A Consensus Tree Approach for Reconstructing Human Evolutionary History and Detecting Population Substructure

Ming-Chi Tsai, Guy Blelloch, R. Ravi, and Russell Schwartz

Abstract—The random accumulation of variations in the human genome over time implicitly encodes a history of how human populations have arisen, dispersed, and intermixed since we emerged as a species. Reconstructing that history is a challenging computational and statistical problem but has important applications both to basic research and to the discovery of genotype-phenotype correlations. We present a novel approach to inferring human evolutionary history from genetic variation data. We use the idea of consensus trees, a technique generally used to reconcile species trees from divergent gene trees, adapting it to the problem of finding robust relationships within a set of intraspecies phylogenies derived from local regions of the genome. Validation on both simulated and real data shows the method to be effective in recapitulating known true structure of the data closely matching our best current understanding of human evolutionary history. Additional comparison with results of leading methods for the problem of population substructure assignment verifies that our method provides comparable accuracy in identifying meaningful population subgroups in addition to inferring relationships among them. The consensus tree approach thus provides a promising new model for the robust inference of substructure and ancestry from large-scale genetic variation data.

Index Terms—Biology and genetics, trees, information theory, graph algorithms.

1 Introduction

Understanding how modern human populations arose from our common ancestors is one of the central questions of human genetics. The completion of the human genome [1], [2], and the subsequent discovery of millions of common genetic variations in the human genome [3] has created an unprecedented opportunity to address this question. Several major studies have recently been undertaken to assess genetic variation in human population groups and enable detailed reconstruction of the ancestry of human population groups [4], [5], [6], [7]. In addition to its importance as a basic research problem, this topic 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 [8].

Past work on human ancestry inference has essentially treated it as two distinct problems: identifying meaningful population groups and inferring evolutionary trees among them. Population groups may be assumed in advance based on common conceptions of ethnic groupings, although the field increasingly depends on computational analysis to make such inferences automatically. Probably, the most well-known system for this problem is STRUCTURE [9], which uses a Markov Chain Monte Carlo (MCMC) method to group sequences into K ancestral population group each with its own allele frequency profile. Another well-known program for this problem is EIGENSOFT [10], 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 [11] and mStruct [12] 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 recombination and mutation 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 propose a new admixture model to identify subgroups by representing each population as *mixtures of ancestral alleles* rather than a single ancestral allele profile.

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 [13], mtDNA genotypes [14], [15], short tandem repeats [14], [16], and Y chromosome polymorphism [17]—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 data sets with little recombination, most do not work well when utilizing

- *M.-C. Tsai is with the Joint Carnegie Mellon University/University of Pittsburgh PhD Program in Computational Biology and Lane Center for Computational Biology, 4400 Fifth Avenue, Pittsburgh, PA 15213. E-mail: mingchit@andrew.cmu.edu.*
- *G. Blelloch is with the Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: guyb@cs.cmu.edu.*
- *R. Ravi is with the Tepper School of Business, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: ravi@cmu.edu.*
- *R. Schwartz is with the Department of Biological Sciences, Carnegie Mellon University, Pittsburgh, PA 15213. E-mail: russells@andrew.cmu.edu.*

Manuscript received 20 July 2010; revised 25 Sept. 2010; accepted 3 Oct. 2010; published online 27 Jan. 2011.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-2010-07-0174.
Digital Object Identifier no. 10.1109/TCBB.2011.23.

genome wide data sets. Furthermore, there has thus far been little crosstalk 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.

We propose a novel approach for reconstructing a species history that is intended to unify these two inference problems. The method is conceptually based on the idea of consensus trees [18], 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 [19], has made it possible to infer likely phylogenies on millions of small regions spanning the human genome. The intuition behind our 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.

In the remainder of this paper, we describe and evaluate our approach. We first present in more detail our mathematical model of the consensus tree problem and a set of algorithms for finding consensus trees from families of local phylogenies. We next evaluate our method on a set of simulated data and two real data sets from the HapMap Phase II [4] and the Human Genome Diversity Project [5]. Finally, we consider some of the implications of the results and future prospects of the consensus tree approach for evolutionary history and substructure inference.

2 Methods

2.1 Consensus Tree Model

We assume that 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 $T_1, \ldots, T_n$.

Our objective function for choosing T_M is based on the task of finding a consensus tree [18] 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 our phylogenies are derived from many variant markers, each only minimally informative, within a single species. Standard consensus tree approaches, such as majority consensus [20] or Adam consensus [21], would not be expected to be effective in this situation as it is likely that 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.

For this purpose, we chose a model of the problem based on the principle of minimum description length (MDL) [22], 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, $\mathcal{D} = \{T_1, T_2, \ldots, T_n\}$ 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 we compute the description length of a tree. For this purpose, we 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. We represent the observed trees and candidate consensus trees as sets of bipartitions for the purpose of calculating description lengths. Once we have identified a set of bipartitions representing the desired consensus tree, we 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. 1, 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 [22] 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:

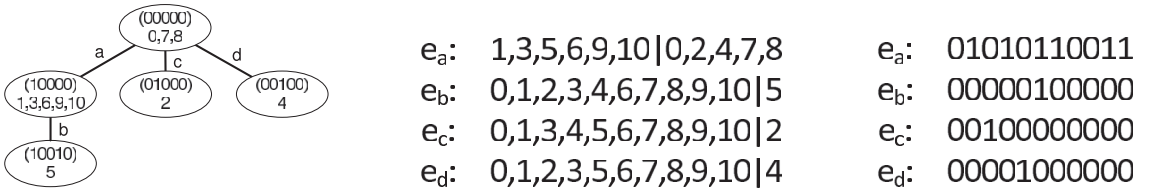


Fig. 1. (a) A maximum parsimony (MP) tree consisting of 11 labeled individuals or haplotypes. (b) The set of bipartitions induced by edges (e_a, e_b, e_c, e_d) in the tree. (c) 0-1 bit sequence representation for each bipartition.

$$L(b) = mH(b) = m(-p_0 \log_2 p_0 - p_1 \log_2 p_1).$$

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:

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

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:

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

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

2.2 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, 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

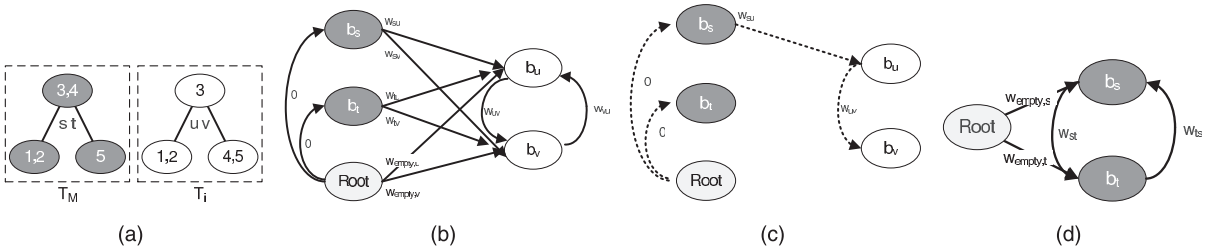


Fig. 2. Illustration of the DMST construction for determining model description length. (a) Hypothetical model tree T_M (gray) and observed tree T_i (white). (b) Graph of possible reference relationships for explaining T_i (white nodes) by reference to T_M (gray nodes). (c) A possible resolution of the graph of (b). (d) Graph of possible reference relationships for explaining T_M by itself.

B_M, representing the fact that the presumed cost of the model tree has already been computed. Fig. 2 illustrates the construction for a hypothetical model tree T_M and observed tree T_i (Fig. 2a), showing the graph of possible reference relationships (Fig. 2b), a possible solution corresponding to a specific explanation of T_i in terms of T_M (Fig. 2c), and the graph of possible reference relationships for T_M by itself (Fig. 2d).

Given the graph construction, the minimum encoding length for both constructions is found by solving for the DMST with the algorithm of Chu and Liu [23] and summing the weights of the edges. This cost is computed for a candidate model tree T_M and for each observed tree T_i, for $i = 1, \ldots, n$, 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 $\mathcal{L}(T_M, T_1, \ldots, T_n)$ 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 (nonunique) 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 the 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 $b_1 \prec b_2 \cdots \prec b_k$ in decreasing order of numbers of splits they explain (i.e., the number of out-edges from their corresponding nodes in the DMST). We 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, we 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. We 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. We 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 then the final consensus tree.

2.3 Validation Experiments

Simulated data set. We first evaluated our method on a simulated data set consisting of three independent populations, each with 150 individuals (300 chromosomes). We generated the genealogy trees for each population using the coalescent simulator MS [24] 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. 3 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 [25]. 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 five consecutive SNPs spanning across the sequences. Given the data set, we ran the algorithms on 10,000 randomly selected trees or their corresponding 33,295 unique SNPs.

Real data. We further evaluated our method by applying it to samples from two real SNP variation data sets. 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 from 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 five

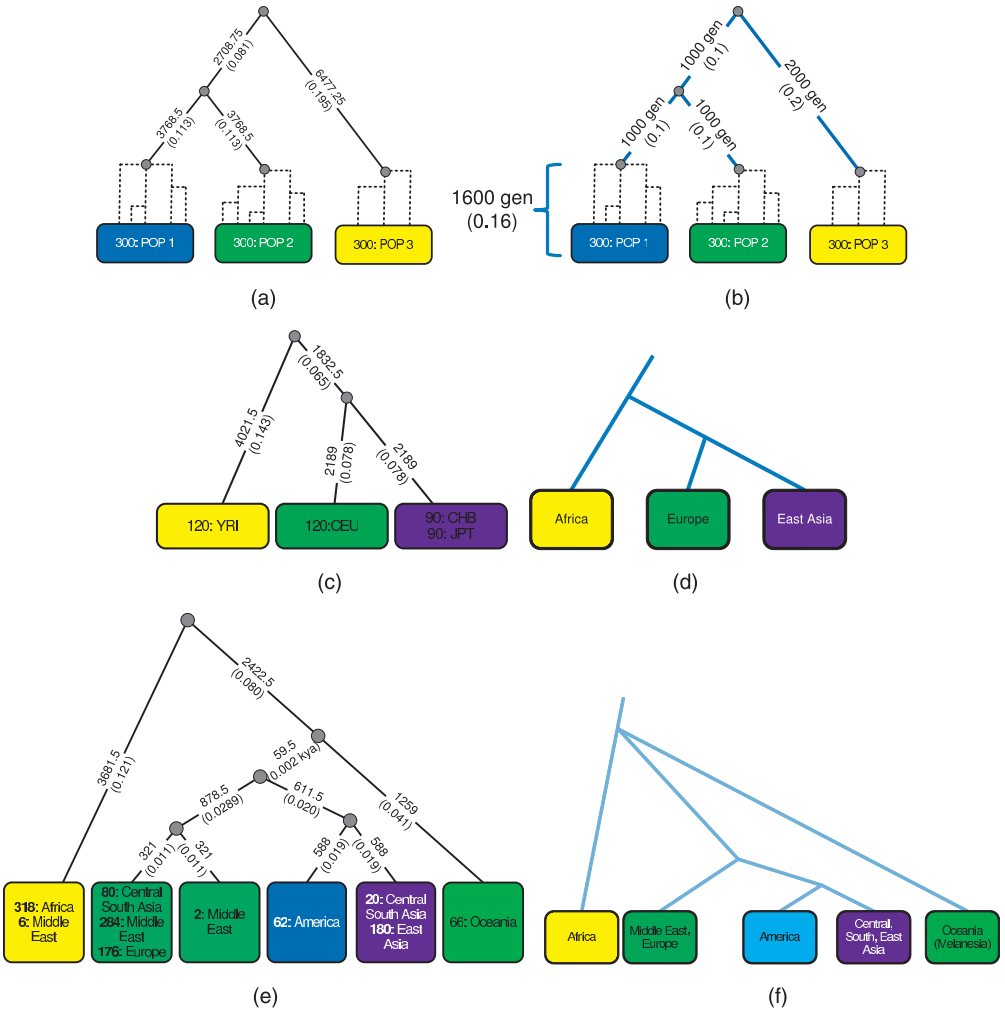


Fig. 3. Inferred consensus trees. Node labels show numbers of haplotypes belonging to each known population. Edges in inferred trees are labeled by the number of splits assigned to each and, in parentheses, the fraction of all splits assigned to each. For the simulated gold standard tree, edges are labeled 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 data set. (d) Trimmed and condensed tree from [26]. (e) Consensus tree obtained from the HGDP data set. (f) Trimmed and condensed tree from [26].

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 our method and the comparative algorithms. We next used phased data (version 1.3) from the Human Genome Diversity Project (HGDP) [5], which genotyped 525,910 SNP sites in 597 individuals from 29 populations categorized into seven regions 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.

Benchmarks. We are not aware of any comparable method to ours and therefore cannot directly benchmark it against any competitor. We therefore assessed it 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 [26], 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 our method. To allow comparison with either of our 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 data set, 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) [9] and Spectrum [11]. We selected these

programs because they are well accepted as leading methods for the substructure problem and are able to handle comparable sizes of data set to our method. We chose to omit EIGENSOFT, despite its wide usage 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 data set 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 our 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 [27], a measure commonly used to assess accuracy of a clustering method relative to a predefined "ground truth." Variation of information (VI) 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 return 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 an 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.

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 our method, STRUCTURE, and Spectrum for four 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.

3 Results

Fig. 3 shows the trees inferred by our method on the simulated data and the two real data sets alongside their corresponding true simulated tree or the condensed Shriver and Kittles "gold standard" trees. Fig. 3a shows the inferred tree produced by our model on the simulated data set. Based on the numbers of observed bipartitions explained by each model bipartition, the tree reconstruction correctly infers the key divergence events across the three populations when compared to Fig. 3b. 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. 3c shows the inferred tree from the HapMap data set. The tree reconstruction infers there to be an initial separation of the YRI (African) subpopulation from the others ($\mathrm{CEU} + \mathrm{JPT} + \mathrm{CHB}$) followed by a subsequent separation of CEU (European) from $\mathrm{JPT} + \mathrm{CHB}$ (East Asian). When collapsed to the same three populations (African, European, and East Asian), the gold standard tree (Fig. 3d) shows an identical structure. Furthermore, these results are consistent with many independent lines of evidence for the out-of-Africa hypothesis of human origins [26], [28], [29]. 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 data set, the trees differ slightly from run to run, so we arbitrarily provide our first run, Fig. 3e, as a

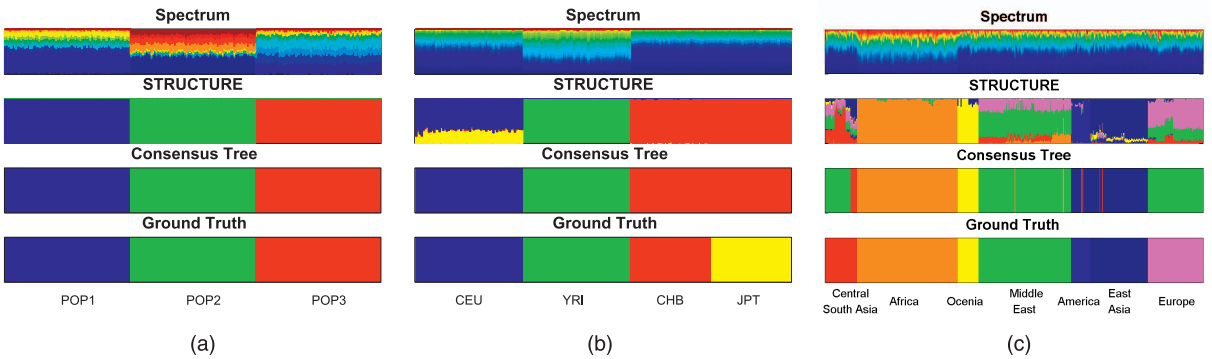


Fig. 4. Inferred population structures. Each colored vertical line shows the assigned population(s) for a single sequence for one method. From top to bottom: Spectrum (with colors representing fractional assignments to distinct ancestral haplotypes), STRUCTURE (with colors representing probabilities 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 the HapMap Phase II data set. (c) Inferred population structures from one trial of 10,000 trees from the HGDP data set.

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 $\text{Asian} + \text{American}$ from $\text{European} + \text{Middle Eastern}$ (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 that derived from Shriver and Kittles for the same population groups (Fig. 3f). The only notable distinctions are that gold standard tree has no equivalent to our purely Middle Eastern node; that the gold standard does not distinguish between the divergence times of Oceanian and other non-African populations from the African, while ours 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 ours splits Central South Asian groups between European and East Asian subgroups (an interpretation supported by more recent analyses [30]). Our 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 [28], [31]. 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 suggest a proportionally longer gap between the divergence of African from non-African subgroups and further divergence between the non-African subgroups.

Fig. 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 subproblem of subpopulation inference. Note that STRUCTURE and our 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. 4a), all of the methods were able to separate the three population groups. For HapMap (Fig. 4b), all three methods consistently identified YRI and CEU as distinct subpopulations but failed to separate CHB and JPT.

Results were more ambiguous for HGDP (Fig. 4c). 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. [32]. 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 our method and STRUCTURE by converting the result to the most likely cluster assignment and computing VI scores and inconsistency scores. Fig. 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, our 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. 6). In all cases, our method consistently ran faster than both STRUCTURE and Spectrum, which both use similar Gibbs sampling approaches.

Fig. 7 shows the consensus trees constructed using different sizes of data set 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

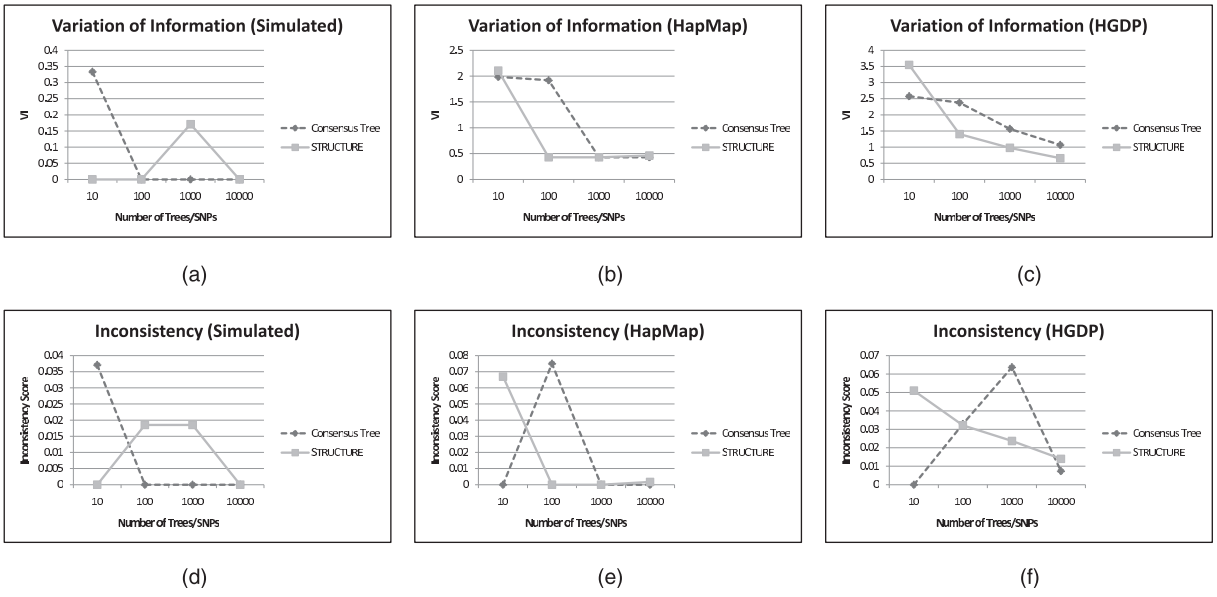


Fig. 5. Variation of information and inconsistency scores. Lower VI reflects higher accuracy in identifying known population structure. Lower inconsistency reflects greater reproducibility between independent samples.

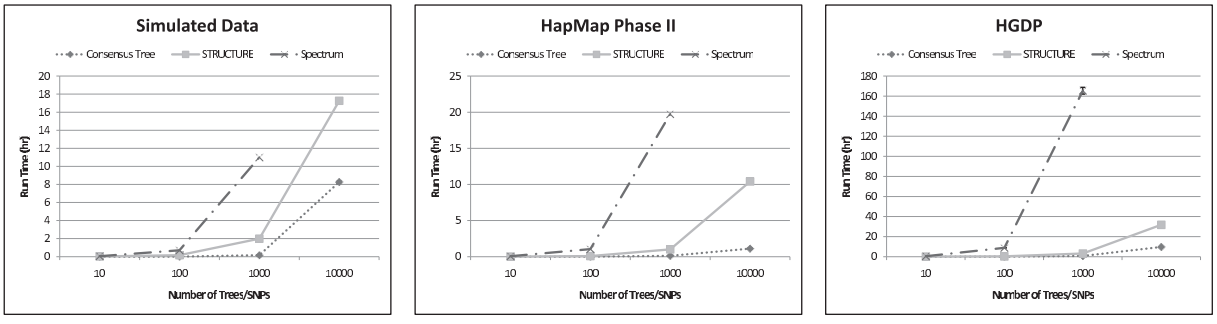


Fig. 6. Average runtime of the algorithms on different data sets and different data set sizes.

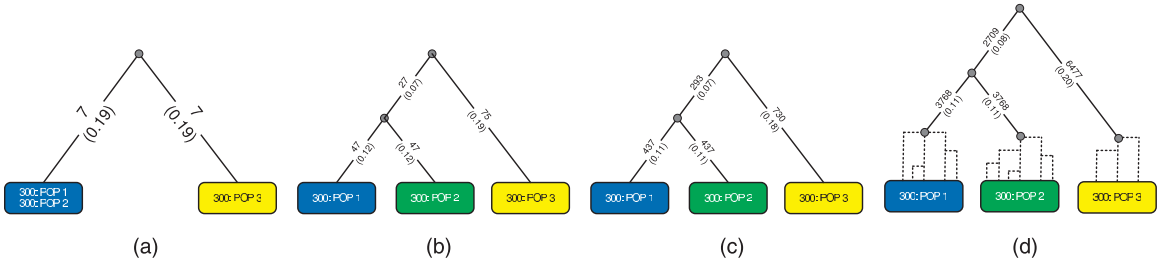


Fig. 7. Consensus trees produced using varying numbers of input trees. Node labels show numbers of haplotypes belonging to each simulated population. Edges are labeled by the number of splits assigned to each and, in parentheses, the fraction of all splits assigned to each. From left to right: Consensus Tree from (a) 10, (b) 100, (c) 1,000, and (d) 10,000 observed trees.

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.

4 Discussion

While population substructure inference is only one facet of the problem solved by our method, it nonetheless provides for a convenient partial validation. Comparison with leading population substructure algorithms shows that our method provides very good performance on the substructure problem. Our 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. Our method is also quite competitive on runtime with these alternatives, although other substructure methods that were not amenable to a direct comparison, such as mStruct [12], can yield substantially superior runtimes for closely related analyses. Our method also shows an ability to automatically adjust to varying amounts of data while avoiding overfitting, as demonstrated by the consistency scores, as would be expected for the chosen MDL approach.

One key advantage of our 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. Our 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 data sets, despite different individuals, populations represented, and markers selected.

Our 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 toward 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 [29], 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 [33], [34], [35].

While the results show that our methods are capable of making robust but sensitive inferences of population structure as well as tree structure, our method does nonetheless have some significant limitations. One such limitation is runtime; while our method is superior in this regard to STRUCTURE and Spectrum, its runtime is still considerable and far worse than 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 runtime is largely due to the many calls our 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, our computations should be easily amenable to parallelization.

Another limitation, noted above, is that our 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 principle, then, our 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 [36]. 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.

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. Our method also provides edge length estimates that can roughly be interpreted as relative times between divergence events, although there appear to be some biases in these estimates. It may be possible to translate these relative times into estimates of absolute elapsed times given more detailed population genetic models or independent sources of data about absolute times along one or more edges of the trees. 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. In future work, we hope to better test these assumptions, in part by developing more accurate models for estimating the branch lengths, and to extend the method to inferences of ancestry in the presence of admixture.

ACKNOWLEDGMENTS

The authors would like to thank Srinath Sridhar for valuable discussions on the ideas behind this work. This work was supported by US National Science Foundation (NSF), IIS award #0612099, and by NIH T32 training grant T32 EB009403 as part of the HHMI-NIBIB Interfaces Initiative.

REFERENCES

[1] F.S. Collins, M. Morgan, and A. Patrinos, "The Human Genome Project: Lessons from Large-Scale Biology," *Science,* vol. 300, no. 5617, pp. 286-290, Apr. 2003.

[2] J. Venter et al., "The Sequence of the Human Genome," *Science,* vol. 291, no. 5507, pp. 1304-1351, 2001.

[3] 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," *Nucleic Acids Research,* vol. 29, no. 1, pp. 308-311, 2001.

[4] K.A. Frazer et al., "A Second Generation Human Haplotype Map of over 3.1 Million Snps," *Nature,* vol. 449, no. 7164, pp. 851-861, Oct. 2007.

[5] M. Jakobsson, S. Scholz, P. Scheet, R. Gibbs, J. Vanliere, H. Fung, Z. Szpiech, J. Degnan, K. Wang, R. Guerreiro, J. Bras, J. Schymick, D. Hernandez, B. Traynor, J. Simon-Sanchez, M. Matarin, A. Britton, J. van de Leemput, I. Rafferty, M. Bucan, H. Cann, J. Hardy, N. Rosenberg, and A. Singleton, "Genotype, Haplotype and Copy-Number Variation in Worldwide Human Populations," *Nature,* vol. 451, no. 7181, pp. 998-1003, Feb. 2008.

[6] D. Behar, S. Rosset, J. Blue-Smith, O. Balanovsky, S. Tzur, D. Comas, R. Mitchell, L. Quintana-Murci, C. Tyler-Smith, and R. Wells, "The Genographic Project Public Participation Mitochondrial DNA Database," *PLoS Genetics,* vol. 3, no. 6, p. e104, 2007.

[7] M. Nelson, K. Bryc, K. King, A. Indap, A. Boyko, J. Novembre, L. Briley, Y. Maruyama, D. Waterworth, G. Waeber, P. Vollenweider, J. Oksenberg, S. Hauser, H. Stirnadel, J. Kooner, J. Chambers, B. Jones, V. Mooser, C. Bustamante, A. Roses, D. Burns, M. Ehm, and E. Lai, "The Population Reference Sample, Popres: A Resource for Population, Disease, and Pharmacological Genetics Research," *Am. J. Human Genetics,* vol. 83, no. 3, pp. 347-358, 2008.

[8] D. Thomas and J. Witte, "Point: Population Stratification: A Problem for Case-Control Studies of Candidate-Gene Associations?," *Cancer Epidemiology Biomarkers and Prevention,* vol. 11, no. 6, pp. 505-512, 2002.

[9] J. Pritchard, M. Stephens, and P. Donnelly, "Inference of Population Structure Using Multilocus Genotype Data," *Genetics,* vol. 155, no. 2, pp. 945-959, June 2000.

[10] N. Patterson, A. Price, and D. Reich, "Population Structure and Eigenanalysis," *PLoS Genetics,* vol. 2, no. 12, pp. e190+, Dec. 2006.

[11] K. Sohn and E. Xing, "Spectrum: Joint Bayesian Inference of Population Structure and Recombination Events," *Bioinformatics,* vol. 23, no. 13, pp. i479-i489, 2007.

[12] S. Shringarpure and E. Xing, "Mstruct: Inference of Population Structure in Light of both Genetic Admixing and Allele Mutations," *Genetics,* vol. 108, pp. 575-593, 2009.

[13] M. Nei and A. Roychoudhury, "Genetic Relationship and Evolution of Human Races," *Evolutionary Biology,* vol. 14, pp. 1-59, 1982.

[14] L. Jorde, M. Bamshad, W. Watkins, R. Zenger, A.E. Fraley, P. Krakowiak, K. Carpenter, H. Soodyall, T. Jenkins, and A. Rogers, "Origins and Affinities of Modern Humans: A Comparison of Mitochondrial and Nuclear Genetic Data," *Am. J. Human Genetics,* vol. 57, pp. 523-538, 1995.

[15] R. Cann, M. Stoneking, and A. Wilson, "Mitochondrial DNA and Human Evolution," *Nature,* vol. 325, no. 6099, pp. 31-36, 1987.

[16] 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,* vol. 271, no. 5254, pp. 1380-1387, 1996.

[17] 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,* vol. 145, no. 3, pp. 787-805, 1997.

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

[19] S. Sridhar, F. Lam, G. Blelloch, R. Ravi, and R. Schwartz, "Direct Maximum Parsimony Phylogeny Reconstruction from Genotype Data," *BMC Bioinformatics,* vol. 8, no. 1, p. 472, 2007.

[20] T. Margush and F. Mcmorris, "Consensus N-Trees," *Bull. Math. Biology,* vol. 43, pp. 239-244, 1981.

[21] E. Adams, "N-Trees as Nestings: Complexity, Similarity, and Consensus," *J. Classification,* vol. 3, no. 2, pp. 299-317, 1986.

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

[23] Y.J. Chu and T.H. Liu, "On the Shortest Arborescence of a Directed Graph," *Science Sinica,* vol. 14, pp. 1396-1400, 1965.

[24] R.R. Hudson, "Generating Samples under a Wright-Fisher Neutral Model of Genetic Variation," *Bioinformatics,* vol. 18, no. 2, pp. 337-338, Feb. 2002.

[25] A. Rambaut and N.C. Grass, "Seq-Gen: An Application for the Monte Carlo Simulation of DNA Sequence Evolution Along Phylogenetic Trees," *Computer Applications in Biosciences,* vol. 13, no. 3, pp. 235-238, 1997.

[26] M. Shriver and R. Kittles, "Genetic Ancestry and the Search for Personalized Genetic Histories," *Nature Rev. Genetics,* vol. 5, pp. 611-618, 2004.

[27] M. Meila, "Comparing Clusterings—An Information Based Distance," *J. Multivariate Analysis,* vol. 98, no. 5, pp. 873-895, 2007.

[28] M. Kayser, M. Krawczak, L. Excoffier, P. Dieltjes, D. Corach, V. Pascali, C. Gehrig, L. Bernini, J. Jespersen, E. Bakker, L. Roewer, and P. de Knijff, "An Extensive Analysis of Y-Chromosomal Microsatellite Haplotypes in Globally Dispersed Human Populations," *Am. J. Human Genetics,* vol. 68, no. 4, pp. 990-1018, 2001.

[29] S. Tishkoff and B. Verrelli, "Patterns of Human Genetic Diversity: Implications for Human Evolutionary History and Disease," *Ann. Rev. Genomics and Human Genetics,* vol. 4, no. 1, pp. 293-340, 2003.

[30] D. Reich, K. Thangaraj, N. Patterson, A. Price, and L. Singh, "Reconstructing Indian Population History," *Nature,* vol. 461, no. 7263, pp. 489-494, 2009.

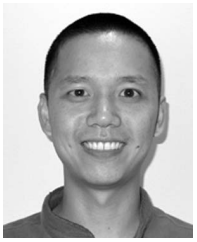
[31] S. Tishkoff and S. Williams, "Genetic Analysis of African Populations: Human Evolution and Complex Disease," *Nature Rev. Genetics,* vol. 3, no. 8, pp. 611-621, 2002.

[32] M. He, J. Gitschier, T. Zerjal, P. de Knijff, C. Tyler-Smith, and Y. Xue, "Geographical Affinities of the HapMap Samples," *PLoS ONE,* vol. 4, no. 3, p. e4684, 2009.

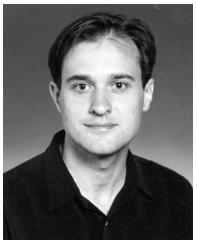
[33] D.E. Reich and D.B. Goldstein, "Genetic Evidence for a Paleolithic Human Population Expansion in Africa," *Proc. Nat'l Academy of Sciences USA,* vol. 95, no. 14, pp. 8119-8123, 1998.

[34] 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,* vol. 64, no. 02, pp. 117-134, 2000.

[35] L.A. Zhivotovsky, "Estimating Divergence Time with the Use of Microsatellite Genetic Distances: Impacts of Population Growth and Gene Flow," *Molecular Biology and Evolution*, vol. 18, no. 5, pp. 700-709, 2001.
[36] D. Gusfield, "Optimal, Efficient Reconstruction of Root-Unknown Phylogenetic Networks with Constrained and Structured Recombination," *J. Computer and System Sciences*, vol. 70, no. 3, pp. 381-398, 2005.



Ming-Chi Tsai received the BA degrees in computer science and molecular and cell biology from the University of California, Berkeley, in 2003. Since 2007, he has been working toward the PhD degree at the Joint CMU-Pitt PhD Program in computational biology. His primary area of research is computational biology.



Guy Blelloch received the BA degree in physics and the BS degree in engineering, in 1983, from Swarthmore College, and the MS and PhD degrees in computer science from the Massachusetts Institute of Technology, in 1986 and 1988, respectively. He is currently a professor of computer science at Carnegie Mellon University and codirector of the ALADDIN center for the study of algorithms. His research interests are in programming languages and applied algorithms.

R. Ravi received the BTech degree in computer science and engineering from the Indian Institute of Technology, Madras, in 1989, and the PhD degree in computer science from Brown University, in 1993. After postdoctoral fellowships at UC Davis and DIMACS, Princeton University, he joined the Operations Research faculty at the Tepper School of Business at Carnegie Mellon University, in 1995, where he is currently Carnegie Bosch Professor of operations research and computer science.

Russell Schwartz received the BS, MEng, and PhD degrees from the Department of Electrical Engineering and Computer Science at the Massachusetts Institute of Technology, the last in 2000. He later worked in the Informatics Research group at Celera Genomics. He joined the faculty of Carnegie Mellon University, in 2002, where he is currently an associate professor of biological sciences.