significantly. 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. 3.1(b) while other regions would have genealogies derived from Fig. 3.1(c). 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 [119] 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. 3.1(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. 3.1(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. 3.1(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. 3.1(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 described in Chapter 2 on learning population histories from non-admixed variation data [119] is conceptually based on the idea of consensus trees [73], 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. 3.1(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 [38].

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, $B^M = \{b_1^M, b_2^M, ...b_r^M\}$, and a set of weight values, $W = \{w_0, w_1, w_2, ..., w_r\}$. 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, $w_1, ..., w_r$, 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 [119] 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 bipartition is a good assignment for the observed bipartition, the bipartition is then assigned to a empty bipartition and attribute to the weight w_0. By matching the observed bipartition, we indirectly estimates 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, the consensus-tree based algorithm should first identify that there are three subpopulations ($K = 3$) in the data. Second, the algorithm should output an inferred evolutionary model $\tilde{G} = (G, V)$, shown in Figure 3.1(d) and characterized by the 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 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: Under the two-parental one admixed population scenario, learning the directed graph $G = (V, E)$ and its label function from the outputs of consensus tree algorithm would be trivial by associating the lowest weighted model bipartition representing one of the three populations to be the admixed population. This would leave us with just Θ to infer. 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 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 generated by removing a single edge of the genealogy. For each bipartition, if $f(b)$ is a function that computes the optimal index assignment of a bipartition b to the model bipartition set using the conditional entropy function described in our prior method [119] 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:

$$P(W|\Theta) = \prod_{i=0}^{3} \int_{l_{b_i^M}=0}^{\infty} \sum_{\mathcal{G}} P(w_i|\Theta, l_{b_i^M}) P(l_{b_i^M}|\mathcal{G}, \Theta) P(\mathcal{G}|\Theta) dl_{b_i^M} \tag{3.1}$$

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

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:

$$P(W|\Theta) \quad = \prod_{i=0}^{3} \sum_{\mathcal{G}} P(w_i|l_{b_i^M}, \theta) P(\mathcal{G}|\Theta) \tag{3.2}$$

As an example, suppose we have an output from the consensus algorithm shown in Figure 3.1(d). If we have a particular parameter value we want to evaluate the likelihood function, we would enumerate through all possible genealogy consistent with the specified t_1, t_2, α, and θ. Suppose a genealogy in Figure 3.1(e) was one possible genealogy being enumerated. We would compute the score by converting the genealogy into a set of bipartitions as shown in Figure 3.1(f) and subsequently compute the optimal assignment of each bipartition to the most related model bipartition using the same scoring mechanism in [119]. Given the optimal assignment of each bipartition, we can then compute the expected branch length $l_{b_i^M}$ associated with model bipartition 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 an 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 compute the expected number of mutations associated with each model bipartition and null bipartition and subsequently the probability $P(w_i|l_{b_i^M})$. The idea behind the model is that a correct parameter set that describes the history should give us the closest match of observed weights associated with each model bipartition and thus the highest 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 [17] and [77] 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, $L_1, L_2, ..., L_{k-1}$, 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 $L_{tot,G}$.

$$L_{tot,\mathcal{G}} = \sum_{i=0}^{n} L_{tot,G_i} = n\left(\frac{1}{n}\sum_{i=0}^{n} L_{tot,G_i}\right) \tag{3.3}$$

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:

$$L_{tot,\mathcal{G}} \approx nE(L_{tot,G}) = n\left(\sum_{j=2}^{k} jE(L_j)\right) \tag{3.4}$$

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, 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 as we observed the improvements in error tapers off quickly as we increased the number of genealogies in our experiments. 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:

$$P(W|\Theta) = \prod_{i=0}^{3} \sum_{\hat{\mathcal{G}}} P(w_i | l_{b_i^M}, \theta) P(\hat{\mathcal{G}}|\Theta) \tag{3.5}$$

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 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 $\mathcal{G}$ 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:

$$\begin{aligned} & P(X_o|W) \propto P(W|X_o) \\ & = \left(\prod_{i=0}^{3} P(w_i|l_{b_i^M})P(l_{b_i^M}|x^o_{\hat{\mathcal{G}}})\right) P(x^o_{\hat{\mathcal{G}}}|x^o_{t_1}, x^o_{t_2}, x^o_{\alpha}) \end{aligned} \tag{3.6}$$

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

$$\begin{aligned} Q(X_n|X_o) &= P\left(x^n_{t_1}|\mu^o_{t_1}, \sigma_{t_1}\right) P\left(x^n_{t_2}|\mu^o_{t_2}, \sigma_{t_2}\right) \\ & \quad \times P\left(x^n_{\alpha}|\mu^o_{\alpha}, \sigma_{\alpha}\right) P\left(\hat{\mathcal{G}}|x^n_{t_1}, x^n_{t_2}, x^n_{\alpha}\right) \end{aligned} \tag{3.7}$$

yielding a Metropolis-Hastings acceptance ratio r of:

$$r = \frac{\left(\prod_{i=0}^{3} P(w_i|l_{b_i^M})P\left(l_{b_i^M}|x^n_{\hat{\mathcal{G}}}\right)\right)}{\left(\prod_{i=0}^{3} P(w_i|l_{b_i^M})P\left(l_{b_i^M}|x^o_{\hat{\mathcal{G}}}\right)\right)} \tag{3.8}$$

3.2 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 [48] 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 [43, 62]. 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* [126]. *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 [17, 77, 126], 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 [11, 77], 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 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 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 datasets 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 and additional four groups of 45 simulated datasets, we evaluated the performance of the algorithm by computing 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 [12], 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 [6]. 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

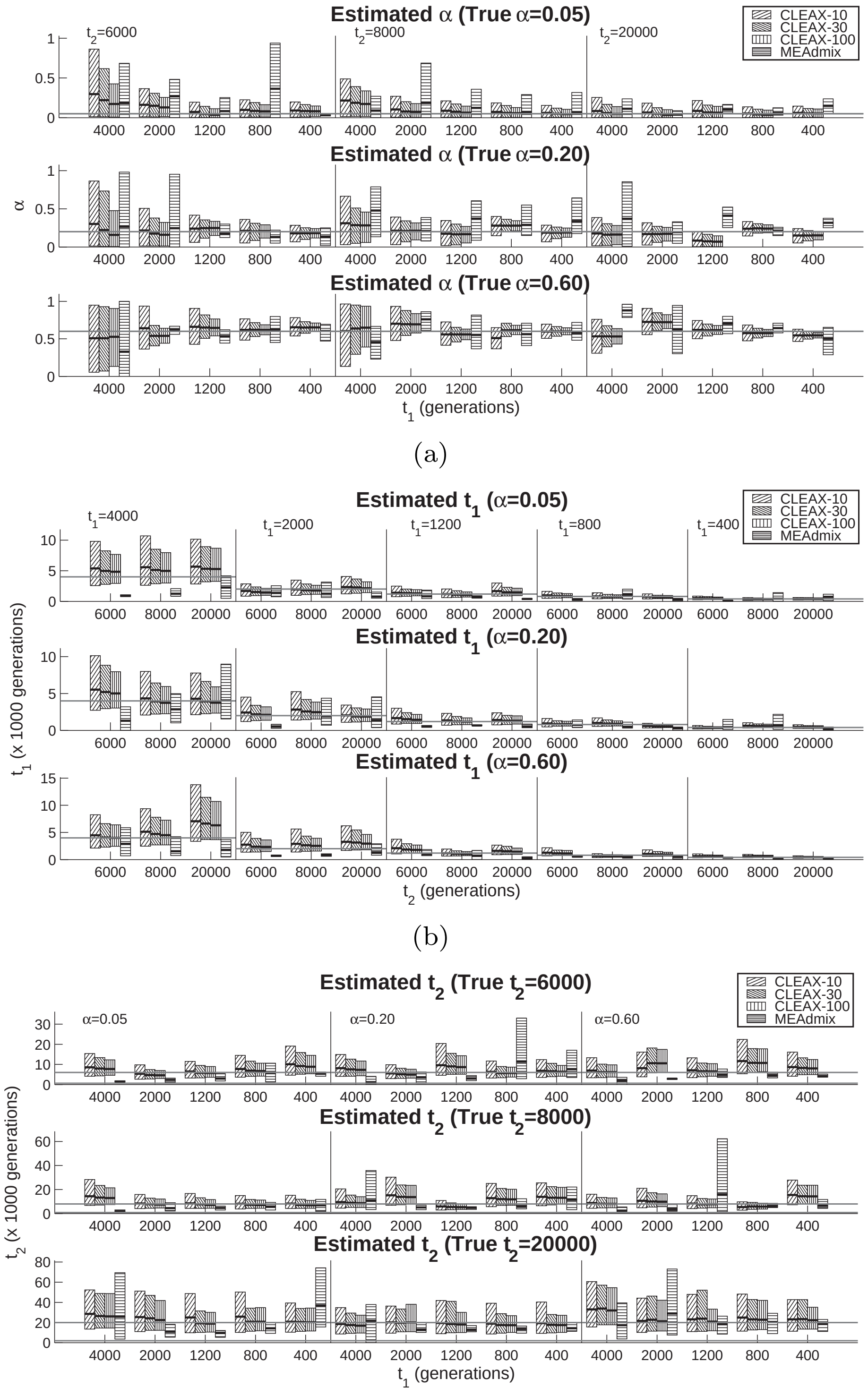
China and Japan. The remaining 181 individuals from America consist mainly of admixed individuals from Mexico, Puerto Rico, and Columbia. Similar to the bovine dataset, we filtered the dataset until only one admixed population and two parental populations remained by removing the 246 individuals having 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) [4] 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 restarted each chain 10 times for bovine and HapMap datasets and 50 times for 1000 Genome data to ensure that it did not become stuck in a local optimum due to poor selection of the starting point.

3.3 Results

Coalescent Simulated Data: Figure 3.2(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

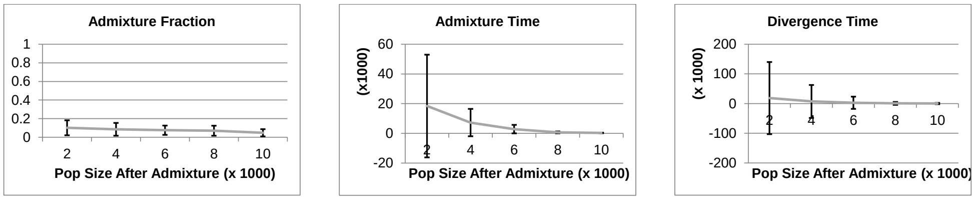


(a)

(b)

(c)

Figure 3.2: 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 in 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



(a) (b) (c)

Figure 3.3: 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.

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 3.2(b)) and t_2 (Figure 3.2(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 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.

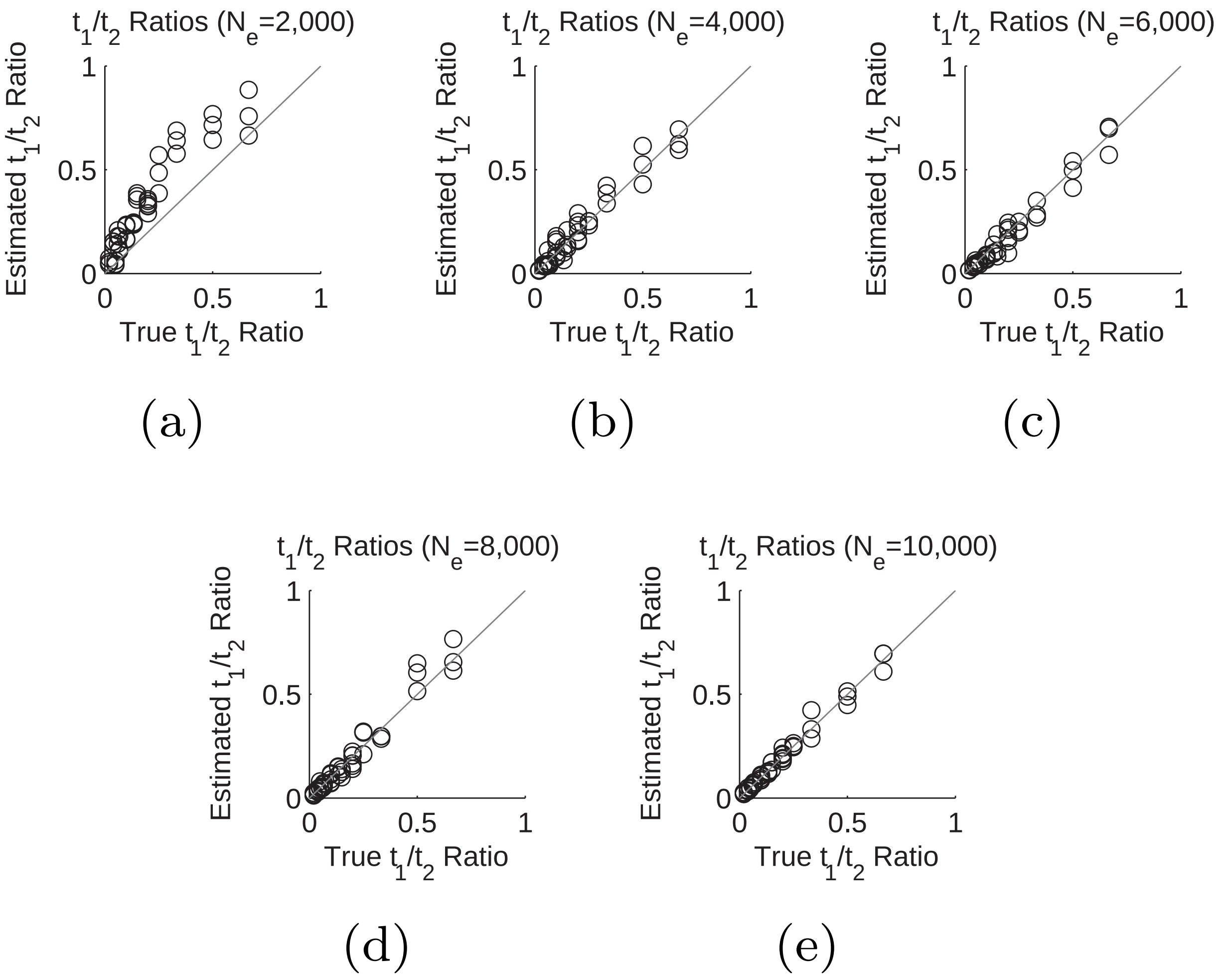


Figure 3.4: Plot of estimated t_1/t_2 ratio against true t_1/t_2 ratio from datasets when the effective population size changes from 10000 to 2000, 4000, 6000, 8000, and 10000 (a-e).

Aggregate quantitative performance is shown in Table 3.1, 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 t_1 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, and instead identifying 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 3.1, 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 3.1 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.

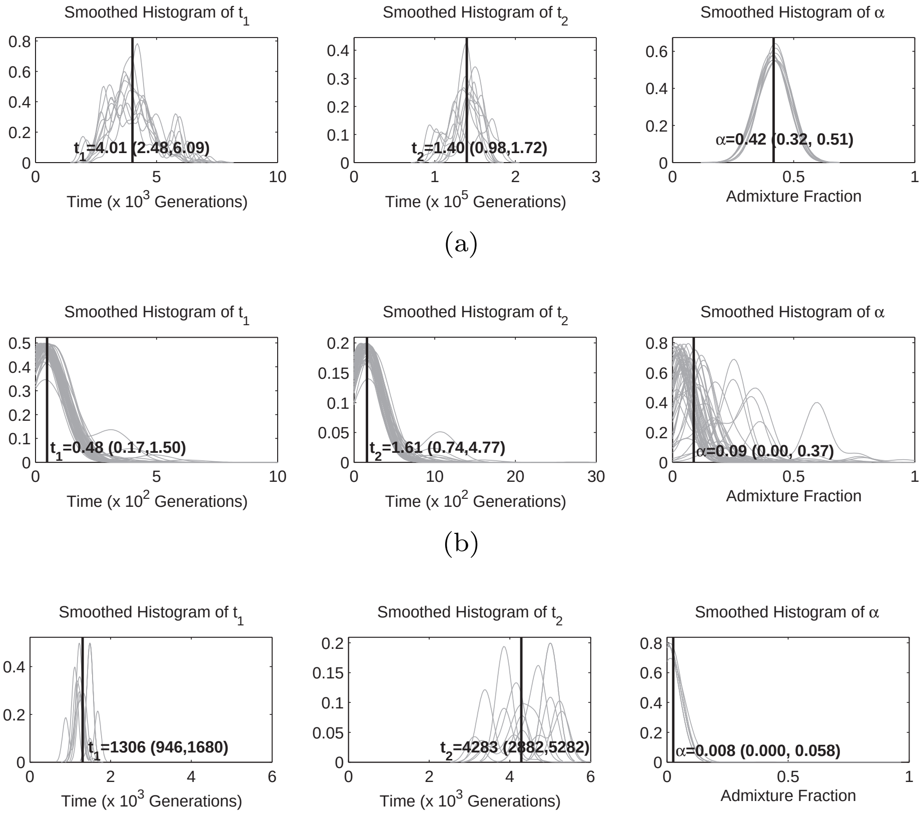
To understand the effect of varying effective population size on the performance of the algorithm, we evaluated our method on datasets with reduced effective population size after admixture events. Figure 3.3 shows the average absolute difference between the estimated and the true parameter values across different reduced effective population size 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 3.4 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 3.5(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 of the ten independent runs of the Markov chain. All ten runs of the chain on the bovine data yielded consistent probability distribution may suggest the confidence of the estimated parameter values. 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 [12]. 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 [12]. The mean estimate 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 [12] 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 [58, 62]. Using an estimated effective population size of 107 [12], 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} [62]. Inaccuracy in the rate might also be due to ascertainment bias or the incomplete detection of the mutations at the sequencing phase.

Figure 3.5(b) shows the distribution of CLEAX estimates for the 1,000 Genomes Project data, interpreting the American group, consisting of individuals from Mexico and Puerto Rico, as admixed from the Asian and European groups. Given the samples generated from the ten chains, CLEAX inferred an average of 9% admixture from the Asian group and 89% from the European group. Admixture fraction α from different runs of the chain are most concentrated around 0.05, although 6 out of 50 chains, likely stuck in local optima, have values of approximately 0.3 for 5 chains and 0.6 for another. While the mean estimate is slightly lower than other literatures [66, 90, 112], the 95% confidence interval is consistent with other estimates. The mean estimate of the admixture time t_1 was 48 generations with a 95% confidence interval between 17 to 150 generations. Assuming 20 years per generations, this would translate to approximately 960 years ago with 95% confidence interval ranging from 340 years ago to 3,000 years ago, a number slightly higher than the 200-500 year ago estimate by Tang *et al.* [112] but within a reasonable range. The mean divergence time t_2 was estimated to be 161 generations ago with 95% confidence interval between 74 and 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 ranging from 1,500 years to 9,500 years ago, a consistent range with Garrigan *et al.* [35] but a more recent estimate than Zhivotvosky *et al.*[132].

Figure 3.5(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 [42], the implied mutation rate would be 2.16×10^{-9} per site per generation, similar to prior estimates [58, 62].



(a)

(b)

(c)

Figure 3.5: Probability density of the estimated parameter values, t_1, t_2, and α (left to right) for the bovine, HapMap, and 1000 Genome datasets. Vertical line represents the mean of the parameter value with 95% confidence interval printed in parenthesis. (a) 10 MCMC chains ran on 76 cattle from the bovine dataset on each of the 10 independent runs [12]. (b) 50 MCMC chains ran on 1092 individuals from 1000 genome data [6]. (c) 10 MCMC chains ran on 210 individuals from HapMap Phase II data [4].

3.4 Discussion

In this chapter, 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 pro-

Table 3.1: 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]

vides 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 3.1, 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 providing increased confidence 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 [12, 40, 132]. 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 by [132] (62-133kya) and the HMM estimation by [61] (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. Estimates of the divergence time between Asian and European between 23.0 kya and 33.6 kya for HapMap are similar to estimates by Gutenkunst *et al.* [40].

Estimates of the divergence time between Asian and Europe from the 1000 Genome dataset is also similar to the estimate from HapMap albeit with slightly more recent range but consistent with estimates from Garrigan (7-13kya) *et al.* [35]. While the mean average of the admixture time of the American group was somewhat higher than expected (980 years), the lower bound of the estimate of 340 years ago is a reasonable estimate of the admixture time. The admixture fraction estimate for the American group is also consistent with existing literature [66, 112].

Similarly, using the bovine dataset, estimates of divergence time and admixture fraction were also consistent with the general consensus [12]. 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 to a few hundred cattle in recent years [12, 59], 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.

Analysis on the effect of varying effective size on performance suggests that the estimation of times of divergence and admixture is sensitive to changes in effective population size but that such changes have modest effect on the admixture fraction estimation. This observation suggests that estimates of the admixture fraction can be more reliably trusted than estimates of divergence and admixture time when one suspects effective population has changed drastically over time. For time estimation, estimates were within an order of magnitude when the change in effective population size was less than or equal to 40%, suggesting estimation could still be trusted if changes in effective population size are modest. Furthermore, estimates of the ratio between t_1 and t_2 seemed 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 if we can anchor at least one time point using external data sources or prior knowledge even if the population size changes significantly.

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 [17, 77], 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.

Chapter 4

Coalescent-based Method for Joint Estimation of Population History, Time, and Admixture[1]

Despite intense study and interest, learning the history by which modern human population groups arose from our common ancestors remains a difficult question. The topic is not only a central issue in basic research into human genetics; it is also an important practical question in medical genetics for better separating functionally significant genetic variations from spurious associations with phenotype due to population substructure [53, 117]. While large data sets suitable for inferences of population-level evolution have proliferated (c.f., [5, 6, 50, 76]), there is as yet no fully automated method to analyze the available data and reconstruct the sequences of events by which ancestral populations have arisen and intermixed to produce the modern diversity of human population groups.

Although there are abundant methods for learning various features of population origin and evolution, these methods can generally be grouped into two broad categories. First, one can infer evolutionary history by learning phylogenetic trees or networks consistent with a set of data

[1] This chapter was developed from material submitted in [122]

using variants of a wide variety of phylogenetic inference algorithms (c.f., [31]). Such methods allow one to build detailed models of evolutionary relationships between individuals in a dataset and the sequences of mutations that might distinguish them. Such approaches, however, have the disadvantage of long computation times that limits their usefulness on large genomic datasets or large sample pools, especially when attempting to reconstruct phylogenetic histories from data presumed to have undergone extensive recombination. The major alternative is to study history at the population level, greatly lowering computational time, but typically requiring considerable preprocessing or manual intervention to identify population subgroups before inferring parameters describing relationships among those subgroups [9, 16, 17, 63, 73, 125]. Once population labels are defined, a simple approach commonly applied is to compute the genetic distances between the populations and build a model of their phylogenetic relationships using distance-based tree reconstruction algorithms, such as UPGMA or neighbor joining [9, 73], with additional parameters such as divergence times inferred subsequently [15, 29, 103] by custom inferences designed for analyzing specific subsets of parameters or presumed population-level events.

A particularly challenging event type is admixture, an evolutionary event in which two or more geographically or culturally separated populations came into contact and establish a new population in which individuals share a mixture of genetic information from multiple parental populations. This process of admixture is believed to be common in human populations, where movements have repeatedly brought together populations that were historically separated. This can be seen, for instance, in the United States where many African Americans contain varying amounts of ancestry from Europe and Africa [80]. To deal with the problem of interpreting evolution in the presence of admixture, several coalescence-based methods have been proposed. One such approach is the *lea* method of Chikhi *et al.* [17], which generates genealogies sampled from backward simulation of coalescence trees of lineages in order to estimate likelihood scores of admixture events for Markov chain Monte Carlo (MCMC) sampling. A similar method known as *MEAdmix* [126], which also uses the coalescent theory to simulate mutation patterns under various admixture scenarios and identify those most consistent with observed variation data. Such methods were significant advances in learning evolutionary history in the presence of admixture. Nonetheless, they are typically limited by the assumption of a fixed evolutionary model known in advance, e.g., a two-parental and one admixture population scenario, and generally to small fixed numbers of subpopulations. In addition, these methods have difficulty scaling to large data sets due to long computation times.

With large whole-genome variation data sets now becoming the norm for the field, new methods are needed that can productively use such data to build more accurate and detailed models of population dynamics. Other methods exist for drawing inferences from large-scale variation data, but each brings its own limitations. Principal component analysis (PCA) has proven a powerful tool for visually capturing relationships within complex population mixtures [13, 34], but typically requires considerable manual intervention and interpretation to translate these visualizations into concrete models of the population history. It further provides only limited quantitative data on relationships between admixed populations (e.g., estimates of admixture fractions but not 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 [84, 85, 98], which provides complementary information with important uses in admixture mapping, but similarly offers little direct insight into the history by which these admixtures occurred.

Here, we describe a generalized Consensus-tree based Likelihood Estimation for AdmiXture (gCLEAX), intended to address the gap in methods for automated inference of population histories for arbitrary number of populations from genome-scale variation data. This method is intended to learn population histories from genetic variation data in a more generalized framework than prior models allowing for automated identification of subpopulations, reconstruction of the history of divergence and admixture events by which those subpopulations likely emerged, and inference of quantitative parameters describing timing and proportions of admixture for these events. The method is based in part on a method described in Chapter 3 for learning parameters of admixture events on two parental and one admixture population scenario [121] but generalizes that approach to learn population history models for arbitrary numbers of populations. This method makes it possible for the first time to perform fully automated inferences of population histories in the presence of admixture on large-scale genomic datasets. In addition, it has the advantage 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 phylogenetic history, allowing it to exploit dependencies among these features of the population history that are inaccessible to prior approaches.

4.1 Material and Methods

As described in Chapter 3, rather than inferring parameters directly from the molecular data [17, 77, 125], gCLEAX first learns a set of summary descriptions of the overall population history from the molecular data that allow more efficient processing of the computationally costly Monte Carlo sampling steps. Once the set of summary descriptions is obtained, it uses a coalescent-based inference model on the summary descriptions to evaluate possible population histories and learn most likely divergence times and admixture fractions. Our method is based on our prior work on learning parameters of a specific three-population admixture scenario [121]. The present work generalizes that approach to infer population models with or without admixture for groups of, in principle, arbitrarily many populations. Here, we briefly describe the details of previous three-population algorithm, with the methods focused primarily on the generalization in the present work.

Although the present work is focused on the general scenario, the three-population scenario consisting of three hypothetical populations P_1, P_2, and P_3 at the current time, is nonetheless useful for illustrating the model. This scenario is shown in Fig. 4.1. At time t_1, population P_2 was formed from a mixture of an α fraction of individuals from parental population P_1 and a $1-\alpha$

fraction of individuals coming from parental population P_3. Further in the past, at time t_2, the two parental populations P_1 and P_3 themselves separated from a common ancestral population. These events define a network of ancestral relationships between the observed populations and the ancestral populations from which they derive.

Given the sequence data derived from the prior example, generalized gCLEAX will first learn that there are three subpopulations in the example dataset using an algorithm developed in previous work [119], which identifies a well-supported set of phylogenetic "splits" defining bipartitions of the population into robustly distinguishable subpopulation groups. At the same time, that algorithm will also identify a set of edges representing the evolutionary history of the population. Given the prior example, the algorithm would identify the edges that represent the separation of population 1 from the other populations (edges e_c and e_b in Fig. 4.1), the separation of population 2 (e_d in Fig. 4.1), and the separation of population 3 (e_e and e_a in Fig. 4.1). In addition to the edges, the algorithm will also infer a number of mutations that have likely occurred along each identified edge. These inferences of meaningful bipartitions, weighted by inferred numbers of mutations, provide a concise summary of the complete variation data set that will be used by our algorithm to estimate the posterior probability distribution of the evolutionary model, event times, and admixture proportions that best describe the data.

4.1.1 Learning Summary Descriptions

As described in detail in Chapter 2, we developed an algorithm for identifying subpopulations and their population-level evolutionary tree from single nucleotide polymorphism (SNP) datasets. The algorithm produces a set of well-supported model bipartitions (i.e., tree edges), $B^M = \{b_1^M, b_2^M, ...b_r^M\}$, and a set of weight values, $W' = \{w_1, w_2, ..., w_r\}$, associated with the model bipartitions, as well as a weight w_0 that represents the number of observed bipartitions that are best explained by none of the model bipartitions. The weights of the bipartitions approximate the numbers of mutations that most likely occurred along the individual branches of the

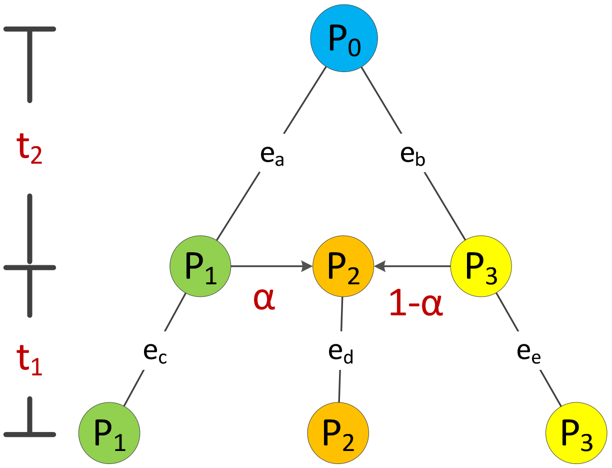


Figure 4.1: Example of a history of two parental populations (P_1 and P_3) and an admixed population (P_2). Ancestral population P_0 diverged at t_2 to form P_1 and P_3, followed by an admixture event at t_1 to form P_2.

population history to which the bipartitions correspond. This set of model bipartitions and its associated weights are then used to reconstruct the evolutionary model.

4.1.2 Admixture Model

To learn the population history from the dataset, we will assume that all populations derive from a single ancestral population. The population is presumed to evolve by a series of discrete events, each either a divergence event in which an ancestral population splits into two subpopulations or an admixture event in which two populations contribute to the formation of a third admixed population. If we have k populations at the present time, then there must have been $k-1$ evolutionary events going backward in time until all populations merge into a single ancestral population. Hence, we would have $k-1$ time parameters ($t_1, ..., t_{k-1}$) to learn. For each admixture event, another admixture parameter (α_i) describing the fraction of ancestry derived from each ancestral population would also have to be learned. In addition to time and admixture parameters, the model is characterized by the topology of the network of population events, $\mathcal{M}$; the mutation rate, μ; and the effective population size (N_{ij}) of each population i at each time t_j

for $j = \{1, 2, ..., k-1\}$. As in Chapter 3, we will assume that the effective population size is constant and identical in each population at each time point (e.g., $N_{ij} = N$). Under this assumption, the free parameters we must learn are $\mathcal{M}$, $\mathcal{T} = (t_1, t_2, ..., t_{k-1})$, $\mathcal{A} = (\alpha_1, ..., \alpha_{k-2})$, and $\theta = N\mu$.

Under the described admixture scenario, the consensus-tree based algorithm should first identify that there are k subpopulations in the data. Second, the algorithm should output a model bipartition set $B = \{b_1^M, b_2^M, ..., b_r^M\}$ characterizing the evolutionary history of the populations. Finally, the algorithm should produce a weight vector $W = \{w_0, w_1, w_2, ..., w_r\}$, representing the numbers of observed variants most likely to correspond to each model bipartition $b_1^M, b_2^M, ..., b_r^M$ as well one additional weight attributed to a "null bipartition," essentially a noise term collecting observed variations that appear not to correspond to any population-level bipartition.

To infer the parameter set $\Theta = \{\mathcal{M}, \mathcal{T}, \mathcal{A}, \theta\}$, we estimate the distribution of the posterior probability of the parameters given the observed weights $W = \{w_0, w_1, w_2, ..., w_r\}$ using a similar model described in Chapter 3

$$P(\Theta|W) \quad = \quad \frac{P(W|\Theta)P(\Theta)}{P(W)}$$

Since we have no prior knowledge of the parameters, we assume a uniform distribution for the prior. Hence, we have

$$P(\Theta|W) \propto P(W|\Theta)$$

If we know the exact genealogy of the individuals at a particular segment of the genome where no recombinations have occurred, under the assumption of an infinite sites model, the number of mutations would be Poisson distributed with mean equal to the length of the genealogy l_G multiplied by the number of base pairs l in the segment and the mutation rate μ. While the assumption of infinite sites is an approximation, it is a reasonable one when the mutation rate per site per generation is much smaller than the inverse length of the genealogy, as recurrent mutations become highly unlikely under such a scenario.

Since a genealogy is a tree, it can be broken down into a set of bipartitions (tree edges) weighted by elapsed times. For each bipartition in the genealogy, we can determine the most likely assignment of that bipartition to a branch in the consensus model, with the total number of bipartitions yielding an estimated length of the branch in the consensus model. Given a branch length, we can then calculate the expected number of mutations assigned to each model branch given a mutation rate and the length of the DNA fragment. If a potential candidate genealogy closely resembles the true genealogy, we would expect the number of observed mutations assigned to each model branch to closely match the number of expected mutations assigned to each model branch. Our likelihood function can thus be broken down as a function of a genealogy and branch lengths assigned to the model branches. Because the entire genome is made of fragments of DNA having different genealogies due to recombinations, our likelihood function will have to sum over possible mixtures of genealogies that might collectively explain the full set of fragments.

To make MCMC sampling of the likelihood function practical, we make two simplifications described in Chapter 3 that drastically reduce the number of steps needed to achieve convergence in exchange for a modest decrease in precision. First, we substitute expected coalescence times for integration over the full distribution of possible times. Second, we assume that the admixed evolutionary scenario as a whole is described by a mixture of a finite number m of distinct genealogies. Chapter 3 showed these assumptions introduce modest errors in accuracy for genome-scale data while substantially simplifying the computational problem. The result is the following likelihood function, a generalization of that derived in [121]:

$$P(W|\Theta) \;\; = \;\; \prod_{i=0}^{r} \sum_{\mathcal{G}} P(w_i|l_{b_i^M}) P(l_{b_i^M}|\mathcal{G}) P(\mathcal{G}|\Theta)$$

where $l_{b_i}^M$ is the total branch length assigned to ith model branch in the consensus model, $P(w_i|l_{b_i^M}) =$

$\mathrm{Poisson}(w_i; \mu \times l_{b_i^M})$, and $\hat{\mathcal{G}}$ denotes the simplified genealogy set reduced to m genealogies.

4.1.3 MCMC Sampling

We estimate likelihoods of potential models by Metropolis sampling, where the state of the model is the set of all parameters $\Theta = \{\mathcal{M}, T, A, \theta\}$ and the set of possible genealogies $\hat{\mathcal{G}}$ spanning the genome, where $|\hat{\mathcal{G}}| = m$. The likelihood of any state $X_o = \{x^o_{\mathcal{M}}, x^o_{\mathcal{T}}, x^o_{\mathcal{A}}, x^o_{\theta}, x^o_{\hat{\mathcal{G}}}\}$ is then expressed as follows:

$$P(X_o|W) \propto \left(\prod_{i=0}^{r} P(w_i|l_{b_i^M})\right) P(l_{b_i^M}|x^o_{\hat{\mathcal{G}}}) P(x^o_{\hat{\mathcal{G}}}|x^o_{\mathcal{M}}, x^o_{\mathcal{T}}, x^o_{\mathcal{A}}, x^o_{\theta}))$$

New states X_n are then sampled by first sampling a new discrete model uniformly from the set of possible models consistent with the number of populations identified in the consensus tree phase and next sampling new continuous values for the quantitative parameters $\mathcal{T}$, $\mathcal{A}$, and θ from independent Gaussian distributions with $\mu_{o,t_i} = x_{o,t_i}$, $\mu_{o,\alpha_i} = x_{o,\alpha}$, and $\mu_{o,\theta} = x_{o,\theta}$ and σ_{t_i}, σ_{α_i}, and σ_{θ}. Using the three-population scenario as an example, the chain will first sample a model from one of the six possible models shown in Fig. 4.1.3. Suppose a model with admixture between P_1 and P_3 was selected (Fig. 4.1.3(b)), the chain will then proceed to sample new x_{n,t_1}, x_{n,t_2}, x_{n,α_1}, and x_{θ} from Gaussian distributions with means μ_{o,t_1}, μ_{o,t_2}, μ_{o,α_1}, $\mu_{o,\theta}$ and standard deviations σ_{t_1}, σ_{t_2}, σ_{α_1}, σ_{θ} respectively. If a model with no admixture was selected, the chain will proceed to sample new x_{n,t_1} and x_{n,t_2} from Gaussian distributions but automatically set $x_{n,\alpha}$ to 1. The variances of the Gaussian distributions are adjusted during the burn-in period but subsequently kept fixed, a heuristic adjustment intended to decrease mixing time by balancing the jump size per step and fraction of jumps accepted. Specifically, variances of the Gaussian distributions are adjusted according to absolute difference between the expected and the observed number of mutations divided by the observed number of mutations. If the difference between the expected and the observed is small, the variances would be close to zero or a minimum value specified by the user. If the difference between the expected and the observed is large, the variances would be set at a maximum value specified by default or by the user. Once the algorithm selects values of parameters for the new state X_n, it then samples a new genealogy set through coalescence simulation given the selected new parameters.

Candidate state transitions are then accepted or rejected based on the likelihood function

$$\begin{aligned} Q(X_n|X_o) &= P(x^n_{\mathcal{M}}|x^o_{\mathcal{M}}) \left(\prod_i^{k-1} P\left(x^n_{t_i}|\mu^o_{t_i}, \sigma_{t_i}\right) P\left(x^n_{\alpha_i}|\mu^o_{\alpha_i}, \sigma_{\alpha_i}\right) \right) \\ &\quad \times P\left(x^n_\theta|\mu^o_\theta, \sigma_\theta\right) P\left(\hat{\mathcal{G}}|x^n_T, x^n_{\mathcal{M}}, x^n_A, x^n_\theta\right) \end{aligned}$$

by the Metropolis criterion.

If states X_o and X_n have the same model topology but differ in the population labeling or in time values, admixture fraction, or θ, then the parameters $\mathcal{M}$, $\mathcal{T}$, $\mathcal{A}$ would be distributed identically in the old and new states, letting us simplify the Metropolis acceptance ratio to:

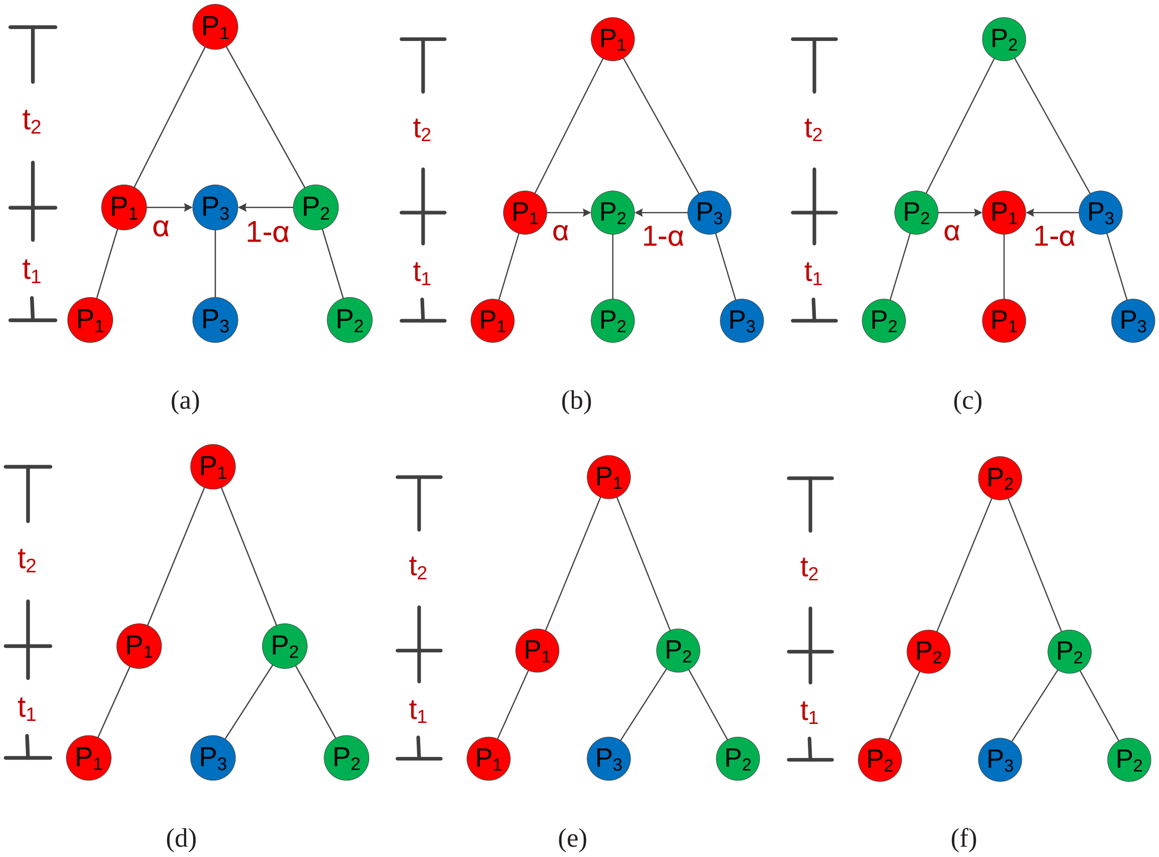
$$r = \frac{\left(\prod_{i=0}^{r} P(w_i|l_{b_i^M})\right)}{\left(\prod_{i=0}^{4} P(w_i|l_{b_i^M})\right)}$$

On the other hand, if X_o and X_n have different admixture events, then the transition probabilities would only cancel out in the acceptance ratio for parameters $\mathcal{M}$, $\mathcal{T}$, and $\alpha_j \in \mathcal{A}$ when the jth event is an admixture event shared by X_n and X_o. Let $\mathcal{A}^-$ be the set of admixture events found in X_o but not X_n and let $\mathcal{A}^+$ be the set of admixture events found in X_n but not X_o. In this case, the Metropolis acceptance ratio can be simplified to:

$$r = \frac{\left(\prod_{i=0}^{r} P(w_i|l_{b_i^M})\right) \prod_{j \in \mathcal{A}^-} P(\alpha^o_j|\alpha^n_j, \sigma_{\alpha_j})}{\left(\prod_{i=0}^{4} P(w_i|l_{b_i^M})\right) \prod_{j \in \mathcal{A}^+} P(\alpha^n_j|\alpha^o_j, \sigma_{\alpha_j})}$$

4.1.4 Validation on Simulated Data

We first validated our method on the classical admixture scenario of two parental populations, P_1 and P_2, and one admixed population, P_3, at the present time. The admixed population P_3 is assumed to have been formed at time t_1 with admixture fraction α_1 from population P_1 and $1 - \alpha_1$ from population P_2. We used a total of 90 different simulated datasets generated in our



(a) (b) (c)

(d) (e) (f)

Figure 4.2: Possible evolutionary models for 3 populations. Top: Models with admixture events at t_1. Bottom: Models with divergence events at t_1. Note that models with different ordering of the populations are excluded because they are isomorphic to one of the six models presented.

prior work[121] consisting of all possible combinations of t_1 = $\{400, 800, 1200, 2000, 4000\}$, t_2 = $\{6000, 8000, 20000\}$, and α_1 = $\{0.05, 0.2, 0.6\}$ on 3.5×10^6-base sequences and on 3.5×10^7-base sequences. Each simulated dataset consists 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 chose this set of datasets as a baseline for comparison to existing coalescence-based algorithms for reconstructing admixture events that are limited to this specific three-population model.

To evaluate the performance of our method on the two-parental and one admixture scenario, we first assessed the quality of the evolutionary model selected by our method by the fraction of trials in which our method assigned the maximum likelihood to the correct evolutionary model. This measure gives us a reasonably stringent assessment of accuracy in selecting the correct topology for the history independent of its quantitative parameters.

We further assessed the quality of the method's time and admixture fraction estimations by comparing results obtained by our method with those of a leading method for learning admixture fractions and divergence times: *MEAdmix* [126]. *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 gCLEAX, one key difference between *MEAdmix* and gCLEAX 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, gCLEAX 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 [17, 77, 126], we chose to compare to *MEAdmix* because it estimates similar continuous parameters to gCLEAX 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 [11, 77], 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 gCLEAX and *MEAdmix* on the 135 simulated datasets and assessed error by averaging the absolute difference between the true and estimated parameter values for each parameter, $(|\hat{\Theta} - \Theta|)$. When running our method on simulated data, we set the number of genealogies for gCLEAX to be m=30. We ran the MCMC chain 10 times with 1,000 steps of burn-in followed by 10,000 steps of sampling for each run. For adjusting for the variances during the burn-in period, we arbitrarily used a minimum and maximum variance of $t = 0.001\theta$ and $t = 0.15\theta$ for time and $\alpha = 0.01$ and $\alpha = 0.15$ for admixture proportion. For θ, we first calculated a rough lower and upper bound by assuming all other parameter values are known. If we set all the divergence and admixture times to have occurred after every lineage within the subpopulations coalesced, we would get a coarse estimate of the lower bound by dividing the expected number of mutations from such a scenario by the observed number of mutations. On the other hand, if we set all the divergence times and admixture times to have occurred at exactly time zero, we would get a coarse estimate of the upper bound of θ by dividing the expected by the observed number of mutations. Using these coarse lower and upper bounds, we set our minimum and maximum variance of θ to be 0.1% and 5% of the difference between the upper bound and the lower bound. 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 next evaluated our method on a number of simulated datasets generated using different evolution scenarios with four modern population groups, denoted P_1, P_2, P_3, and P_4. We simulated data consisting of three evolutionary events at times t_1, t_2, and t_3 that resulted in the formation of four population groups at the present time. At the most ancient time, t_3, we simulated a divergence event that splits the ancestral population into P_1 and the parental population, P_{234}, consisting of P_2, P_3, and P_4. Then, at time t_2, either a divergence event splits the parental population, P_{234}, into P_2 and parental population P_{34} or an admixture event occurs between P_1

and P_{234} with admixture proportion α_2 from P_1 and $1-\alpha_2$ from P_{234} to yield an admixed population P_2 and parental populations P_1 and P_{34}. Finally, at time t_1, either a divergence event occurs to form P_3 and P_4 or an admixture event occurs between P_2 and P_{34} with α_1 admixture proportion from P_2 to form admixed population P_4 and parental population P_2 and P_3. We generated a total of 36 different datasets from combinations of $t_1 = \{2000, 4000\}$, $t_2 = \{4000, 10000\}$, $t_3 = \{6000, 20000\}$, $\alpha_1 = \{0, 0.05, 0.2\}$, and $\alpha_2 = \{0, 0.3, 0.6\}$ where $t_1 < t_2 < t_3$. We chose the coalescence simulator MS [48] for generating the simulated datasets. In all of our simulations, we assumed the effective population size of each population to be 10,000. We set the mutation rate to be 10^{-9} per base pair per generation, the recombination rate to be 10^{-8} per generation for simulations, and the length of the chromosome to be 7×10^7 base pairs.

Using the same configurations as with the evaluation of the three-population scenarios, we evaluated our method's ability to find the right evolutionary model on the simulated four-population datasets by computing the percentage of correctly inferred evolutionary models. Because we are not aware of any other algorithm that performs joint inference of evolutionary model, time, and admixture parameters for four or more populations, we instead compared our estimates of the time parameters to those of a commonly applied divergence estimator based on Wright's F_{st} statistic described by Reynolds *et al.* [93]. Letting $\bar{P}_{ijk}$ be the frequency at the ith site of the jth allele in the kth population, a commonly applied estimator of F_{ST} between any two populations with n total samples is

$$\hat{F}_{ST} \quad = \quad \frac{\sum_i \left(\frac{1}{2} \sum_j \left(\bar{P}_{ij1} - \bar{P}_{ij2} \right)^2 - \frac{1}{2(2n-1)} \left(2 - \sum_j \left(\bar{P}_{ij1}^2 + \bar{P}_{ij2}^2 \right) \right) \right)}{\sum_j \left(1 - \sum_j \bar{P}_{ij1} \bar{P}_{ij2} \right)}$$

Under a model of neutral divergence from an ancestral population, we can estimate the time of the divergence using the following formula:

$$\hat{t} = -\log(1 - \hat{F}_{ST})$$

where t is in unit of 2N generations. We computed the divergence times between all pairs of populations and used the UPGMA tree reconstruction algorithm [106] to generate an evolutionary history for these times. The estimated time for each divergence point in the tree was then used for comparison to our method. For each divergence or admixture time point (t_1, t_2, and t_3), we computed the average difference between the true and estimated time from the 36 different simulated datasets we generated.

4.1.5 Validation on Real Data

We further evaluated our method by applying it to two real large scale genome variation datasets. We first evaluated our method using sequence data from 1,000 Genomes Project Phase I release version 3 in NCBI build 37 [6]. The dataset consists of 1,092 individuals from a number of different ethnic backgrounds that can be largely grouped into four broad subpopulations by continent 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 individual have Asian ancestry from China and Japan. The remaining 181 individuals have American ancestry and are mainly admixed individuals from Mexico, Puerto Rico, and Columbia. We note that the data deviates somewhat from the assumptions of our model because we do not have a sample descended from the Native American subpopulation that would have contributed to the ancestral admixed American populations. Rather, we use modern Asian individuals as a proxy for a modern Native American population. Due to computational issues, we ran our consensus tree algorithm on a uniformly selected subsample of variant sites across the whole genome consisting of 100,000 sites to derive a summary description of the dataset as a set of model bipartitions representing the population clusters. Once we identified a set of model bipartitions, we then used the bipartition set on the complete genome to compute the weight assigned to each model bipartition and use the set of pairs of model bipartitions and weights to estimate the evolutionary history and its parameters.

In addition to sequencing datasets from the 1,000 Genomes Project, we also tested our method on the HapMap Phase III dataset (phased, release 2) [5] which genotyped over 1.6 million SNP sites 1,184 individuals from 11 global populations. Instead of using the whole 1,184 individuals, we chose three non-admixed populations and one admixed population among the 11 global populations to minimize the effect of ascertainment bias. The three non-admixed populations were the 117 individuals with African ancestry from Ibadan, Nigeria (YRI); 115 individuals with Asian ancestry from Beijing, China and Tokyo, Japan; and 170 individuals with European ancestry from Utah, USA. The admixed population consisted of 52 individuals with Mexican ancestry from Los Angeles, CA. Again, the data is not ideal for our model because of the lack of a modern non-admixed Native American population. We would expect the model to treat modern Asian populations as a proxy for this Native American population, with some error in inferences to be expected as a result.

Because a large number of SNPs sites used in genotyping the HapMap samples were collected from other sources, such as dbSNP [99], distribution of the allele frequencies were heavily skewed due to ascertainment bias. In addition, HapMap initially filtered SNP sites with less than 5% minor allele frequency in the sample pools, but later switched to a two-hit criterion strategy where at least two counts of the minor allele must have occurred in the sample in order for the site to be considered a SNP site. Furthermore, HapMap additionally resequenced a portion of the samples on ten 500-kb ENCODE regions [19]. As a result, because of the utilization of multiple ascertainment mechanisms, we chose not to use the whole genome data. Instead, we tried to minimize the number of ascertainment sources by using chromosome 1 SNPs, as chromosome 1 contains no SNPs identified via the ten 500-kb ENCODE regions. In addition, since the published HapMap data did not identify the specific sites obtained using the 5% filtering approach versus the two-hit criterion, we therefore assumed that there is equal chance that a site with frequency lower than 5% is derived from either criterion. Given this SNP ascertainment model, we then modified our likelihood function to correct for ascertainment bias. Let $P(asc|b_j)$ be the probability of ascertaining a variant site given that the site is generated from a mutation occurring along branch $b_j \in \mathcal{G}$. We can then estimate the probability of ascertainment by converting each branch b_j to an allele count and use the method of [78] to estimate an ascertainment probability. Given this ascertainment probability, we correct for ascertainment bias in any branch length of b_j by multiplying the inferred length of b_j by its estimated ascertainment probability. This would give us an ascertainment-corrected branch length $l^{asc}_{b^M_i}$:

$$l^{asc}_{b^M_i} = \sum_{b_j \in \{b|f(b)=i\}} l_{b_j} P(asc|b_j)$$

By substituting $l^{asc}_{b^M_i}$ for $l_{b^M_i}$, we can correct our likelihood model for the presumed ascertainment bias.

For both datasets, we performed 50 trials of MCMC sampling with 2,000 steps of burn-in followed by 10,000 steps of sampling per trial. We set the number of genealogies $|\mathcal{G}|$ to be 30 and the minimum and maximum value of the variances of the Guassian distributions for sampling next parameter values to be the same as used for the simulation study.

4.2 Results

4.2.1 Simulated Data

Two Parental and One Admixed Populations

Table 4.1-4.3 shows the inferred posterior probabilities for all six possible evolutionary models for the three-population scenarios extracted from the outputs of the MCMC chains. Of the 45 datasets simulated from two parental and one admixed population scenarios with 3.5×10^6-base sequences, 43 yielded correct identification of the true evolutionary model as the most likely evolutionary model. The two datasets for which the algorithm incorrectly inferred the most likely evolutionary model were datasets simulated with low admixture fraction ($\alpha = 0.05$), suggesting the variance in 3.5×10^6-base sequences may still be too large in some instances to completely distinguish the different evolutionary models. If we compare the probability distribution between the three different sets of admixture fractions ($\alpha = 0.05, \alpha = 0.2, \alpha = 0.6$), datasets with low admixture fractions ($\alpha = 0.05$) tend to yield flatter distributions, and thus lower-confidence predictions, than do datasets with moderate admixture fractions ($\alpha = 0.2, 0.6$).

When we tested our method on longer sequences, we observed improved estimation of the most likely evolutionary model. Table 4.4-4.6 shows the probability distribution for all six possible evolutionary models on the 45 3.5×10^7-base sequences generated from the three populations scenario. For 3.5×10^7 bases sequences, the algorithm inferred the correct evolutionary model for all the 45 simulated datasets. As with the shorter sequence lengths, datasets with low admixture fractions ($\alpha = 0.05$) tend to yield flatter probability distributions than did datasets with moderate admixture fractions ($\alpha = 0.2, 0.6$).

Quality of the method's parameter estimation on time and admixture fraction also compares favorably to *MEAdmix*. Table 4.7 shows the mean and standard deviation of the absolute difference between estimated parameter values and true parameter values from the 45 simulated datasets with 3.5×10^6-base sequences and 45 simulated datasets with 3.5×10^7-base sequences. Because *MEAdmix* did not show any progress for more than 48 hours for 3.5×10^7-base sequences, we did not obtain results for *MEAdmix* on 3.5×10^7-base sequences. From Table 4.7, we observed gCLEAX generally yields a slight improvement of the average absolute difference between the true and estimated value for t_1, t_2, and α over *MEAdmix* on 3.5×10^6-base sequences. Furthermore, average absolute difference of the parameters tends to improve as we increase from 3.5×10^6-base sequences to 3.5×10^7-base sequences.

Four-Population Scenarios

We next examine the performance on a set of more complicated evolutionary scenarios with four populations having either zero, one, or two admixture events. Table 4.8 shows the inferred probability distribution of the top six evolutionary models for each of the datasets with no admixture events. Our method correctly inferred the true evolutionary model to be the most likely evolutionary model on all four of the datasets with no admixture events. The results for the admixture-free datasets showed that scenarios with long divergence times (t_3) tend to have sharper probability densities, and thus more confident predictions for the correct evolutionary model.

Results on datasets with one admixture event at t_2 followed by a divergence event at t_1 showed similar a trend (Table 4.9). For all the datasets with one admixture event at t_2 followed by a divergence event at t_1, the true model was correctly inferred to be the most likely model. Scenarios with shorter divergence times at t_3 tended to have flatter probability densities compared to scenarios with longer divergence times at t_3.

When we changed the simulated scenario to a divergence event at t_2 followed by an admixture event at t_1, the correct evolutionary model became harder to infer. As shown in Table 4.10, five out of eight datasets with one divergence event at t_2 followed by one admixture event at t_1 were correctly inferred by our method. The three datasets in which the method failed to infer the correct evolutionary model each produced an error of failing to detect the admixture and misinterpreting the formation of the admixed population as being purely a divergence from its major ancestral population. This particular error occurred only in scenarios for which the admixture fraction was low ($\alpha = 0.05$).

For the remainder of the four-population datasets containing two admixture events, the results showed a similar trend. Of the 16 datasets with two admixture events, 13 were correctly inferred by our method. Among the three datasets in which our method failed to infer the correct model as the most likely one, two were scenarios in which a low-proportion admixture event was incorrectly identified as a divergence event from the major population. The remaining case also consisted of an incorrect identification of admixture as divergence, but in a scenario with a comparatively high admixture proportion ($\alpha = 0.2$).

When pooling all the four-population results, we found that our method was able to infer the correct evolutionary model in 83.3% of scenarios. Compared to the results obtained from datasets with three populations, the four-population scenarios tend to have much flatter probability distributions and thus lower confidence in the correct models. In four-populations datasets, the most likely models have an average of 32% inferred posterior probability compared to probabilities of nearly 100% for three-population scenarios. Of the six datasets for which the method failed to infer the correct evolutionary model, five of them were models with low admixture proportions. In these cases, our model mistakenly inferred the admixture event as a divergence event from the major ancestral population.

To assess the quality of the parameter estimation, Table 4.12 shows average error between the estimated parameter values and the true parameter values. The table shows a trend of increasing absolute difference between the estimated and true parameter values as the evolutionary events date further back in time. Estimates of t_1 show the least error, followed by t_2 and t_3. Estimates of admixture proportion likewise show increasing error as events become more ancient. When comparing the average error for our method with that of the F_{ST} approach, we found that gCLEAX yields comparable but slightly superior time estimates. While the differences were almost negligible for t_1 and t_2, our estimates of t_3 improved approximately 60% compared to those based on F_{ST}. The F_{ST} approach does not estimate admixture fractions, though, and we therefore could not use it as a basis for comparison to the quality of our admixture proportion estimates.

4.2.2 Real Data

Running our method on the 1,000 Genomes dataset yielded four clusters of individuals, as shown in Fig. 4.3. 95 chromosomes from the AMR group showed closer resemblance to the EUR group than the AMR group and were classified with the EUR group. 5 chromosomes from the AMR

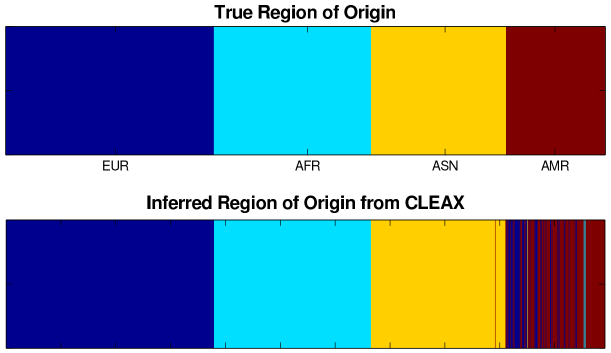
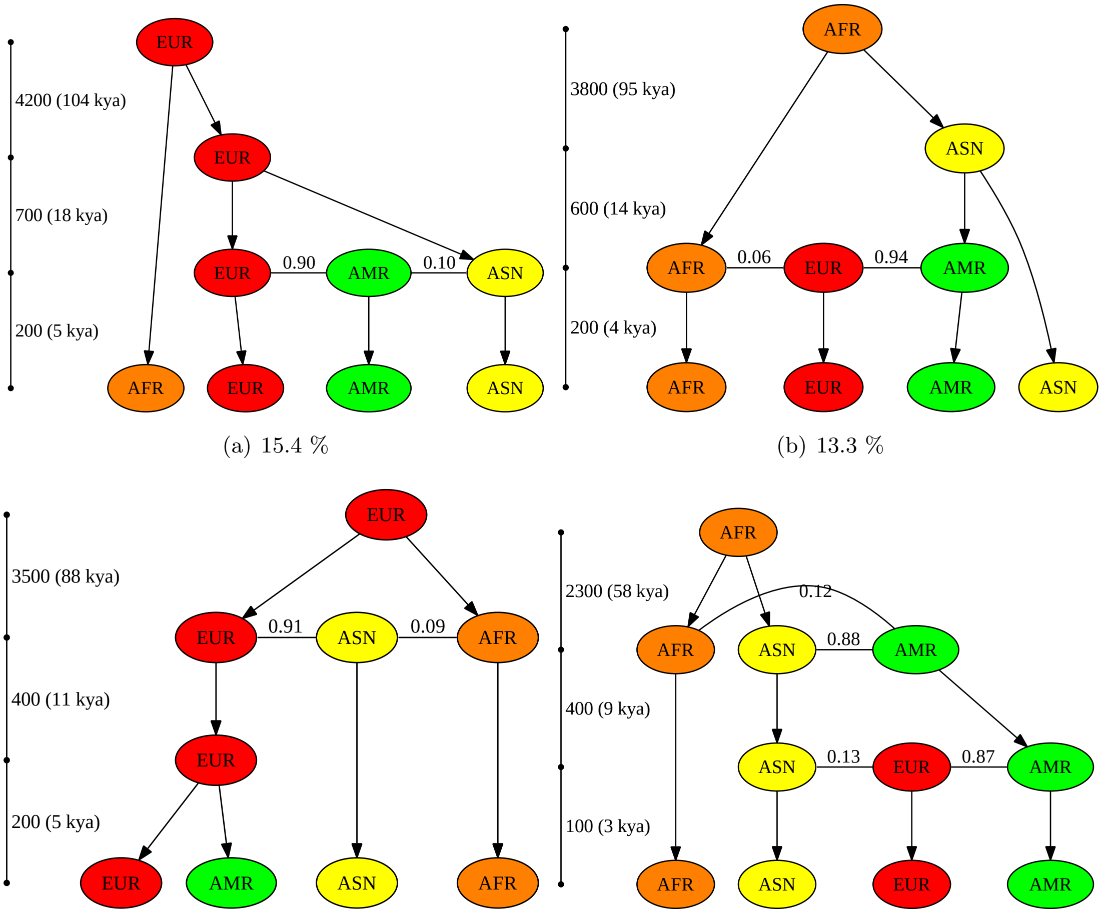


Figure 4.3: Top: True population assignment based on region of origin for 1000 Genome Project dataset. Bottom: Population assignment identified by gCLEAX.

group were grouped with the AFR group. One chromosome from AMR group and one from ASN exchanged membership. While the assignments of individuals to populations do not always match the region of origin exactly, we used the assignment identified by our algorithm rather than the true regions of origin to learn the evolutionary history from the dataset. Running the MCMC sampling on the whole-genome dataset from the 1,000 Genomes Project yielded 21 evolutionary models with non-zero probability. Because many of the models have low posterior probabilities, we only show the four models with the highest posterior probabilities in Fig. 4.4. The most likely evolutionary model from the 1,000 Genomes Project datasets leads to the inference that the AFR group diverged initially from the rest of the populations around 103 kya assuming the effective population size of human to be 10,000 [42] and 25 years per generation. After the divergence of the AFR group from the remaining populations, the EUR and ASN populations diverged at roughly 17 kya. Finally, the AMR group was inferred to be recently admixed from the ancestors of the EUR and ASN groups, with an admixture proportion of 90% EUR and 10% ASN at an estimated time of 5 kya. Other slightly less probable models suggest that the EUR group may have low admixture proportion from AFR or the ASN may have low admixture proportion from the AFR group.

To verify our findings, we also performed our method with ascertainment bias correction on the HapMap Phase 3 dataset. Fig. 4.5 shows the population assignment identified by gCLEAX



(a) 15.4 %

(b) 13.3 %

(c) 12.2 %

(d) 11.8 %

Figure 4.4: Top four evolutionary models obtained from running 50 MCMC chains on whole genome dataset from 1000 Genome Project. The labels represents the four different regions of origins the dataset consisted of. AFR represents groups of individuals with majority having African ancestry. EUR represents group of individuals with majority having European ancestry. ASN represents group of individuals with majority having Asian ancestry. AMR represents group of individuals from the Americas that is believed to be an admixed from Native Americans, African, and Europeans.

There were 26 individuals from the MXL group that were assigned to the CEU group, with the rest of the sample correctly assigned to the MXL group. Running gCLEAX on chromosome 1 of the HapMap dataset using the learned population assignments, gCLEAX identified 16 evolution-

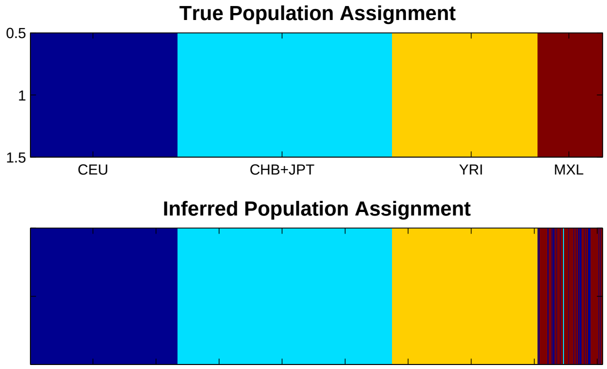
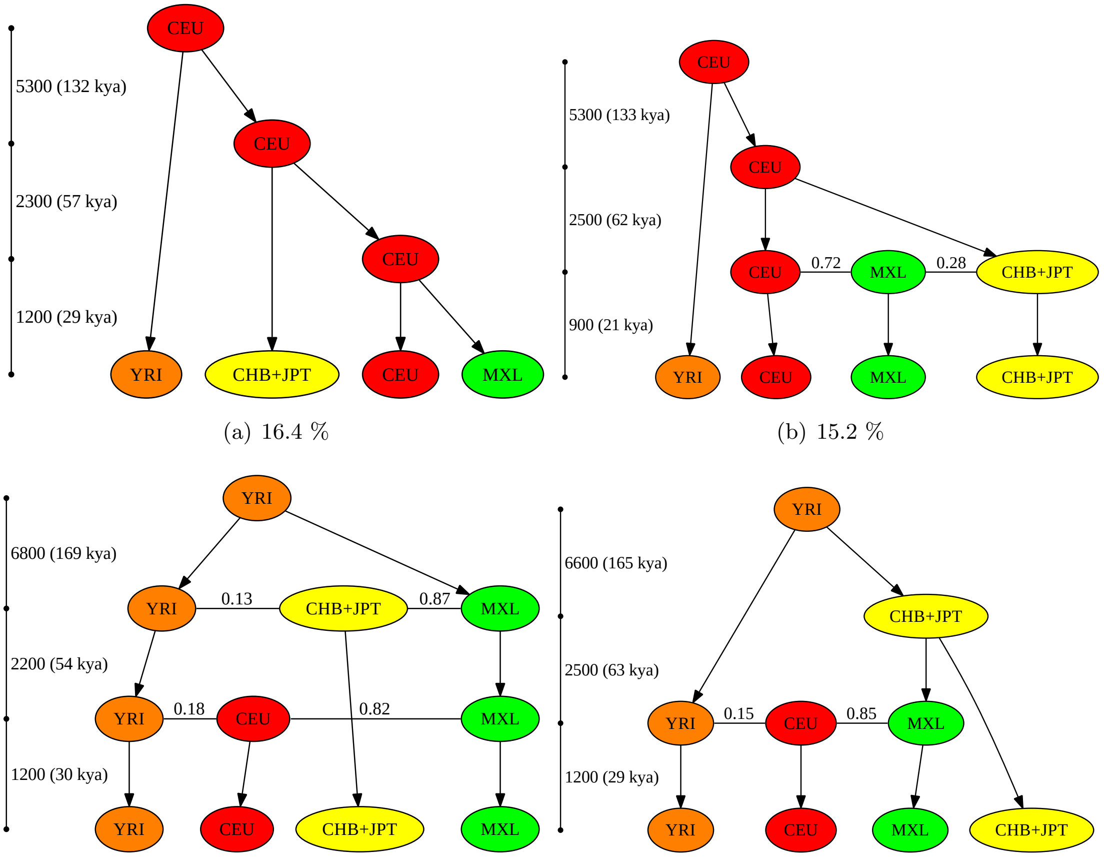


Figure 4.5: Top: True population assignment for HapMap 3 dataset. Bottom: Population assignment identified by gCLEAX.

ary models with non-zero probability. Similar to the results obtained from the 1,000 Genomes Project dataset, Fig. 4.6 shows that the most likely models share nearly the same topology with the exception of missing the admixture of the Mexican group. Instead, the most likely model suggests the MXL group diverged from the CEU group 29 kya. An evolutionary model correctly inferring the admixture event was identified as the second most likely model, with a similar probability value to the most likely one (15.2% vs. 16.4%). As with the 1,000 Genomes dataset, other less probable models suggested that the CEU group or the CHB+JPT groups might have contained admixture from the YRI group. The divergence proportions and admixture times estimated from gCLEAX were also similar to the values obtained from the 1,000 Genomes dataset with the divergence of YRI from the rest of the world at ~ 130 kya, followed by the divergence of CEU and CHB+JPT at ~ 60 kya, and the divergence or admixture event producing MXL at $\sim 20-30$ kya.

4.3 Discussion

Efforts to date at resolving the history of divergence and admixture events by which modern human population structure has emerged have involved a complicated process requiring intensive



(a) 16.4 %

(b) 15.2 %

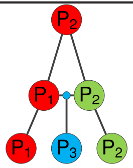
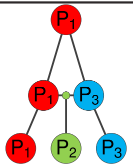
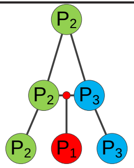
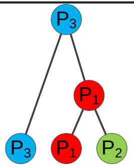
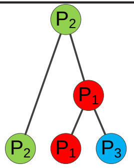
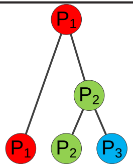
(c) 13.6 %

(d) 11.7 %

Figure 4.6: Top four evolutionary models obtain from running 50 MCMC chains on the chromosome 1 dataset from HapMap Phase 3 Release 2. Labels: YRI=African from Kenya, Nigeria, CEU=European ancestry from Utah, USA, CHB+JPT=Asian from Beijing, China and Tokyo, Japan, MXL=Mexican from California, US

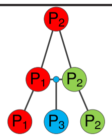
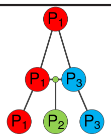
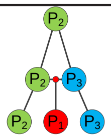
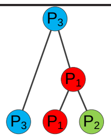
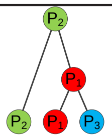
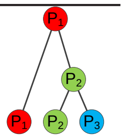
expert intervention and manual integration of numerous software tools and analysis efforts. As we seek to develop population history models of ever greater scope and finer resolution and from ever larger data sets, such manual expert efforts can be expected to become increasingly impractical and error-prone. In this chapter, we propose a first attempt to automate the process of using genetic variation data to infer complicated population history models that capture population-level divergence and admixture events. We have shown that an approach for computing concise summary statistics from large-scale genetic variation data sets and using them to evaluate popu-

Table 4.1: Probability distribution of the possible evolutionary models for the three population scenarios (two parental + one admixed) estimated by gCLEAX for $\alpha = 0.05$ on 3.5×10^6 base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of six possible scenarios (top). In each case, the left-most scenario is correct.

t_1	t_2	α_1						
0.01	0.15	0.05	**0.893**	0.027	0.000	0.041	0.025	0.014
0.02	0.15	0.05	**0.720**	0.120	0.000	0.040	0.070	0.051
0.03	0.15	0.05	**0.384**	0.370	0.000	0.112	0.085	0.049
0.05	0.15	0.05	0.227	**0.545**	0.000	0.082	0.072	0.075
0.10	0.15	0.05	**0.505**	0.396	0.029	0.022	0.037	0.012
0.01	0.20	0.05	**0.510**	0.270	0.000	0.067	0.102	0.051
0.02	0.20	0.05	**0.612**	0.148	0.000	0.089	0.107	0.043
0.03	0.20	0.05	**0.641**	0.192	0.000	0.068	0.064	0.036
0.05	0.20	0.05	**0.524**	0.295	0.000	0.068	0.076	0.037
0.10	0.20	0.05	0.413	**0.500**	0.018	0.020	0.034	0.016
0.01	0.50	0.05	**0.726**	0.075	0.000	0.061	0.070	0.067
0.02	0.50	0.05	**0.424**	0.212	0.000	0.135	0.130	0.099
0.03	0.50	0.05	**0.641**	0.305	0.000	0.024	0.015	0.015
0.05	0.50	0.05	**0.955**	0.024	0.000	0.013	0.007	0.001
0.10	0.50	0.05	**0.921**	0.056	0.000	0.002	0.005	0.016

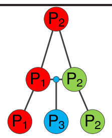
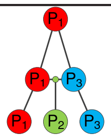
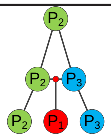
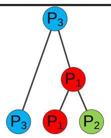
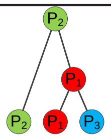
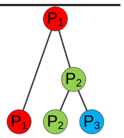
lation models relative to a novel likelihood model provides a feasible method for reconstructing simple scenarios with comparable accuracy to leading methods on well-defined subproblems and without the need for manual intervention to identify population groups or history topologies or

Table 4.2: Probability distribution of the possible evolutionary models for the three population scenario (two parental + one admixed) estimated by gCLEAX for $\alpha = 0.20$ on 3.5×10^6 base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of six possible scenarios (top). In each case, the left-most scenario is correct.

t_1	t_2	α_1						
0.01	0.15	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.15	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.03	0.15	0.20	**0.997**	0.000	0.000	0.003	0.000	0.000
0.05	0.15	0.20	**0.986**	0.010	0.000	0.004	0.000	0.001
0.10	0.15	0.20	**0.582**	0.329	0.053	0.014	0.014	0.009
0.01	0.20	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.20	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.03	0.20	0.20	**0.956**	0.024	0.000	0.010	0.004	0.005
0.05	0.20	0.20	**0.988**	0.003	0.000	0.001	0.000	0.008
0.10	0.20	0.20	**0.898**	0.075	0.000	0.017	0.006	0.004
0.01	0.50	0.20	**0.999**	0.000	0.000	0.000	0.001	0.000
0.02	0.50	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.03	0.50	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.05	0.50	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.10	0.50	0.20	**0.848**	0.054	0.000	0.059	0.022	0.018

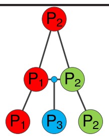
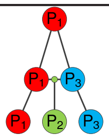
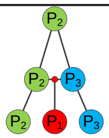
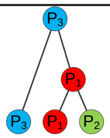
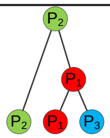
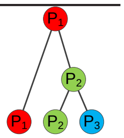
to synthesize results of multiple prediction methods. While our MCMC strategy is similar to a number of prior approaches [17, 77], our algorithm is distinguished by its capability to learn an evolutionary model for an in principle arbitrary number of populations and by its novel strategies for simplifying the likelihood model in ways especially suited to genomic-scale variation

Table 4.3: Probability distribution of the possible evolutionary models for the three population scenario (two parental + one admixed) estimated by gCLEAX for $\alpha = 0.60$ on 3.5×10^6 base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of six possible scenarios (top). In each case, the left-most scenario is correct.

t_1	t_2	α_1						
0.01	0.15	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.15	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.03	0.15	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.05	0.15	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.10	0.15	0.60	**0.726**	0.116	0.142	0.008	0.006	0.001
0.01	0.20	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.20	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.03	0.20	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.05	0.20	0.60	**0.993**	0.000	0.003	0.003	0.000	0.000
0.10	0.20	0.60	**0.867**	0.000	0.105	0.024	0.004	0.000
0.01	0.50	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.50	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.03	0.50	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.05	0.50	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.10	0.50	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000

data sets, trading off large increases in performance for small compromises in accuracy on the assumption of large numbers of variant sites. 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

Table 4.4: Probability distribution of the possible evolutionary models for the three population scenario (two parental + one admixed) estimated by gCLEAX for $\alpha = 0.05$ on 3.5×10^7 base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of six possible scenarios (top). In each case, the left-most scenario is correct.

t_1	t_2	α_1						
0.01	0.15	0.05	**0.702**	0.028	0.000	0.035	0.114	0.120
0.02	0.15	0.05	**0.976**	0.000	0.000	0.011	0.000	0.013
0.03	0.15	0.05	**0.523**	0.198	0.000	0.080	0.063	0.137
0.05	0.15	0.05	**0.628**	0.251	0.000	0.057	0.042	0.022
0.10	0.15	0.05	**0.533**	0.368	0.000	0.052	0.015	0.031
0.01	0.20	0.05	**0.996**	0.003	0.000	0.001	0.000	0.000
0.02	0.20	0.05	**0.697**	0.068	0.000	0.035	0.005	0.194
0.03	0.20	0.05	**0.403**	0.132	0.000	0.154	0.090	0.221
0.05	0.20	0.05	**0.888**	0.000	0.000	0.011	0.000	0.101
0.10	0.20	0.05	**0.429**	0.420	0.000	0.052	0.034	0.065
0.01	0.50	0.05	**0.624**	0.022	0.000	0.019	0.009	0.326
0.02	0.50	0.05	**0.665**	0.003	0.000	0.073	0.010	0.249
0.03	0.50	0.05	**0.542**	0.130	0.000	0.097	0.024	0.207
0.05	0.50	0.05	**0.491**	0.193	0.000	0.052	0.053	0.212
0.10	0.50	0.05	**0.659**	0.077	0.000	0.026	0.069	0.170

aspect of that inference is dependent on the answers to the other two. While the automated models still have a long way to go before they can match the scope of expert-curated analysis, they do establish a proof-of-concept for a principled automated approach and help identify avenues for future work. Continued efforts in this direction will be an important component of advancing

Table 4.5: Probability distribution of the possible evolutionary models for the three population scenario (two parental + one admixed) estimated by gCLEAX for $\alpha = 0.20$ on 3.5×10^7 base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of six possible scenarios (top). In each case, the left-most scenario is correct.

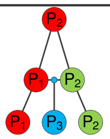
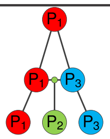
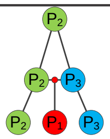
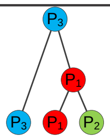
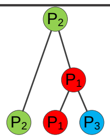
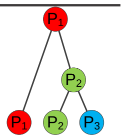
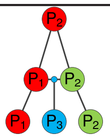
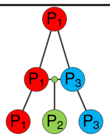
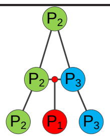
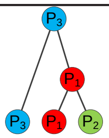
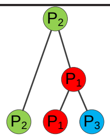
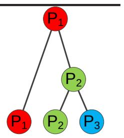
t_1	t_2	α_1						
0.01	0.15	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.15	0.20	**0.974**	0.000	0.000	0.000	0.000	0.026
0.03	0.15	0.20	**0.900**	0.000	0.000	0.000	0.000	0.100
0.05	0.15	0.20	**0.932**	0.030	0.000	0.000	0.000	0.037
0.10	0.15	0.20	**0.571**	0.242	0.044	0.020	0.020	0.103
0.01	0.20	0.20	**0.997**	0.000	0.000	0.000	0.000	0.003
0.02	0.20	0.20	**0.990**	0.000	0.000	0.008	0.000	0.002
0.03	0.20	0.20	**0.994**	0.000	0.000	0.000	0.000	0.006
0.05	0.20	0.20	**0.698**	0.061	0.000	0.024	0.007	0.210
0.10	0.20	0.20	**0.524**	0.379	0.000	0.052	0.012	0.034
0.01	0.50	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.50	0.20	**1.000**	0.000	0.000	0.000	0.000	0.000
0.03	0.50	0.20	**0.990**	0.000	0.000	0.000	0.000	0.010
0.05	0.50	0.20	**0.924**	0.000	0.000	0.005	0.000	0.070
0.10	0.50	0.20	**0.992**	0.000	0.000	0.000	0.000	0.007

Table 4.6: Probability distribution of the possible evolutionary models for the three population scenario (two parental + one admixed) by gCLEAX for $\alpha = 0.60$ on 3.5×10^7 base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of six possible scenarios (top). In each case, the left-most scenario is correct.

t_1	t_2	α_1						
0.01	0.15	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.15	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.03	0.15	0.60	**0.993**	0.000	0.000	0.007	0.000	0.000
0.05	0.15	0.60	**0.987**	0.000	0.011	0.002	0.000	0.000
0.10	0.15	0.60	**0.624**	0.093	0.172	0.007	0.004	0.100
0.01	0.20	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.20	0.60	**0.989**	0.000	0.000	0.011	0.000	0.000
0.03	0.20	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.05	0.20	0.60	**0.927**	0.000	0.000	0.073	0.000	0.000
0.10	0.20	0.60	**0.988**	0.000	0.005	0.006	0.000	0.000
0.01	0.50	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.02	0.50	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.03	0.50	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.05	0.50	0.60	**1.000**	0.000	0.000	0.000	0.000	0.000
0.10	0.50	0.60	**0.999**	0.000	0.000	0.001	0.000	0.000

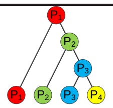
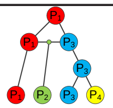
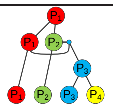
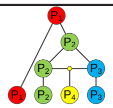
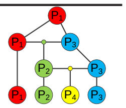
the study of human population history in an era of plentiful genomic sequences.

Validation on real and simulated data demonstrates the generally high accuracy of the methods on at least simple scenarios, with some caveats. Estimates from simulated data show our method to yield comparable and often superior accuracy to leading methods for specific sub-

Table 4.7: Mean and standard deviation of absolute difference between estimated parameter values and true parameter values from the 45 datasets simulated from the three-population scenario (two-parental, one admixed) with 3.5×10^6-base sequences.

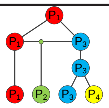
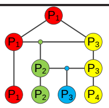
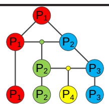
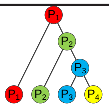
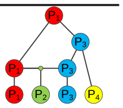
	MEAdmix	gCLEAX	gCLEAX
	(3.5×10^6)	(3.5×10^6)	(3.5×10^7)
$\lvert\hat{\alpha_1} - \alpha_1\rvert$	0.0448 ± 0.0469	0.0436 ± 0.0403	0.0417 ± 0.0360
$\lvert\hat{t_1} - t_1\rvert$	485 ± 384	375 ± 474	208 ± 268
$\lvert\hat{t_2} - t_2\rvert$	2880 ± 4373	2700 ± 2390	2690 ± 2675

Table 4.8: Probability distribution of the top six evolutionary models for four-population scenarios with no admixture events on 7×10^7-base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of the six most likely possible scenarios (top). In each case, the left-most scenario is correct.

t_1	t_2	t_3	α_1	α_2					
0.04	0.08	0.15	0.00	0.00	**0.094**	0.040	0.093	0.085	0.053
0.04	0.08	0.50	0.00	0.00	**0.840**	0.160	0.000	0.000	0.000
0.04	0.25	0.50	0.00	0.00	**0.444**	0.158	0.075	0.035	0.021
0.08	0.25	0.50	0.00	0.00	**0.320**	0.100	0.059	0.020	0.060

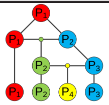
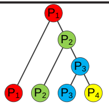
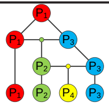
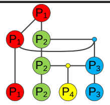
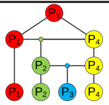
problems, such as reconstructing the specific three-population admixture scenario for which prior methods were designed. Similarly, estimates from our method proved slightly superior to those of a standard approach based on F_{ST} statistics at estimating divergence times in these scenarios. Analysis on the two real datasets also compared favorably with existing literature in most respects. Evolutionary models learned from the 1,000 Genomes Project dataset closely matched the general consensus on the history of modern human populations with the inference of ancestral splits of African from non-African followed by European from Asian, and with the inference

Table 4.9: Probability distribution of the top 6 evolutionary models for four population scenario with one admixture event followed by one divergence event on 7×10^7-base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of the six most likely possible scenarios (top). In each case, the left-most scenario is correct.

t_1	t_2	t_3	α_1	α_2					
0.04	0.08	0.15	0.00	0.30	**0.180**	0.113	0.060	0.060	0.020
0.04	0.08	0.50	0.00	0.30	**0.240**	0.110	0.140	0.060	0.220
0.04	0.25	0.50	0.00	0.30	**0.352**	0.040	0.091	0.192	0.000
0.08	0.25	0.50	0.00	0.30	**0.380**	0.040	0.080	0.200	0.000
0.04	0.08	0.15	0.00	0.60	**0.167**	0.153	0.139	0.018	0.039
0.04	0.08	0.50	0.00	0.60	**0.280**	0.199	0.060	0.000	0.040
0.04	0.25	0.50	0.00	0.60	**0.440**	0.021	0.034	0.000	0.030
0.08	0.25	0.50	0.00	0.60	**0.320**	0.060	0.000	0.000	0.020

of Mexicans, Puerto Ricans, and Colombians as admixed groups from these ancestral populations [66, 102, 112]. The HapMap dataset supported a similar model, although with an incorrect inference of the Mexican group as diverged from Europeans rather than admixed from European and Asians, with the correct model having slightly lower posterior probability. As the simulated data results suggest, the method can have difficulty distinguishing admixture from divergence when the minor ancestral population's contribution is small. Parameter estimates on these data are generally well supported by the literature [61, 132]. Estimates of the African-European divergence time of 103 kya and 133 kya from 1000 Genomes and HapMap data respectively are consistent with the STR estimation by [132] (62-133kya) and the HMM estimation by [61] (60-120 kya). Admixture proportion estimation of the Mexican group was also generally consistent with prior studies. While the 90% admixture proportion from the EUR group for the AMR group is high relative to prior estimates, it is not dramatically off from the ranges supported by prior

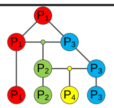
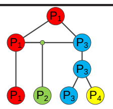
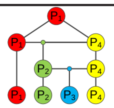
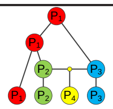
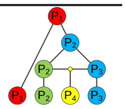
Table 4.10: Probability distribution of the possible evolutionary model for four-population scenarios with one divergence event followed by one admixture event on 7×10^7-base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of the six most likely possible scenarios (top). In each case, the left-most scenario is correct.

t_1	t_2	t_3	α_1	α_2					
0.04	0.08	0.15	0.05	0.00	0.100	**0.180**	0.060	0.060	0.100
0.04	0.08	0.50	0.05	0.00	0.080	**0.340**	0.043	0.000	0.100
0.04	0.25	0.50	0.05	0.00	**0.320**	0.080	0.220	0.100	0.020
0.08	0.25	0.50	0.05	0.00	0.026	**0.340**	0.060	0.000	0.100
0.04	0.08	0.15	0.20	0.00	**0.320**	0.040	0.080	0.140	0.020
0.04	0.08	0.50	0.20	0.00	**0.383**	0.100	0.080	0.056	0.060
0.04	0.25	0.50	0.20	0.00	**0.573**	0.000	0.180	0.120	0.000
0.08	0.25	0.50	0.20	0.00	**0.450**	0.020	0.105	0.119	0.000

studies by Martinez-Cortes *et al.* [66] and Tang *et al.* [112], which both estimated admixture proportions in the ranges of 60-70% European, 10-20% African, and 10-20% Native American. These deviations could be explained by the assumption of our model that admixture events occur only between pairs of populations and by the lack of a non-admixed Native American sample from which our method could learn.

The most significant error of our model on the real data is in substantially overestimating the age of the admixture time for the Mexican, Puerto Rican, and Columbian groups. The most likely source of error here is, again, the lack of a modern Native American sample in the data set, leading the method to approximate American samples as admixtures of the European and Asian populations it had available. We would expect this inference to lead to a misattribution of a large number of mutations distinguishing Asian from Native American as mutations distinguishing

Table 4.11: Probability distribution of the top six possible evolutionary models for four-population scenarios with two admixture events on 7×10^7-base sequences. Each row shows one input parameter set followed by the estimated posterior probabilities for each of the six most likely possible scenarios (top). In each case, the left-most scenario is correct.

t_1	t_2	t_3	α_1	α_2					
0.04	0.08	0.15	0.05	0.30	**0.220**	0.201	0.080	0.020	0.020
0.04	0.08	0.50	0.05	0.30	**0.293**	0.188	0.155	0.000	0.060
0.04	0.25	0.50	0.05	0.30	**0.200**	0.080	0.120	0.000	0.120
0.08	0.25	0.50	0.05	0.30	**0.164**	0.107	0.093	0.006	0.138
0.04	0.08	0.15	0.05	0.60	**0.145**	0.083	0.116	0.060	0.000
0.04	0.08	0.50	0.05	0.60	0.140	**0.240**	0.040	0.040	0.000
0.04	0.25	0.50	0.05	0.60	0.120	**0.200**	0.060	0.080	0.020
0.08	0.25	0.50	0.05	0.60	**0.206**	0.148	0.082	0.066	0.007
0.04	0.08	0.15	0.20	0.30	**0.199**	0.160	0.140	0.000	0.120
0.04	0.08	0.50	0.20	0.30	**0.360**	0.320	0.060	0.000	0.040
0.04	0.25	0.50	0.20	0.30	0.280	0.020	0.020	0.020	**0.416**
0.08	0.25	0.50	0.20	0.30	**0.260**	0.160	0.120	0.000	0.160
0.04	0.08	0.15	0.20	0.60	**0.320**	0.100	0.040	0.280	0.000
0.04	0.08	0.50	0.20	0.60	**0.498**	0.060	0.000	0.140	0.000
0.04	0.25	0.50	0.20	0.60	**0.420**	0.000	0.000	0.180	0.000
0.08	0.25	0.50	0.20	0.60	**0.340**	0.007	0.143	0.259	0.000

the ancestral from modern admixed American populations, and in turn to an overestimated time since admixture. Another source of error may be the assumption that the effective population size is constant and equal for all population groups when in fact we would expect smaller bottlenecks for the newly admixed groups.

Table 4.12: Mean and standard deviation of the absolute difference between the estimated parameter values and the true parameter values from the 36 datasets simulated from different four-population scenarios with zero, one, or two admixture events on 7×10^7-base sequences.

	gCLEAX	$\hat{F_{st}}$
$\lvert\hat{\alpha_1} - \alpha_1\rvert$	0.0272 ± 0.0389	–
$\lvert\hat{\alpha_2} - \alpha_2\rvert$	0.0576 ± 0.0588	–
$\lvert\hat{t_1} - t_1\rvert$	203 ± 228	264 ± 194
$\lvert\hat{t_2} - t_2\rvert$	1100 ± 930	1130 ± 847
$\lvert\hat{t_3} - t_3\rvert$	3570 ± 2440	5790 ± 3110

Comparing the results between three-population and four-population scenarios does suggest that scaling to large models is problematic due to the combinatorial explosion in numbers of possible models as the number of populations increases. That explosion in possible models would be expected to impact both the true sharpness of the probability density and the mixing time of the MCMC method. More data can in principle help to address the former problem, leading to greater support for correct models although at the cost of increased computation time. The current tests examined data sets up to the size of a single large chromosome, but there is no reason in principle the method cannot extend to all variants in a full human genome. Similarly, the datasets examined ranged to the size of hundreds of individuals, but adding more individuals and especially individuals from a more diverse set of modern populations can also be expected to allow better discrimination of correct from incorrect histories and finer resolution of those histories. Further algorithmic improvements may be needed to address slow mixing time, though, especially as more data is added. Our primary criterion of success, inferring exactly the right model with high probability, may also be too stringent a criterion for larger and more complicated models. The MCMC approach makes it trivial to define more modest but achievable goals, though, such as identifying those specific events that are well-supported by a given data set and estimating their parameters.

Chapter 5

Structured Testing for Disease Association

Case-control association mapping has been one of the most widely used methods for identifying loci involved in disease inheritance [3, 94, 127]. The method typically tests for association between a marker locus and trait by measuring the amount of allele-frequency differences between individuals with the phenotype of interest (cases) and unrelated healthy individuals (controls). A strong statistical association between genotypes at the marker locus and the phenotype is usually considered evidence for a potential candidate region where the disease locus may be located.

Since the availability of large scale genomic datasets, a number of methods to test for disease association has been proposed [26, 131]. One classic test used in genome-wide association studies (GWAS) is the Pearson's chi-square test [60]. The count of the allele for each biallelic SNP site can be summarized in a 2×2 contingency table which tests the null hypothesis that the disease has no effect on the distribution of the allele counts. Other tests, such as likelihood ratio test, logistic regression, and Cohran-Armitage test for trend, are also frequently used in GWAS [60, 131].

Despite its popularity, loci identified by association mapping often require close scrutiny and careful analysis to exclude false-positives that are the consequence of stratification differences between the cases and the controls [3, 20]. This stratification is most often due to differences in population substructure in cases and controls. For example, when cases and controls have different fractions of individuals from different subpopulations, this would increase the likelihood of a marker locus being falsely identified by the statistical test as a candidate locus due to significant differences in the allele frequencies between the subpopulations.

To address this problem, one trivial approach to eliminating spurious associations in case-control studies would be to avoid selecting samples where structure is clearly present or to balance the samples such that population structure is evenly distributed between cases and controls. Despite successfully avoiding the effect of population stratification, selecting individuals free of population substructures can be a tedious and, in many occasions, impossible task. A more desirable alternative is to apply a test statistic that corrects for population substructures [83, 86, 115]. One popular method is the EIGENSTRAT [83] which uses principal component analysis (PCA) to identify the axes of variation. The axes of the highest variation should provide clues as to the ancestry of individuals and population structure exhibited in the dataset. By learning the ancestry proportion of each sample based on where each sample maps to the axes of variation, EIGENSTRAT effectively corrects for population stratification by readjusting the genotypes and phenotypes of each sample according to the ancestries of the sample. While this method corrects each SNP for population structure according to a global ancestry assignment for each sample, the method does not consider the relationships between ancestral populations and the local substructure for each SNP. Another common approach is to partition the individuals into their respective subpopulations and perform a stratified association test, such as the Cochran-Mantel-Haenszel test [115]. On the other hand, given the history output learned automatically from genetic data using methods described in our previous chapters, we could potentially take advantage of the structural and relational information learned from the data itself to reduce the number of false discovery of candidate loci due to population stratification in association studies. In this chapter, we propose two simple structured association tests, as a proof of concept, that correct for population stratification effects using the outputs of our minimum description length (MDL)-based consensus tree algorithm in Chapter 2.

5.1 Methods

In Chapter 2, we proposed an MDL algorithm for learning population histories that identifies a set of robust edges persistent across the entire genomic dataset. The algorithm specifically produces a set of model edges or bipartitions $B^M = \{B_1^M, ..., B_r^M\}$ by optimizing the following MDL-based objective function:

$$\mathcal{L}(T_M, T_1, T_2, ..., T_n) \quad = \quad \underset{B^M \in \mathcal{T}}{\arg\min} \left(L(B^M) + \sum_{i=0}^{n} L(B_i | B^M) + f(B^M) \right)$$

where $B_1, B_2, ..., B_n$ are the observed bipartition sets derived from phylogenies generated from a SNP dataset partitioned into n windows of k SNPs. $L(B^M)$ computes the minimum description length of the model edges B_M and $L(B_i|B^M)$ computes the minimum description length required to explain the observed (input) bipartitions B_i given the model edges (B^M). The function $f(B^M)$ defines an additional penalty for proposing a model tree that is overly complex.

At the end of the optimization procedure, the method would produce a set of robust model bipartitions that represent the edges of the population history. At the same time, finding the optimal model bipartitions indirectly provide local structural information through the computation of the cost function. During the computation of $L(B_i|B^M)$, the method computes the conditional entropy of each observed bipartition $b \in B_i$ relative to each model bipartition $b^M \in B^M$. This computation not only allows us to compute the minimum description length but also effectively estimates the probability distribution over population subdivision events from which the variant site may have most recently arisen from. For example, suppose that we identify an observed bipartition derived from a variant site closely resembling the model bipartition representing $P_2|P_1P_3$ in a dataset with three populations (P_1, P_2, and P_3). The fact that the observed bipartition has the closest resemblance to the model bipartition representing $P_2|P_1P_3$ would indicate that the mutation may have likely occurred simultaneously with or after P_2 diverged from other populations. At the same time, the resemblance to $P_2|P_1P_3$ would also suggest the variant site may have a substructure effect between P_2 and rest of the populations. Given the information that the mutation may be a result of the population split of P_2 with other populations, we could then correct the substructure effect for the test by separating the samples into P_2 and rest of the populations and test for association separately.

5.1.1 Localized Structured Association Test

To correct for population stratification, we would first identify the set of model bipartitions B^M. We would then use our function $L(B_i|B^M)$ to find the most likely model bipartition $b^M \in B^M$ showing the closest resemblance to each variant site i from the genomic dataset. Let b_i be the observed bipartition derived from biallelic variant site i, B^M be the set of model bipartition, and $H(b_i|b_j^M)$ be the conditional entropy of the observed bipartition b_i given model bipartition $b_j^M \in B^M$. Then, we can find the optimal model bipartition $b_j^{M,*}$ showing the closest resemblance to b_i using the following formula:

$$b_i^{M,*} \quad = \quad \underset{b_j^M \in B^M}{\arg\min} \left(H(b_i|b_j^M) \right)$$

Given that $b_i^{M,*}$ is the closest model bipartition or population edge resembling the ith variant site, we would split the chromosome copies from individuals (2 from each individual) into two partitions according to b_i and perform a Cochran-Mantel-Haenszel (CMH) test of association conditional on the model partition $b_i^{M,*}$. The intuition behind such an approach is that, by splitting the samples into its respective substructures using the closest matching model bipartition, we should effectively remove the most prominent population substructure affecting the variant site.

Suppose the best model bipartition $b_i^{M,*}$ resembling variant site i split the chromosomes in the dataset into part p_0 and part p_1, then we can count the number of chromosomes having any specific genotype and specific phenotype. If the genotype is biallelic and the phenotype is either having the disease (cases) or healthy (controls), then we can set up two 2×2 contingency tables shown in Fig. 5.1.1.

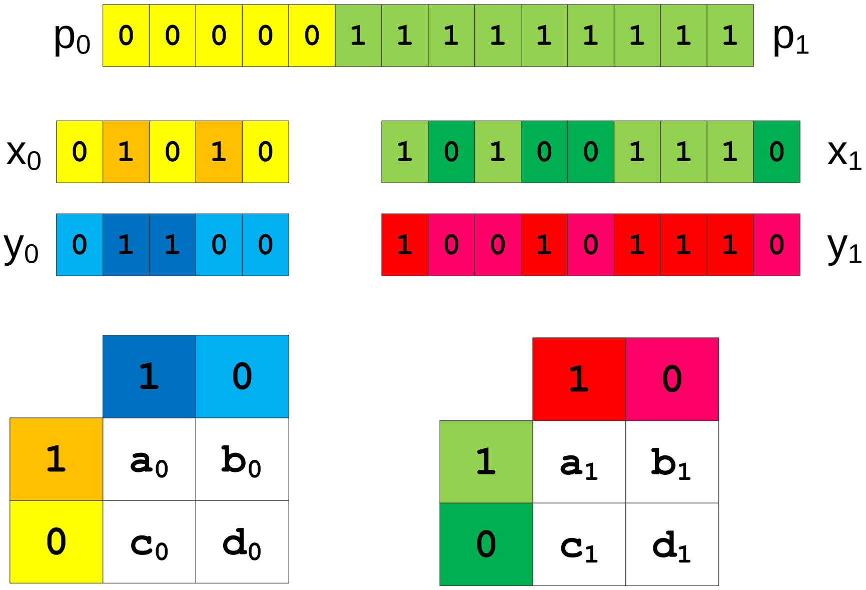


Figure 5.1: Example of a Cochran-Mantel-Haenszel test applied using separation of chromsomes into two partition sets using the a model partition. Chromosomes partitioend into part p_0 and part p_1 by model bipartition b_i are separately tested for association of the phenotype Y with genotype X using Cochran-Mantel-Haenszel test by counting the number of chromosomes having each specific phenotype with each possible genotype in 2×2 contingency tables. a_0, b_0, c_0, and d_0 represents the number of individuals in part p_0 having genotype and phenotype pairs (y_1, x_1), (y_0, x_1), (y_1, x_0), and (y_0, x_1) respectively. a_1, b_1, c_1, and d_1 represents the number of chromosomes in part p_1 having genotype and phenotype pairs (y_1, x_1), (y_0, x_1), (y_1, x_0), and (y_0, x_1) respectively.

Given a_0, b_0, c_0, d_0 as the number of chromosomes in partition p_0 having phenotype-genotype pair (y_1, x_1), (y_0, x_1), (y_1, x_0), (y_0, x_0) respectively and a_1, b_1, c_1, d_1 as the number of chromosomes in partition p_1 having phenotype-genotype pair (y_1, x_1), (y_0, x_1), (y_1, x_0), (y_0, x_0) respectively, the Cochran-Mantel-Haenszel test statistic can be computed as:

$$\chi^2_{MH} = \frac{\left(\left|a_0 - \frac{(a_0+b_0)(a_0+c_0)}{a_0+b_0+c_0+d_0}\right| + \left|a_1 - \frac{(a_1+b_1)(a_1+c_1)}{a_1+b_1+c_1+d_1}\right| - 0.5\right)^2}{\frac{(a_0+b_0)(a_0+c_0)(b_0+d_0)(c_0+d_0)}{(a_0+b_0+c_0+d_0)^3-(a_0+b_0+c_0+d_0)^2} + \frac{(a_1+b_1)(a_1+c_1)(b_1+d_1)(c_1+d_1)}{(a_1+b_1+c_1+d_1)^3-(a_1+b_1+c_1+d_1)^2}}$$

The numerator computes the squared sum of the deviations between the observed and expected values with a continuity correction added. The denominator estimates the variance of the squared differences. The test statistic follows a chi-squared distribution with one degree of freedom. When the observed and expected values are similar, the test statistics would be smaller thus reducing the chance to reject the null hypothesis that the phenotype and genotype are independent. When the variance of the squared differences is large, suggesting that observed can significantly deviates from the mean under the null hypothesis because of uncertainties, the test statistic would also be small. By identifying sites that have large differences between expected and observed allele frequencies in both partitions through the test statistic, we should be able to identify sites that are associated with the phenotype without the confounding effect of population substructure.

5.1.2 Weighted Localized Structured Association Test

One issue with using the most likely model bipartition for variant site is that the most likely model bipartition can have a very similar score to the other model bipartitions in the population model. If the scores of the model bipartitions are similar to each other, using the most likely one can leave out slightly less prominent substructure effects on the variant sites, which may lead to more false-positive associations. To address this issue, we note that the conditional entropy we computed for each variant site given a model bipartition can roughly be treated as the negative log probability

[38] that the observed bipartition is best explained by the model bipartition. Equivalently, we can view this as the probability that bipartition b_j^M is the key substructure effect for this variant site. Under this assumption, we can compute a weight w_j that the model bipartition b_j^M is the key substructure effect for the variant site i. We would then compute the p-value for each split of the population by model bipartition $b_j^M \in B^M$ using the Cochran-Mantel-Haenszel test and computing each p-value using a weighted sum of the possible bipartitions producing the effect. To justify the formulation, we claim that what we are interested in learning is the probability of observing a value equal to or greater than the respective test statistic given that the genotype X and phenotype Y are independent (H_0) and that zero or one division of individuals by a model bipartition is influencing the allele frequencies. Suppose M_j is a model that specifies how the sample will split into two partitions using the jth model bipartition $b_j^M \in B^M$ and $\mathcal{M}$ is the set of all models that specifies how we can split the sample into two partitions. Then, we can rewrite the probability as the following:

$$P(X^2_{MH} \geq c|H_0, \mathcal{M}) \;\; = \;\; \sum_{M_j \in \mathcal{M}} P(M_j)P(X^2_{MH,j} \geq c_j|H_0, M_j)$$

Since $P(M_j|D)$ is the belief that b_j^M is the prominent substructure effect, $P(M_j|D)$ would simply be the weight w_j we computed from the conditional entropy:

$$P(M_j|D) = w_j \;\; = \;\; \frac{2^{-H(b_i|b_j^M)}}{\sum_k 2^{-H(b_i|b_k^M)}}$$

Combining the definition above, we get:

$$P(X^2_{MH} \geq c|H_0, \mathcal{M}) \;\; = \;\; \sum_{j=1}^{|\mathcal{M}|} \frac{2^{-H(b_i|b_j^M)}}{\sum_k 2^{-H(b_i|b_k^M)}} P(X^2_{MH,j} \geq c_j|H_0, M_j)$$

5.1.3 Validation

To validate our approach, we first generated a simulated dataset using the coalescent simulator *MS* [48]. We simulated genotypes of 150 affected and 150 control individuals resulting in a total of 600 chromosomes on a sequence with 3.5×10^6 bases using a mutation rate of 10^{-9} per site per generation, a recombination rate of 10^{-8} per generation, and effective population size of 10,000. We assume that the individuals were sampled from three subpopulations (P_1, P_2, and P_3). At time $t_1 = 2,000$ generations ago, P_2 and P_3 diverged into their respective subpopulations. At time $t_2 = 6,000$ generations ago, P_1 and the ancestral population of P_2 and P_3 diverged. We assume the causal mutation occurred after the divergence of P_1 and the parental population of P_2 and P_3. This scenario should in theory result in a much higher relative risk to ascertain a case from P_2 and P_3 than a case from P_1. Among the 150 affected, we arbitrarily assigned 75 individuals to P_2 and 75 individuals to P_3. In this dataset, we assumed no one from P_1 contracted the disease. Of the 150 controls, 50 individuals were assigned to each of the three populations.

To simulate the candidate locus, we sample the genotype by computing the probability of observing genotype X given phenotype Y and population assignment K, $P(X|Y,K)$. Through conditional probability and Bayes' theorem, we can rewrite the probability as:

$$\begin{aligned}
P(X|Y,K) &= \frac{P(X|Y,K)}{\sum_{X\in\{0,1\}} P(X|Y,K)} \\
&= \frac{P(K,Y|X)P(X)}{\sum_{X\in\{0,1\}} P(K,Y|X)P(X)} \\
&= \frac{P(Y|K,X)P(K|X)P(X)}{\sum_{X\in\{0,1\}} P(Y|K,X)P(K|X)P(X)} \\
&= \frac{P(Y|K,X)P(X|K)P(K)}{\sum_{X\in\{0,1\}} P(Y|K,X)P(X|K)P(K)} \\
&= \frac{P(Y|X)P(X|K)}{\sum_{X\in\{0,1\}} P(Y|X)P(X|K)}
\end{aligned}$$

where $P(X|K)$ would correspond to the allele frequency of each allele in each population and $P(Y|X)$ would be the probability of having disease or normal phenotype given the genotype.

Given the equation above, we generated genotypes for 120 independent disease loci using a probability of $P(Y = 1|X = 1) = \mathcal{N}(0.7, 0.1)$, $P(Y = 1|X = 0) = \mathcal{N}(0.3, 0.1)$, $P(X = 1|K = 1) = 0.1$, $P(X = 1|K = 2) = 0.6$, and $P(X = 1|K = 3) = 0.6$.

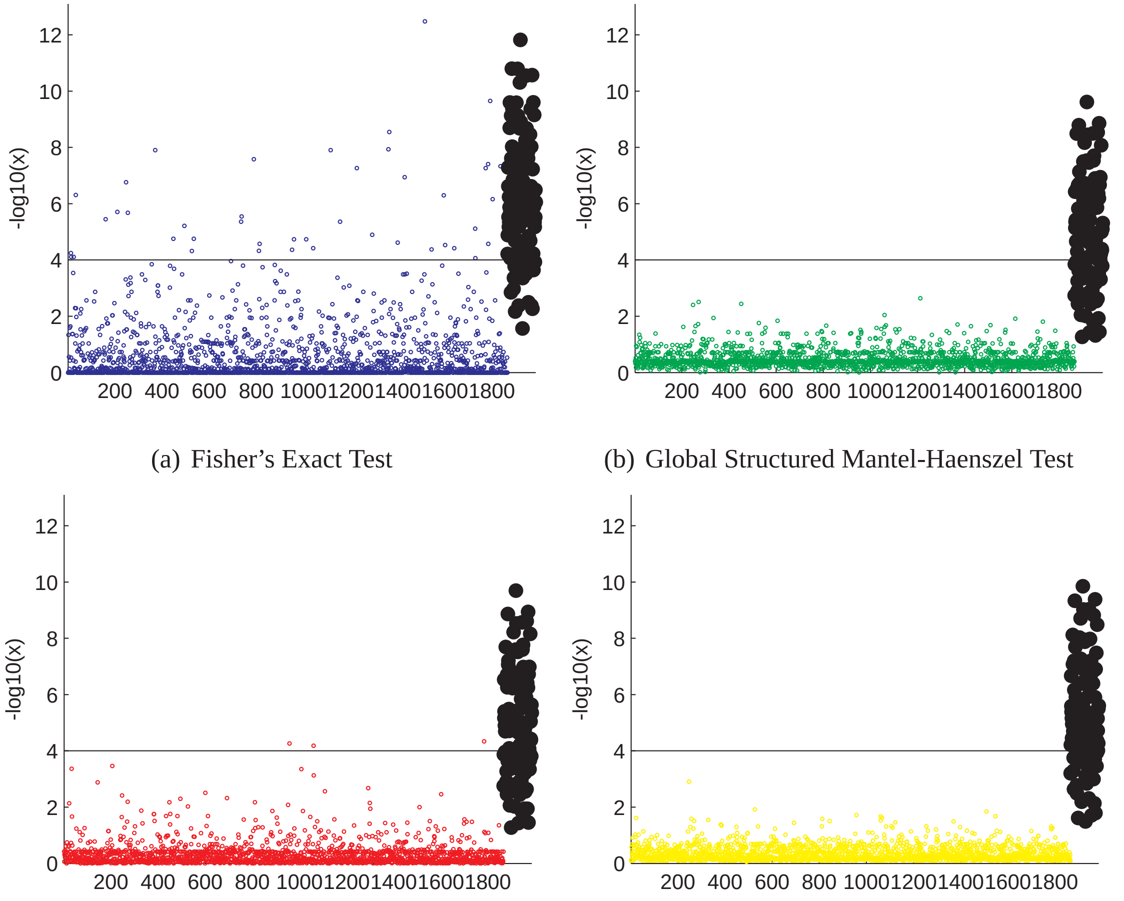
In addition to simulated data, we also used the HapMap phase 2 chromosome 21 dataset as a test dataset [4]. The dataset consisted of 90 Utah residents with ancestry from Northern and Western Europe (CEU); 90 individuals with African ancestry from Ibadan, Nigeria (YRI); 45 Han Chinese from Beijing, China (CHB); and 45 Japanese in Tokyo, Japan (JPT). For the CEU and YRI groups, which consist of trio data (parents and a child), we used only the 60 unrelated parents with haplotypes as inferred by the HapMap consortium. Given that our algorithm identified three populations from the same dataset in Chapter 2, we artificially assigned 5 individuals from YRI, 40 individuals from CEU, and 40 individuals from CHB+JPT as cases. The remaining 125 individuals were labeled as controls. Using the same approach as with the fully simulated data, we then generated the genotypes for the 120 candidate locus using a probability of $P(Y = 1|X = 1) = \mathcal{N}(0.7, 0.1)$, $P(Y = 1|X = 0) = \mathcal{N}(0.3, 0.1)$, $P(X = 1|K = YRI) = 0.1$, $P(X = 1|K = CEU) = 0.66$, and $P(X = 1|K = CHB + JPT) = 0.66$.

To compare the performance of our proposed methods, we tested both the non-weighted and weighted localized structured association test against a simple and common statistical test used for association testing known as Fisher's exact test [33] that does not correct for population stratification. In addition to Fisher's exact test, we further tested our methods against a simple stratified association test commonly used in disease association testing by software such PLINK [87]. Rather than testing for association using the most prominent substructure effect for each variant site, we instead perform the Cochran-Mantel-Haenszel test on the global substructures identified. In this approach, the data would be partitioned into all known subpopulations identified. In both the simulated dataset and HapMap dataset, this would mean that we partitioned the individuals into three subsets and test for associations separately in the subsets.

5.2 Results

Fig. 5.2 shows the computed negative log p-value using Fisher's exact test, the global Cochran-Mantel-Haenszel test, the localized Cochran-Mantel-Haenszel test, and the weighted localized Cochran-Mantel-Haenszel test. From the plot, we observed that negative log p-values for loci not associated with the disease locus from the three Cochran-Mantel-Haenszel tests are much lower than for the Fisher's exact test on the entire sample. On the other hand, candidate loci showed a slightly lower negative log p-values using the three Cochran-Mantel-Haenszel tests than Fisher's exact test. If we reject the null hypothesis H_0 using a p-value of 0.0001, we achieved a type I error of 0.023, 0.000, 0.002, and 0.000 for Fisher's exact, global, local, weighted-local Cochran-Mantel-Haenszel test respectively suggesting the stratified test have a higher specificity. The Fisher's exact test has a power of 0.94 while the Cochran-mantel-Haenszel tests have power of 0.88, 0.86, 0.88 for global, local, weighted local respectively indicating the stratified test may have lower sensitivity than the non-stratified test. However, when comparing the three stratified tests, the weighted-local test has the same power as the global test while decreasing the type I error.

Semi-simulated data from the HapMap phase 2 dataset also showed similar results compared to the fully simulated data. Fig. 5.2 shows the negative log p-values for the genetic markers typed on chromosome 21. Similar to the results from fully simulated data, the results from the HapMap dataset showed lower number of false positives using the Cochran-Mantel-Haenszel tests. The negative log p-value for the candidate loci were also lower in the three Cochran-Mantel-Haenszel tests compared to the Fisher's exact test. When using a reject threshold of 0.0001, the weighted-local Cochran-Mantel-Haenszel test gave the lowest type I error (0.000) while Fisher's exact test gave the worst type I error (0.048). Global Cochran-Mantel-Haenszel, however, outperformed non-weighted local Cochran-Mantel-Haenszel with type I error of 0.0005 compared to 0.0027. Power analysis for the four test showed similar results compared to fully simulated data. Power of Fisher's exact test remained the best of the four with 0.93 followed by both weighted local



(a) Fisher's Exact Test

(b) Global Structured Mantel-Haenszel Test

(c) Localized Structured Mantel-Haenszel Test

(d) Weighted Localized Structured Mantel-Haenszel

Figure 5.2: Negative log p-values computed for all 1866 variant sites generated from coalescent simulator, MS, and 120 simulated loci. We arbitrarily assigned 150 individuals out of 300 individuals as cases. The simulated candidate locus is attached to the end of the dataset. Black lines represent the rejection threshold of 0.0001. (a) Negative log p-values obtained from Fisher's Exact Test. (b) Negative p-value obtained from Mantel-Haenszel test on the three subpopulations identified. (c) Negative log p-values obtained from Mantel-Haenszel test on the best local split of populations using our MDL-based scoring function. (d) Negative log p-values obtained from Mantel-Haenzel test using a weighted local split of populations using our MDL-based scoring function.

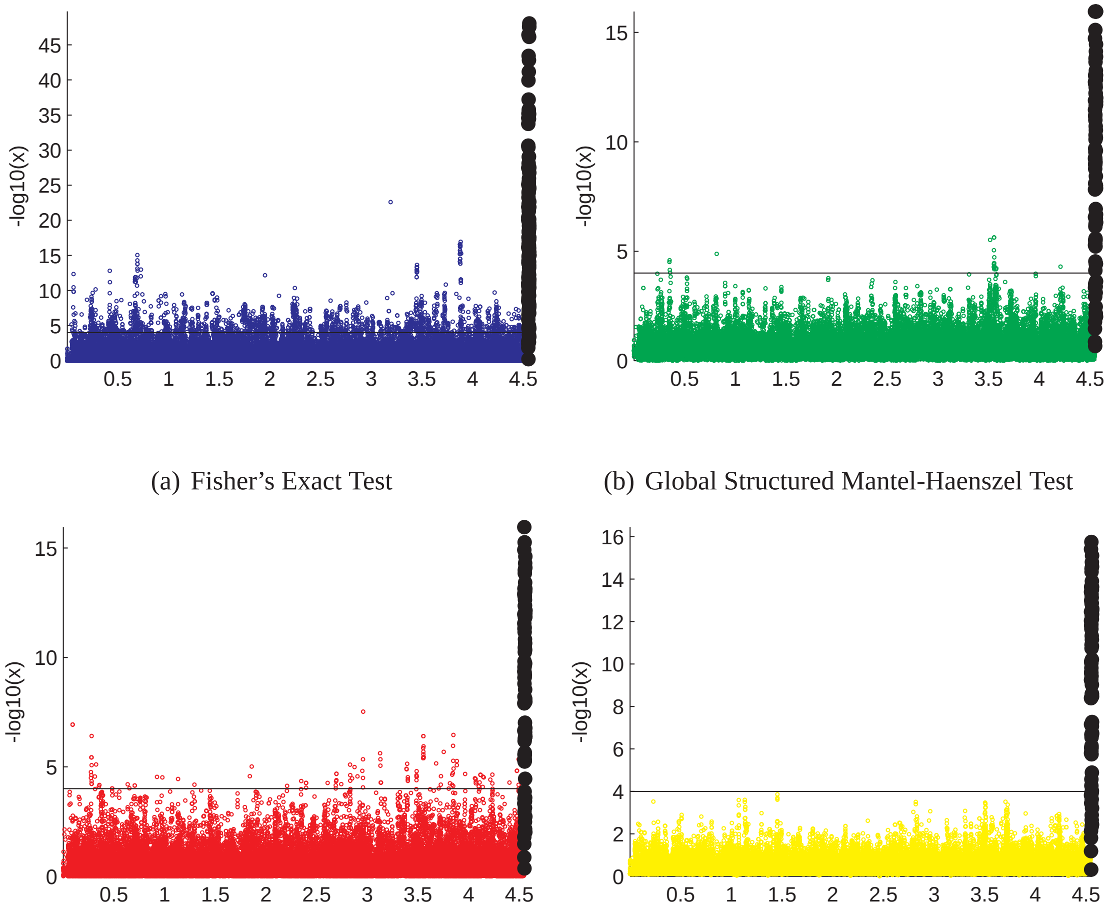
and global Cochran-Mantel-Haenszel of 0.90. Local Cochran-Mantel-Haenszel test achieved the lowest power of the four with a power of 0.89.

5.3 Discussion

In this chapter, we proposed two methods to perform structured association tests. The novelty of the methods comes from the fact that the methods take structural and relational information learned from the dataset using our MDL-based consensus tree algorithm to correct for population substructure. Intuitively, the methods map each variant site to a branch in the population history learned from the dataset. Using that branch as the most prominent local population structure, the method then corrects for the effect of population substructure by separately testing for association in each local subpopulation. In one approach, the method uses the most likely population edge resembling the observed variant site to locally correct for population substructure. In the second approach, we instead correct for population substructure effects using each of the population splits identified weighted by the probability belief that the population split is the key stratification effect.

Results from both fully simulated and semi simulated data suggest that structured association tests using Cochran-Mantel-Haenszel test statistic are capable of removing spurious discoveries due to population structure. From the results, we observed that the global structured association test seemed to perform better than the non-weighted localized structured association test. Such trends may likely result from the fact that the most likely population substructure computed using our minimum description length function may not always be the correct one, especially when the observed bipartition may be quite similar to two or more model bipartitions. However, by weighting and combining the p-values performed in our weighted test, we effectively reduce the effect of such issues and achieve the best type I error among the four methods.

From the results, we also observed that the three stratified association tests tend to have lower power than Fisher's exact test that does not take the effect of population substructure



(a) Fisher's Exact Test

(b) Global Structured Mantel-Haenszel Test

(c) Localized Structured Mantel-Haenszel Test

(d) Weighted Localized Structured Mantel-Haenszel

Figure 5.3: Negative log p-values computed for all 45487 variant sites obtained from HapMap phase 2 dataset and 120 simulated candidate loci. We arbitrarily assigned 105 individuals out of 210 as cases. The simulated candidate locus is attached to the end of the dataset. Black line in the plots represents the rejection threshold of 0.0001. (a) Negative log p-values obtained from Fisher's Exact Test. (b) Negative p-value obtained from Mantel-Haenszel test on the three subpopulations identified. (c) Negative log p-values obtained from Mantel-Haenszel test on the best local split of populations using our MDL-based scoring function. (d) Negative log p-values obtained from Mantel-Haenszel test using weighted local splits of populations using our MDL-based scoring function.

into account. Such pattern, while undesirable, may be a limitation of the stratified tests since samples are divided into smaller group with fewer samples in each group. As a result, this can reduce the confidence of the statistical test due to fewer data to support association in each group. Despite lower power when comparing to non-stratified association test, our weighted-localized Cochran-Mantel-Haenszel test achieved the same power as the global test suggesting the weighted-localized test is able achieve a better type I error while maintaining the same power.

Although further comparison with existing methods for dealing with substructure, such as STRAT [86] or EIGENSTRAT [83] is needed, results nonetheless show a promising approach to test for association under the effect of population substructure. Furthermore, comparison to Fisher's exact test shows that there are fewer false positives and overall reduction of p-values that are not truly associated with the disease phenotype. Such results suggest that our method could potentially provide a promising direction to improve association tests under the effect of population structure. Given that our prior methods in learning population history can obtain time and admixture information, a possible improvement would be to incorporate such information into the test statistics to further improve the performance of the disease association test under the presence of population substructure.

Chapter 6

Conclusions and Future Directions

Automatically identifying subpopulations and learning their evolutionary history from the ever growing genomic datasets is an important but challenging problem that, to the best of our knowledge, no existing algorithms have yet to solve. In this thesis, we have developed novel algorithms specifically designed to address the problem of learning population histories from large scale genomic datasets. Starting with a basic model based on the theory of minimum description length, we have shown that it is computational feasible to automatically identify population substructures and infer evolutionary history at the same time from large scale datasets under the assumption that the underlying histories have little or no admixture between the populations. Validation results from Chapter 2 showed the minimum description length-based consensus tree algorithm is capable of identifying the distinct substructures within the dataset and the relationships between the substructures with high accuracy. We have also found the algorithm to be robust over a wide range of parameter variations, providing confidence that the ability of the algorithm to identify the correct substructures and histories. Furthermore, population history produced by the algorithm from real human datasets is consistent with existing beliefs about human evolution. Analysis of the computational time needed to run the algorithm further demonstrated its ability to handle large-scale datasets in a reasonable time frame, establishing the method as the first practical algorithm for joint inference of population structure and its history from large genomic datasets. Comparing to existing consensus tree algorithms, the proposed algorithm is also one of the first practical algorithms able to identify a meaningful consensus tree from noisy datasets in which observed splits between populations often involve some individuals being classified in different population groups at different observed data points.

The thesis has further addressed the problem of admixture, a common phenomenon in humans of mixing genetic materials between populations, by developing a method for learning the parameters describing the population history in the presence of admixture. By building upon the algorithm described in Chapter 2, we have demonstrated the feasibility of accurately quantifying the timing and fractions of admixture and divergence events between populations throughout the history. Beginning with an initial model of learning parameter values pertaining to a simple two-parental and one admixed population scenario, we showed that estimation of parameters was at least as good as or better than existing models on simulated sequences with lengths ranging from roughly equivalent to large fragments of DNA chromosomes to complete chromosome lengths. Sensitivity analysis on varying population sizes suggests the method is moderately robust to time-varying population sizes. In addition, the minimal loss of accuracy on the long sequence datasets also suggests that the mean length assumption used for the coalescent time of n lineages is a good strategy that could benefit coalescent simulations and other simulations utilizing repeated draws from exponential distributions.

Given the model for quantifying parameters of admixture event for two-parental and one-admixed population scenarios, the thesis further generalized the approach to automatically learning population histories and their associated parameters for three or more populations. Analysis on the simulated data suggests the generalized approach shared similar accuracy for quantifying the parameters of admixture and divergence events with the initial model for two-parental and one-admixed population scenarios. The method's ability to correctly infer the correct population model and its associated parameters suggests its applicability as the first practical algorithm for automatically learning histories of three or more populations under the presence of admixture.

Finally, as a proof of concept, the thesis tested the applicability of incorporating population history information into association test statistics to remove the effects of population substructure. Analyses on simulated and semi simulated datasets suggested the incorporation of population history can indeed reduce type I errors due to population stratification differences. The success of the structured association test demonstrates one of the practical applications of population histories learned from genetic data.

6.1 Future Work

The work presented in the thesis provides one possible strategy for learning population histories with or without the presence of admixture for large-scale genomic datasets. While the models presented in the thesis yielded reasonably accurate solutions under various scenarios, the problem of efficiently solving for all scenarios of population history, and perhaps at a much higher resolution remains a challenge. Despite the difficulty of learning population histories from genetic data, the models here nonetheless act as a stepping stone to reach the holy grail of accurately learning detailed history of populations in all possible scenarios directly from genetic variation datasets.

Two scenarios of population history the thesis did not investigate are the issues of migration between different populations and varying effective population size. Migration of individuals in large fractions can have a large impact on the pattern of genetic variations and thus the accuracy of the inference algorithms. Significant changes in effective population size, as demonstrated in our analysis in Chapter 3, can also have large impacts on the accuracy of the current model. Despite the fact that our current method does not incorporate migration parameters into the model, an extension to include migration events could in theory be achieved by introducing migration rate parameters, M, and simulatig the coalescent process with the specified migration rates. Incorporation of migration and variable effective population sizes should provide a more realistic and in-depth depiction of the human population history, but such enhancements may also introduce additional challenges as additional parameters can also increase the cost of computations.

Our current model only considers divergence and admixture events as two possible evolutionary events. While these two evolutionary events can recreate most population scenarios, events such as convergence can also exist. By incorporating other evolutionary events, we can achieve a more accurate depiction of the evolutionary history of humans. Such an incorporation into our model can be naively done by simply adding additional population models into the set of possible model $\mathcal{M}$. However, introduction of new evolutionary events can increase the number of models exponentially, thus making the inference expensive due to the number of steps needed to sample from the MCMC chain. Nonetheless, such issues can be remedied by developing prior distributions on the set of possible models to steer the chain to visit models that one believes to be more probable, thus reducing the number of steps needed.

In addition to addressing the fundamental questions in population genetics, learning population history provides important information for understanding disease origin. With the ability to automatically learn historic information from molecular data, the algorithms open up the possibility of using that information to improve power in identifying disease-causing alleles through genome-wide association studies (GWAS). One other possible direction in harnessing such historic information is to incorporate the population history into regularized linear regression ($Y = \beta X + \lambda|\beta|_{L_1/L_2}$) for identifying weakly associated loci X that are associated with a complex disease Y. By setting up the regression coefficients for each marker and each population in the dataset, we can, for example, formulate our objective such that an associated or causal mutation that occurs in two or more populations would be more likely to be selected given a population history that specifies the populations were grouped as one ancestral population for a long time. By doing so, mutations associated with the disease that may not have been picked up by simple regression could potentially be more likely identified with the historic information.

Another possible direction in utilizing the population history is to incorporate historic information into a new test statistic for identifying loci associated with a phenotype of interest. While

Chapter 5 demonstrated an improvement of the association test using the population structure information, we did not incorporate any other historic information, such as time and admixture fractions into our test statistic. By incorporating the length of each model bipartition or branch, for example, we might reduce the chance of over-correcting for the population substructure effect when the most likely model bipartition resembling the observed bipartition does not have a high confidence.

While we have only considered a few possibilities for utilizing the historic information learned from the models presented in this thesis, the ability to automatically identify and learn histories of populations from genetic variation data should provide tremendous opportunities for solving numerous problems in population and medical genetics.

Bibliography

[1] Pan asian single nucleotide polymorphism. Online. `http://www4a.biotec.or.th/PASNP`. 1.1

[2] The international hapmap project. *Nature*, 426(6968):789–796, 2003. 1.2

[3] Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007. 10.1038/nature05911. 5

[4] A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, 2007. 1.2, 2, 2.2.2, 3.2, 3.5, 5.1.3

[5] Integrating common and rare genetic variation in diverse human populations. *Nature*, 467(7311):52–58, 2010. 1.2, 4, 4.1.5

[6] An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012. 1.2, 3.2, 3.5, 4, 4.1.5

[7] E.N. Adams. N-trees as nestings: Complexity, similarity, and consensus. *Journal of Classification*, 3(2):299–317, 1986. 2.1.1

[8] David Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome Research*, 19(9):1655–1664, 2009. 1.4

[9] David Balding, Martin Bishop, and Chris Cannings. *Handbook of Statistical Genetics*. John Wiley and Sons, 2007. ISBN 9780470997628. 4

[10] D.M. Behar, S. Rosset, J. Blue-Smith, O. Balanovsky, S. Tzur, D. Comas, R.J. Mitchell,

L. Quintana-Murci, C. Tyler-Smith, R.S. Wells, and The Genographic Consortium. The genographic project public participation mitochondrial DNA database. *PLoS Genet*, 3(6): e104, 2007. 2

[11] G Bertorelle and L Excoffier. Inferring admixture proportions from molecular data. *Molecular Biology and Evolution*, 15(10):1298–1311, 1998. 1.5, 3.1, 3.2, 4.1.4

[12] The Bovine HapMap Consortium. Genome-wide survey of snp variation uncovers the genetic structure of cattle breeds. *Science*, 324(5926):528–532, 2009. 3.2, 3.3, 3.5, 3.4

[13] Katarzyna Bryc, Adam Auton, Matthew R. Nelson, Jorge R. Oksenberg, Stephen L. Hauser, Scott Williams, Alain Froment, Jean-Marie Bodo, Charles Wambebe, Sarah A. Tishkoff, and Carlos 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. 3, 4

[14] R.L. Cann, M. Stoneking, and A.C. Wilson. Mitochondrial DNA and human evolution. *Nature*, 325(6099):31–36, 1987. 2

[15] L. L. Cavalli-Sforza and W. F. Bodmer. *The Genetics of Human Populations*. W. H. Freeman and Company, 1971. 4

[16] Ranajit Chakraborty. Gene admixture in human populations: Models and predictions. *American Journal of Physical Anthropology*, 29(S7):1–43, 1986. 3, 3.1, 4

[17] L. Chikhi, M.W. Bruford, and M.A. Beaumont. Estimation of admixture proportions: A likelihood-based approach using markov chain monte carlo. *Genetics*, 158(3):1347–1362, 2001. 1.4, 1.5, 3, 3.1, 3.1, 3.2, 3.4, 4, 4.1, 4.1.4, 4.3

[18] Y. J. Chu and T. H. Liu. On the shortest arborescence of a directed graph. *Science Sinica*, 14:1396–1400, 1965. 2.1.2

[19] A. G. Clark, M. J. Hubisz, C. D. Bustamante, S. H. Williamson, and R. Nielsen. Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res*, 15(11):

1496–502, 2005. 4.1.5

[20] David G. Clayton, Neil M. Walker, Deborah J. Smyth, Rebecca Pask, Jason D. Cooper, Lisa M. Maier, Luc J. Smink, Alex C. Lam, Nigel R. Ovington, Helen E. Stevens, Sarah Nutland, Joanna M. M. Howson, Malek Faham, Martin Moorhead, Hywel B. Jones, Matthew Falkowski, Paul Hardenbol, Thomas D. Willis, and John A. Todd. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet*, 37(11):1243–1246, 2005. 10.1038/ng1653. 5

[21] F. S. Collins, M. Morgan, and A. Patrinos. The human genome project: lessons from large-scale biology. *Science*, 300(5617):286–290, April 2003. 2

[22] Francis S. Collins, Ari Patrinos, Elke Jordan, Aravinda Chakravarti, Raymond Gesteland, LeRoy Walters, the members of the DOE, and NIH planning groups. New goals for the u.s. human genome project: 1998-2003. *Science*, 282(5389):682–689, 1998. 1

[23] The 1000 Genomes Project Consortium. 1000 genome project. Online, . `http://www.1000genomes.org/`. 1.1

[24] The Hapmap Consortium. International hapmap project. Online, . `http://hapmap.ncbi.nlm.nih.gov/`. 1.1

[25] The HUGO Pan-Asian SNP Consortium. Mapping human genetic diversity in asia. *Science*, 326(5959):1541–1545, 2009. 1.2

[26] Anna L. Dixon, Liming Liang, Miriam F. Moffatt, Wei Chen, Simon Heath, Kenny C. C. Wong, Jenny Taylor, Edward Burnett, Ivo Gut, Martin Farrall, G. Mark Lathrop, Goncalo R. Abecasis, and William O. C. Cookson. A genome-wide association study of global gene expression. *Nat Genet*, 39(10):1202–1207, 2007. 10.1038/ng2109. 5

[27] Isabelle Dupanloup, Giorgio Bertorelle, Louns Chikhi, and Guido Barbujani. Estimating the impact of prehistoric admixture on the genome of europeans. *Molecular Biology and Evolution*, 21(7):1361–1372, 2004. 3

[28] R. B. Eckhardt. Matching molecular and morphological evolution. *Human Evolution*, 4 (4):317–319, 1989. 1

[29] L Excoffier, P E Smouse, and J M Quattro. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial dna restriction data. *Genetics*, 131(2):479–91, 1992. 4

[30] Joseph Felsenstein. Phylip - phylogeny inference package (version 3.2). *Cladistics*, 5: 164–166, 1989. 1.3.2, 1.5

[31] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, 2003. 1.3.2, 1.5, 4

[32] Lars Feuk, Andrew R. Carson, and Stephen W. Scherer. Structural variation in the human genome. *Nat Rev Genet*, 7(2):85–97, 2006. 1.1

[33] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922. 5.1.3

[34] Olivier Franois, Mathias Currat, Nicolas Ray, Eunjung Han, Laurent Excoffier, and John Novembre. Principal component analysis under population genetic models of range expansion and admixture. *Molecular Biology and Evolution*, 27(6):1257–1268, 2010. 3, 4

[35] D. Garrigan, S. B. Kingan, M. M. Pilkington, J. A. Wilder, M. P. Cox, H. Soodyall, B. Strassmann, G. Destro-Bisol, P. de Knijff, A. Novelletto, J. Friedlaender, and M. F. Hammer. Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, x and y chromosome resequencing data. *Genetics*, 177(4):2195–2207, 2007. 3.3, 3.4

[36] N. J. Gawel, R. L. Jarret, and A. P. Whittemore. Restriction fragment length polymorphism (rflp)-based phylogenetic analysis of musa. *Theoretical and Applied Genetics*, 84(3-4): 286–290, 1992. 1, 1.1

[37] David B. Goldstein and Louns 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. 3

[38] P.D. Grnwald, I.J. Myung, and M.A. Pitt. *Advances in Minimum Description Length: Theory and Applications*. The MIT Press, 2005. 2.1.1, 2.1.1, 3.1, 5.1.2

[39] D. Gusfield. Optimal, efficient reconstruction of root-unknown phylogenetic networks with constrained and structured recombination. *Journal of Computer and System Sciences*, 70(3):381–398, 2005. 2.4

[40] Ryan N. Gutenkunst, Ryan D. Hernandez, Scott H. Williamson, and Carlos D. Bustamante. Inferring the joint demographic history of multiple populations from multidimensional snp frequency data. *PLoS Genetics*, 5(10):e1000695, 2009. 3.4

[41] M. F. Hammer, A. B. Spurdle, T. Karafet, M. R. Bonner, E. T. Wood, A. Novelletto, P. Malaspina, R. J. Mitchell, S. Horai, T. Jenkins, and S. L. Zegura. The geographic distribution of human Y chromosome variation. *Genetics*, 145(3):787–805, 1997. 2

[42] Michael F. Hammer. A recent common ancestry for human y chromosomes. *Nature*, 378 (6555):376–378, 1995. 3.3, 4.2.2

[43] Ross C Hardison, Krishna M Roskin, Shan Yang, Mark Diekhans, W James Kent, Ryan Weber, Laura Elnitski, Jia Li, Michael O'Connor, Diana Kolbe, and et al. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Research*, 13(1):13–26, 2003. 3.2

[44] M. J. Havey and F. J. Muehlbauer. Variability for restriction fragment lengths and phylogenies in lentil. *Theoretical and Applied Genetics*, 77(6):839–843, 1989. 1, 1.1

[45] M. He, J. Gitschier, T. Zerjal, P. de Knijff, C. Tyler-Smith, and Y. Xue. Geographical affinities of the HapMap samples. *PLoS ONE*, 4(3):e4684, 03 2009. 2.3

[46] Mika Hirakawa, Toshihiro Tanaka, Yoichi Hashimoto, Masako Kuroda, Toshihisa Takagi, and Yusuke Nakamura. Jsnp: a database of common gene variations in the japanese

population. *Nucleic Acids Research*, 30(1):158–162, 2002. 1.2

[47] R. Hudson. Generating samples under a wright-fisher neutral model of genetic variation. *Bioinformatics (Oxford, England)*, 18(2):337–338, February 2002. 2.2.1

[48] R.R. Hudson. Gene genealogies and the coalescent process. *Oxford Surveys in Evolutionary Biology*, 7:1–44, 1990. 3.2, 4.1.4, 5.1.3

[49] John P. Huelsenbeck and Fredrik Ronquist. Mrbayes: Bayesian inference of phylogenetic trees. *Bioinformatics*, 17(8):754–755, 2001. 1.3.2

[50] Mattias Jakobsson, Sonja W. Scholz, Paul Scheet, J. Raphael Gibbs, Jenna M. VanLiere, Hon-Chung Fung, Zachary A. Szpiech, James H. Degnan, Kai Wang, Rita Guerreiro, Jose M. Bras, Jennifer C. Schymick, Dena G. Hernandez, Bryan J. Traynor, Javier Simon-Sanchez, Mar Matarin, Angela Britton, Joyce van de Leemput, Ian Rafferty, Maja Bucan, Howard M. Cann, John A. Hardy, Noah A. Rosenberg, and Andrew B. Singleton. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature*, 451 (7181):998–1003, 2008. 10.1038/nature06742. 1.2, 2, 2.2.2, 4

[51] L. Jin, M.L. Baskett, L.L. Cavalli-Sforz, L.A. Zhivotovsky, M.W. Feldman, and N.A. Rosenberg. Microsatellite evolution in modern humans: a comparison of two data sets from the same populations. *Annals of Human Genetics*, 64(02):117–134, 2000. 2.4

[52] L.B. Jorde, M.J. Bamshad, W.S. Watkins, R. Zenger, Fraley. A.E., P.A. Krakowiak, K.D. Carpenter, H. Soodyall, T. Jenkins, and A.R. Rogers. Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. *American Journal of Human Genetics*, 57:523–538, 1995. 2

[53] L.B. Jorde, W.S. Watkins, and M.J. Bamshad. Population genomics: a bridge from evolutionary history to genetic medicine. *Human Molecular Genetics*, 10(20):2199–2207, 2001. 4

[54] M. Kayser, M. Krawczak, L. Excoffier, P. Dieltjes, D. Corach, V. Pascali, C. Gehrig,

L.F. Bernini, J. Jespersen, E. Bakker, L. Roewer, and P. de Knijff. An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *American Journal of Human Genetics*, 68(4):990–1018, 2001. 2.3

[55] James Kent, Charles Sugnet, Terrence Furey, Krishna Roskin, Tom Pringle, Alan Zahler, and And Haussler. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002. 1.2

[56] Jan O. Korbel, Alexander Eckehart Urban, Jason P. Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M. Kim, Dean Palejev, Nicholas J. Carriero, Lei Du, Bruce E. Taillon, Zhoutao Chen, Andrea Tanzer, A. C. Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P. Carter, Matthew E. Hurles, Sherman M. Weissman, Timothy T. Harkins, Mark B. Gerstein, Michael Egholm, and Michael Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 2007. 1.1

[57] 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. 3

[58] Sudhir Kumar and Sankar Subramanian. Mutation rates in mammalian genomes. *Proceedings of the National Academy of Sciences*, 99(2):803–808, 2002. 3.3

[59] 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. 3.4

[60] Cathryn M. Lewis and Jo Knight. Introduction to genetic association studies. *Cold Spring Harbor Protocols*, 2012(3):pdb.top068163, 2012. 5

[61] Heng Li and Richard Durbin. Inference of human population history from individual

whole-genome sequences. *Nature*, 475(7357):493–496, 2011. 3, 3.4, 4.3

[62] George Liu, Lakshmi Matukumalli, Tad Sonstegard, Larry Shade, and Curtis Van Tassell. Genomic divergences among cattle, dog and human estimated from large-scale alignments of genomic sequences. *BMC Genomics*, 7(1):140, 2006. 3.2, 3.3

[63] Jeffrey C. Long and Peter E. Smouse. Intertribal gene flow between the Ye'cuana and Yanomama: Genetic analysis of an admixed village. *American Journal of Physical Anthropology*, 61(4):411–422, 1983. 4

[64] Matthew D. Mailman, Michael Feolo, Yumi Jin, Masato Kimura, Kimberly Tryka, Rinat Bagoutdinov, Luning Hao, Anne Kiang, Justin Paschall, Lon Phan, Natalia Popova, Stephanie Pretel, Lora Ziyabari, Moira Lee, Yu Shao, Zhen Y. Wang, Karl Sirotkin, Minghong Ward, Michael Kholodov, Kerry Zbicz, Jeffrey Beck, Michael Kimelman, Sergey Shevelev, Don Preuss, Eugene Yaschenko, Alan Graeff, James Ostell, and Stephen T. Sherry. The ncbi dbgap database of genotypes and phenotypes. *Nat Genet*, 39(10):1181–1186, 2007. 1.2

[65] T. Margush and F.R. Mcmorris. Consensus n-trees. *Bulletin of Mathematical Biology*, 43: 239–244, 1981. 2.1.1

[66] Gabriela Martinez-Cortes, Joel Salazar-Flores, Laura Gabriela Fernandez-Rodriguez, Rodrigo Rubi-Castellanos, Carmen Rodriguez-Loya, Jesus Salvador Velarde-Felix, Jose Francisco Munoz-Valle, Isela Parra-Rojas, and Hector Rangel-Villalobos. Admixture and population structure in mexican-mestizos based on paternal lineages. *Journal of Human Genetics*, 2012. 3.3, 3.4, 4.3

[67] M. Meila. Comparing clusterings–an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895, 2007. 2.2.3

[68] Marta Mel, Asif Javed, Marc Pybus, Pierre Zalloua, Marc Haber, David Comas, Mihai G. Netea, Oleg Balanovsky, Elena Balanovska, Li Jin, Yajun Yang, RM. Pitchap-

pan, G. Arunkumar, Laxmi Parida, Francesc Calafell, Jaume Bertranpetit, and The Genographic Consortium. Recombination gives a new insight in the effective population size and the history of the old world human populations. *Molecular Biology and Evolution*, 2011. 3.1

[69] Laurence R. Meyer, Ann S. Zweig, Angie S. Hinrichs, Donna Karolchik, Robert M. Kuhn, Matthew Wong, Cricket A. Sloan, Kate R. Rosenbloom, Greg Roe, Brooke Rhead, Brian J. Raney, Andy Pohl, Venkat S. Malladi, Chin H. Li, Brian T. Lee, Katrina Learned, Vanessa Kirkup, Fan Hsu, Steve Heitner, Rachel A. Harte, Maximilian Haeussler, Luvina Guruvadoo, Mary Goldman, Belinda M. Giardine, Pauline A. Fujita, Timothy R. Dreszer, Mark Diekhans, Melissa S. Cline, Hiram Clawson, Galt P. Barber, David Haussler, and W. James Kent. The ucsc genome browser database: extensions and updates 2013. *Nucleic Acids Research*, 2012. 1.2

[70] C. D. Michener and R. R. Sokal. A quantitative approach to a problem in classification. *Evolution*, 11(2):130–162, 1957. 1

[71] Yusuke Nakamura. Jsnp database. Online. `http://snp.ims.u-tokyo.ac.jp/`. 1.1

[72] NCBI. Database of single nucleotide polymorphisms. Online. `www.ncbi.nlm.nih.gov/snp`. 1.2

[73] M. Nei and S. Kumar. *Molecular Evolution and Phylogenetics*. Oxford University Press, 2000. 2, 2.1.1, 3.1, 4

[74] M. Nei and A.K. Roychoudhury. Genetic relationship and evolution of human races. *Evolutionary Biology*, 14:1–59, 1982. 2

[75] Matthew R. Nelson. Popres: Population reference sample. Online. `http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000145.v1.p1`. 1.1

[76] Matthew R. Nelson, Katarzyna Bryc, Karen S. King, Amit Indap, Adam R. Boyko, John Novembre, Linda P. Briley, Yuka Maruyama, Dawn M. Waterworth, Grard Waeber, Peter Vollenweider, Jorge R. Oksenberg, Stephen L. Hauser, Heide A. Stirnadel, Jaspal S. Kooner, John C. Chambers, Brendan Jones, Vincent Mooser, Carlos D. Bustamante, Allen D. Roses, Daniel K. Burns, Margaret G. Ehm, and Eric H. Lai. The population reference sample, popres: A resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics*, 83(3):347–358, 2008. 1.2, 2, 4

[77] R. Nielsen and J. Wakeley. Distinguishing migration from isolation: A markov chain monte carlo approach. *Genetics*, 158(2):885–896, 2001. 1.4, 3, 3.1, 3.2, 3.4, 4.1, 4.1.4, 4.3

[78] Rasmus Nielsen, Melissa J. Hubisz, and Andrew G. Clark. Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. *Genetics*, 168(4):2373–2382, 2004. 4.1.5

[79] R. D. M. Page and E. C. Holmes. *Molecular evolution. A phylogenetic approach*. Blackwell Science Ltd., 1998. 1

[80] Esteban J. Parra, Amy Marcini, Joshua Akey, Jeremy Martinson, Mark A. Batzer, Richard Cooper, Terrence Forrester, David B. Allison, Ranjan Deka, Robert E. Ferrell, and Mark D. Shriver. Estimating african american admixture proportions by use of population-specific alleles. *American Journal of Human Genetics*, 63(6):1839–1851, 1998. 3, 4

[81] Nick Patterson, Alkes L. Price, and David Reich. Population structure and eigenanalysis. *PLoS Genet*, 2(12):e190, 2006. 1.3.1, 1.4, 2

[82] Daniel Pinkel and Donna G. Albertson. Array comparative genomic hybridization and its applications in cancer. *Nat Genet*, 37:S11–S17, 2005. 1.1

[83] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A.

Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–909, 2006. 10.1038/ng1847. 5, 5.3

[84] Alkes L. Price, Arti Tandon, Nick Patterson, Kathleen C. Barnes, Nicholas Rafaels, Ingo Ruczinski, Terri H. Beaty, Rasika Mathias, David Reich, and Simon Myers. Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS Genet*, 5(6):e1000519, 2009. 1.4, 1.5, 3, 4

[85] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000. 1.3.1, 1.5, 2, 2.2.3, 3, 4

[86] Jonathan K. Pritchard, Matthew Stephens, Noah A. Rosenberg, and Peter Donnelly. Association mapping in structured populations. *The American Journal of Human Genetics*, 67(1):170–181, 2000. 5, 5.3

[87] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel A. R. Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul I. W. de Bakker, Mark J. Daly, and Pak C. Sham. Plink: A tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics*, 81(3):559–575, 2007. 5.1.3

[88] David C. Queller, Joan E. Strassmann, and Colin R. Hughes. Microsatellites and kinship. *Trends in Ecology and Evolution*, 8(8):285–288, 1993. 1.1

[89] A Rambaut and N.C. Grass. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.*, 13(3):235–238, 1997. 2.2.1

[90] H. Rangel-Villalobos, J. F. Muoz-Valle, A. Gonzlez-Martn, A. Gorostiza, M. T. Magaa, and L. A. Pez-Riberos. Genetic admixture, relatedness, and structure patterns among mexican populations revealed by the y-chromosome. *American Journal of Physical An-*

thropology, 135(4):448–461, 2008. 3.3

[91] D. Reich, K. Thangaraj, N. Patterson, A.L. Price, and L. Singh. Reconstructing indian population history. *Nature*, 461(7263):489–494, 2009. 2.3

[92] D.E. Reich and D.B. Goldstein. Genetic evidence for a Paleolithic human population expansion in Africa. *Proceedings of the National Academy of Sciences of the United States of America*, 95(14):8119–8123, 1998. 2.4

[93] John Reynolds, B. S. Weir, and C. Clark Cockerham. Estimation of the coancestry coefficient: Basis for a short-term genetic distance. *Genetics*, 105(3):767–779, 1983. 4.1.4

[94] Neil Risch and Kathleen Merikangas. The future of genetic studies of complex human diseases. *Science*, 273(5281):1516–1517, 1996. 5

[95] Noah Rosenberg. The human genome diversity project. Online. `http://www.stanford.edu/group/rosenberglab/diversity.html`. 1.1

[96] Sridhar S. Mixed integer linear programming for maximum-parsimony phylogeny inference. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5(3): 323–331, 2008. 1.3.2, 2

[97] R. K. Saiki, S. Scharf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich, and N. Arnheim. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. *Science*, 230(4732):1350–4, 1985. 1, 1.1

[98] Sriram Sankararaman, Srinath Sridhar, Gad Kimmel, and Eran Halperin. Estimating local ancestry in admixed populations. *American journal of human genetics*, 82(2):290–303, 2008. 1.4, 1.5, 3, 4

[99] S. T. Sherry, M. Ward, and K. Sirotkin. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res*, 9(8):677–9, 1999. 1.2, 4.1.5

[100] S. T. Sherry, M. H. Ward, M. Kholodov, J. Baker, L. Phan, E. M. Smigielski, and

K. Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucl. Acids Res.*, 29(1):308–311, 2001. 2

[101] S. Shringarpure and E.P. Xing. mstruct: Inference of population structure in light of both genetic admixing and allele mutations. *Genetics*, 108:100222, 2009. 1.3.1, 1.5, 2.4

[102] M.D. Shriver and R.A. Kittles. Genetic ancestry and the search for personalized genetic histories. *Nature Reviews Genetics*, 5:611–618, 2004. 2.2.3, 2.3, 2.3, 4.3

[103] M Slatkin. A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, 139(1):457–62, 1995. 4

[104] Oliver Smithies. *How It All Began: A Personal History of Gel Electrophoresis*, volume 869 of *Methods in Molecular Biology*, chapter 1, pages 1–21. Humana Press, 2012. 1

[105] K.A. Sohn and E.P. Xing. Spectrum: joint bayesian inference of population structure and recombination events. *Bioinformatics*, 23(13):i479–489, 2007. 1.3.1, 2.2.3

[106] R. R. Sokal and C. D. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958. 4.1.4

[107] K. M. Song, T. C. Osborn, and P. H. Williams. Brassica taxonomy based on nuclear restriction fragment length polymorphisms (rflps). *Theoretical and Applied Genetics*, 75 (5):784–794, 1988. 1, 1.1

[108] Pawel Stankiewicz and James R. Lupski. Structural variation in the human genome and its role in disease. *Annual Review of Medicine*, 61(1):437–455, 2010. 1.1

[109] David L. Swofford. Paup*: phylogenetic analysis using parsimony, version 4.0b10. Software, 2002. `www.sinauer.com/detail.php?id=8060`. 1.3.2

[110] Hua Tang, Jie Peng, Pei Wang, and Neil J. Risch. Estimation of individual admixture: Analytical and study design considerations. *Genetic Epidemiology*, 28(4):289–301, 2005. 1.4

[111] Hua Tang, Marc Coram, Pei Wang, Xiaofeng Zhu, and Neil Risch. Reconstructing genetic

ancestry blocks in admixed individuals. *American journal of human genetics*, 79(1):1–12, 2006. 1.4, 1.5

[112] Hua Tang, Shweta Choudhry, Rui Mei, Martin Morgan, William Rodriguez-Cintron, Esteban Gonzlez Burchard, and Neil J. Risch. Recent genetic selection in the ancestral admixture of puerto ricans. *American Journal of Human Genetics*, 81(3):626–633, 2007. 3.3, 3.4, 4.3

[113] Albert Tenesa, Pau Navarro Ben J. Hayes, David L. Duffy, Geraldine M. Clarke, Mike E. Goddard, and Peter M. Visscher. Recent human effective population size estimated from linkage disequilibrium. *Genome Research*, 17(4):520–526, 2007. 3.1

[114] D.C. Thomas and J.S. Witte. Point: Population stratification: A problem for case-control studies of candidate-gene associations? *Cancer Epidemiol Biomarkers Prev*, 11(6):505–512, 2002. 2

[115] Chao Tian, Peter K. Gregersen, and Michael F. Seldin. Accounting for ancestry: population substructure and genome-wide association studies. *Human Molecular Genetics*, 17 (R2), 2008. 5

[116] S. A. Tishkoff, E. Dietzsch, W. Speed, A. J. Pakstis, J. R. Kidd, K. Cheung, B. Bonn-Tamir, A. S. Santachiara-Benerecetti, P. Moral, M. Krings, S. Pbo, E. Watson, N. Risch, T. Jenkins, and K. K. Kidd. Global patterns of linkage disequilibrium at the cd4 locus and modern human origins. *Science*, 271(5254):1380–1387, 1996. 2

[117] S.A. Tishkoff and B.C. Verrelli. Patterns of human genetic diversity: Implications for human evolutionary history and disease. *Annual Review of Genomics and Human Genetics*, 4(1):293–340, 2003. 2.3, 2.4, 4

[118] S.A. Tishkoff and S.M. Williams. Genetic analysis of African populations: human evolution and complex disease. *Nat Rev Genet*, 3(8):611–621, 2002. 2.3

[119] Ming-Chi Tsai, Guy E. Blelloch, R. Ravi, and Russell Schwartz. A consensus tree ap-

proach for reconstructing human evolutionary history and detecting population substructure. *IEEE/ACM Trans. Comput. Biol. Bioinformatics*, 8:918–928, July 2011. 1, 3.1, 3.1, 4.1

[120] Ming-Chi Tsai, Guy Blelloch, R. Ravi, and Russell Schwartz. Coalescent-based method for learning parameters of admixture events from large-scale genetic variation data. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '12, pages 90–97, New York, NY, USA, 2012. ACM. 1

[121] Ming-Chi Tsai, Guy Blelloch, R. Ravi, and Russell Schwartz. Coalescent-based method for learning parameters of admixture events from large-scale genetic variation data [extended journal version]. Under Review, 2013. 1, 4, 4.1, 4.1.2, 4.1.4

[122] Ming-Chi Tsai, Guy Blelloch, R. Ravi, and Russell Schwartz. Coalescent-based method for joint estimation of population history, time, and admixture. Submitted, 2013. 1

[123] J. Craig Venter, Mark D. Adams, Granger G. Sutton, Anthony R. Kerlavage, Hamilton O. Smith, and Michael Hunkapiller. Shotgun sequencing of the human genome. *Science*, 280 (5369):1540–1542, 1998. 1

[124] J.C. Venter, M.A. Adams, and Eugene W. Myers *et al.* The sequence of the human genome. *Science*, 291(5507):1304–1351, 2001. 2

[125] J. Wang. Maximum-likelihood estimation of admixture proportions from genetic data. *Genetics*, 164(2):747–765, 2003. 4, 4.1

[126] J. Wang. A coalescent-based estimator of admixture from dna sequences. *Genetics*, 173 (3):1679–1692, 2006. 1.4, 1.5, 3, 3.1, 3.2, 4, 4.1.4

[127] William Y. S. Wang, Bryan J. Barratt, David G. Clayton, and John A. Todd. Genome-wide association studies: theoretical and practical concerns. *Nat Rev Genet*, 6(2):109–118, 2005. 5

[128] Joachim Weischenfeldt, Orsolya Symmons, Francois Spitz, and Jan O. Korbel. Phenotypic

impact of genomic structural variation: insights from and for human disease. *Nat Rev Genet*, 14(2):125–138, 2013. 10.1038/nrg3373. 1.1

[129] Carl R. Woese and George E. Fox. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11):5088–5090, 1977. 1

[130] Junshi Yazaki, Brian D. Gregory, and Joseph R. Ecker. Mapping the genome landscape using tiling array technology. *Current Opinion in Plant Biology*, 10(5):534–542, 2007. 1

[131] D. V. Zaykin, P. H. Westfall, S. S. Young, M. A. Karnoub, M. J. Wagner, and M. G. Ehm. Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity*, 53(2):79–91, 2002. 5

[132] L.A. Zhivotovsky. Estimating Divergence Time with the Use of Microsatellite Genetic Distances: Impacts of Population Growth and Gene Flow. *Mol Biol Evol*, 18(5):700–709, 2001. 2.4, 3.3, 3.4, 4.3