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

Ming-Chi Tsai[1], Guy Blelloch[2], R. Ravi[3], and Russell Schwartz[4]

[1] Joint CMU-Pitt Computational Biology Program,
Carnegie Mellon University and University of Pittsburgh, Pittsburgh, PA 15213, USA
[2] Department of Computer Science
[3] Tepper School of Business
[4] Department of Biological Science,
Carnegie Mellon University, Pittsburgh, PA 15213, USA

**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. In this study, we present a novel approach to inferring human evolutionary history from genetic variation data. Our approach uses the idea of consensus trees, a technique generally used to reconcile species trees from divergent gene trees, adapting it to the problem of finding the robust relationships within a set of intraspecies phylogenies derived from local regions of the genome. We assess the quality of the method on two large-scale genetic variation data sets: the HapMap Phase II and the Human Genome Diversity Project. Qualitative comparison to a consensus model of the evolution of modern human population groups shows that our inferences closely match our best current understanding of human evolutionary history. A further comparison with results of a leading method for the simpler problem of population substructure assignment verifies that our method provides comparable accuracy in identifying meaningful population subgroups in addition to inferring the relationships among them.

## 1 Introduction

The advent of high-throughput genotyping methods and their application in large-scale genetic variation studies have made it possible to determine in unprecedented detail how the modern diversity of the human species arose from our common ancestors. 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 [22]. Past work on human ancestry inference has essentially treated it as two distinct problems: identifying meaningful population groups

M. Borodovsky et al. (Eds.): ISBRA 2010, LNBI 6053, pp. 167–178, 2010.

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 [16], which uses a Markov Chain Monte Carlo (MCMC) clustering method to group sequences into subpopulations characterized by similar allele frequencies across variation sites. A variety of other computational and statistical methods have been developed to perform population substructure inference or similar analyses, including EIGENSOFT [15], Spectrum [19], and SABER [21]. 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 [14], mtDNA genotypes [9,2], short tandem repeats [9,23], and Y chromosome polymorphism [5] — supplemented by extensive manual analysis informed by population genetics theory. 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.

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 [13], 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 [20], 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 the HapMap Phase II [7] and Human Genome Diversity Project [8] datasets. 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 we are given a set of $m$ taxa, $S$, 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 $T_1, \ldots, T_n$.

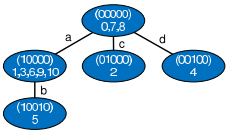
Our objective function for choosing $T_M$ is based on the task of finding a consensus tree [13] from a set of phylogenies each describing inferred ancestry of a small region of a genome. Our problem is, however, fairly different from standard uses of consensus tree algorithms in that our phylogenies are derived from many variant markers, each only minimally informative, within a single species. Standard consensus tree approaches, such as majority consensus [11] or Adam consensus [1], 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.

For this purpose, we chose a model of the problem based on the principle of minimum description length (MDL)[4], a standard technique for avoiding overfitting when making inferences from noisy data sets. An MDL method seeks 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 $T_j$. 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 $f(T_M) = |T_M| \log_2 m$ defines an additional penalty on model edges used to set a minimum confidence level on edge predictions.

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". Fig. 1a shows an example of a hypothetical tree, its description as a set of bipartitions, and




(a)

$e_a$: 1,3,5,6,9,10|0,2,4,7,8
$e_b$: 0,1,2,3,4,6,7,8,9,10|5
$e_c$: 0,1,3,4,5,6,7,8,9,10|2
$e_d$: 0,1,2,3,5,6,7,8,9,10|4

(b)

$e_a$: 01010110011
$e_b$: 00000100000
$e_c$: 00100000000
$e_d$: 00001000000

(c)

**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.

representations of the bipartitions as bit strings. Such a bit representation allows us to compute the encoding length of a bipartition $b$ as the entropy of its corresponding bit string. If we define $p_0$ to be the fraction of bits of $b$ that are zero and $p_1$ as the fraction that are one, then:

$$L(b) = m\left(-p_0 \log_2 p_0 - p_1 \log_2 p_1\right)$$

Similarly, we can encode the representation of one bipartition $b_1$ given another $b_2$ using the concept of conditional entropy. If we let $p_{00}$ 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:

$$L(b_1|b_2) = m\left[\sum_{s,t\in\{0,1\}} -p_{st} \log_2 p_{st} \; + \sum_{u\in\{0,1\}} (p_{0u} + p_{1u}) \log_2 (p_{0u} + p_{1u})\right]$$

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 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:

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

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

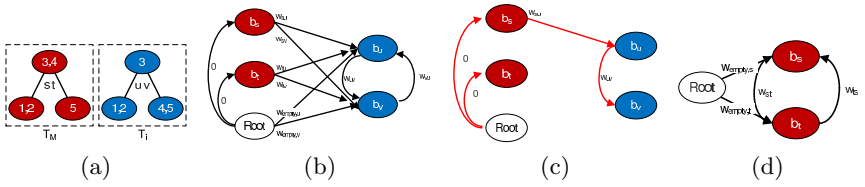



**Fig. 2.** Illustration of the DMST construction for determining model description length. (a) Hypothetical model tree $T_M$ (red) and observed tree $T_i$ (blue). (b) Graph of possible reference relationships for explaining $T_i$ (blue nodes) by reference to $T_M$ (red nodes). (c) A possible resolution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

### 2.2 Algorithms

**Encoding Algorithm.** We pose the problem of computing $L(T_M)$ and $L(T_i|T_M)$ as a weighted directed minimum spanning tree (DMST) problem, illustrated in Fig. 2. 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 illustrates the construction for a hypothetical model tree $T_M$ and observed tree $T_i$ (Fig. 2(a)), showing the graph of possible reference relationships (Fig. 2(b)), a possible solution corresponding to a specific explanation of $T_i$ in terms of $T_M$ (Fig. 2(c)), and the graph of possible reference relationships for $T_M$ by itself (Fig. 2(d)).

For both constructions, the minimum encoding length is found by solving for the DMST with the algorithm of Chiu and Liu [3] and summing the weights of the edges. This cost is computed for a candidate model tree $T_M$ and for each observed tree $T_i$ 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 compute the score of 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 $T'_M$. If the algorithm chooses to insert a new bipartition $b$, it then performs an additional expectation-maximization-like local optimization to improve the fit. It repeatedly identifies the set $B$ of bipartitions explained by $b$ and then locally improves $b$ by iteratively flipping any bits that lower the cost of explaining $B$, 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. 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). 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. 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. The number of bipartitions $w_j$ explained by each model bipartition $b_j$ provides a rough estimate of the number of mutations that occurred after the population diverged, which can be interpreted as an estimated elapsed time scaled by population size. We attribute this scaled time equally to the two branches to assign branch lengths to the tree. Given a weight $w_j$ for the $j$-th model bipartition, the branch length of $e = (s_j, v_{j1})$ and $(s_j, v_{j2})$ would then be $w_j/2$ and the branch length of $e = (t, s_j)$ for $t = \text{parent}(v_j)$ would be $w_{j-1}/2 - w_j/2$.

### 2.3 Validation Experiments

We evaulated our methods by applying them to samples from two SNP variation datasets. We first used the phase II HapMap data set (phased, release 22) [7] 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 44 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 genotypes as inferred by the HapMap consortium. For each run, we randomly sampled 8,000 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 8,000 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) [8], 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.

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 by reference to a expert-curated model of human evolution derived from a review by Shriver and Kittles[18], 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 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 one of the leading substructure algorithms, STRUCTURE (version 2.2) [16]. STRUCTURE requires that the user specify a desired number of populations, for which we supplied the true number for each data set (four for HapMap and seven for HGDP). For each run, we performed 2,000 iterations of burn-ins and 10,000 iterations of the STRUCTURE MCMC sampling, assigning each individual to the population group of

highest likelihood as determined by STRUCTURE. We did not make use of STRUCTURE's capacity to infer admixture or to use additional data on linkage disequilibrium between sites. We assessed the quality of the results based on variation of information [12], a method commonly used to assess accuracy of a clustering method relative to a pre-defined "ground truth." Variation of information is defined as $2H(X,Y) - H(X) - H(Y)$, 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. We also assessed robustness of the methods 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 of $\sum_{i,j} \min\{1 - \frac{2b_{ij}}{\lfloor (a_{ij}+b_{ij})\rfloor}, 1 - \frac{2a_{ij}}{\lfloor (a_{ij}+b_{ij})\rfloor}\} / \binom{n}{2}$. 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.

## 3 Results

Fig. 3 shows the trees inferred by our method on the two data sets alongside their corresponding condensed Shriver and Kittles "gold standard" trees. Fig. 3(a) shows the inferred tree produced by our model. Based on the numbers of bipartitions explained by each method, 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. 3(b)) shows an identical structure. Furthermore, these results are consistent with many independent lines of evidence for the out-of-Africa hypothesis of human origins [10,24,18].

For the HGDP dataset, the trees differ slightly from run to run, so we arbitrarily provide our first run, Fig. 3(c), 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. 3(d)). 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

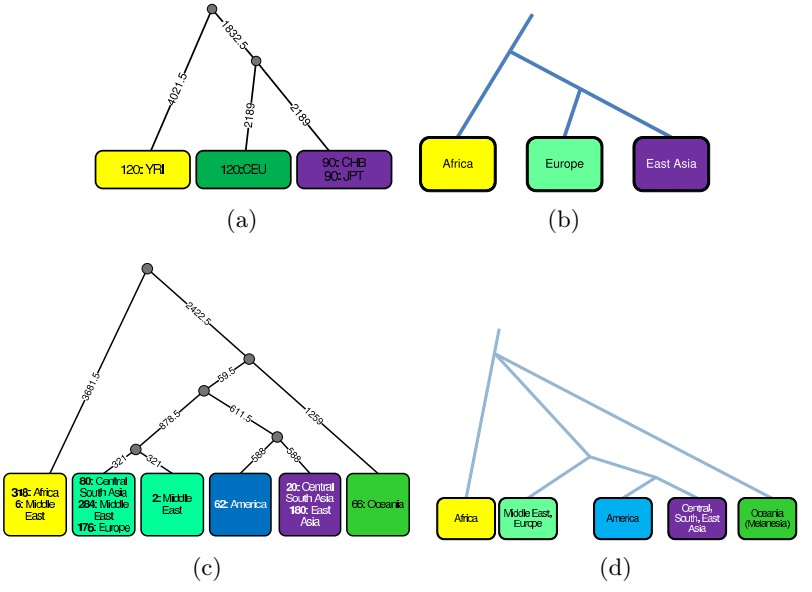
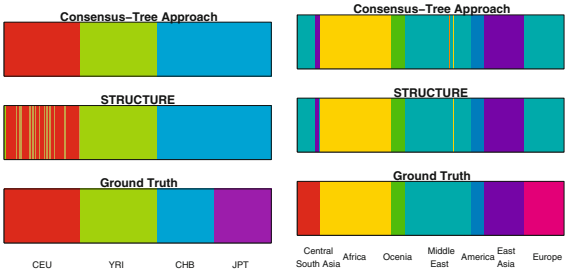



**Fig. 3.** Inferred consensus trees. Node labels show numbers of haplotypes belonging to each known populations. Edge labels can be interpreted as estimates of scaled time since each divergence. (a) Consensus tree obtained from HapMap dataset. (b) Trimmed and condensed tree from [18]. (c) Consensus tree obtained from HGDP dataset. (d) Trimmed and condensed tree from [18].

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 [17]). 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 [25,10].

Fig. 4 shows the corresponding cluster assignments for our method and STRUCTURE in order to provide a secondary assessment of our method's utility for the simpler sub-problem of subpopulation inference relative to STRUCTURE and the presumed ground truth. Each inferred cluster is assigned a distinct label, with colors chosen to maximize agreement with the true population structure. For HapMap (Fig. 4(a)), our method consistently identified YRI and CEU as distinct subpopulations but failed to separate CHB (Chinese) and JPT (Japanese). STRUCTURE produced generally identical output except in one run where it grouped a subset of the CHB and JPT populations in a separate cluster. Tab. 1(a) quantifies these observations, suggesting marginally better performance for the consensus tree method by both measures. Results were more ambiguous for HGDP (Fig. 4(b)) with STRUCTURE showing generally greater sensitivity but still worse consistency than our method. STRUCTURE usually at least approximately finds six of the annotated seven population groups, having




(a) (b)

**Fig. 4.** Inferred population structures from the consensus tree method and STRUCTURE. From top to bottom: consensus-tree, STRUCTURE, and ground truth. (a): Inferred population structures from a single trial of 8,000 trees from HapMap Phase II dataset. (b): Inferred population structures from one trial of 10,000 trees.

**Table 1.** Variation of information (VI) and inconsistency score. Lower VI reflects higher accuracy in identifying known population structure. Higher consistency reflects greater reproducibility between independent samples.

(a) Hapmap

| | VI | Consistency |
|---|---|---|
| STRUCTURE | 0.5039 | 0.0226 |
| Consensus Tree | **0.4286** | **0.0000** |

(b) HGDP

| | VI | Consistency |
|---|---|---|
| STRUCTURE | **0.8949** | 0.1341 |
| Consensus Tree | 0.9265 | **0.0765** |

difficulty only in identifying Central South Asians as a distinct group, consistent with a similar outcome from He *et al.* [6]. The consensus tree method reliably finds five of the seven populations, usually conflating Middle Eastern and European in addition to failing to recognize Central South Asians. Tab. 1(b) quantifies these observations, with the consensus tree method showing slightly worse variation of information but better consistency than STRUCTURE. We note that our methods also provide comparable runtimes to STRUCTURE despite solving a more involved inference problem. Our methods required approximately 1.4 hours for the HapMap data and 30 hours for the HGDP data, compared to approximately 2.5 hours and 48 hours for STRUCTURE.

## 4 Discussion

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 the

leading STRUCTURE method. The method also provides edge length estimates that can roughly be interpreted as estimates of time since divergence on the crude assumption that effective population sizes are equal along all sibling tree edges. The addition of an outgroup to determine likely ancestral states at internal nodes of the tree should in principle allow us to drop that assumption and estimate both divergence times and effective population sizes along the tree edges. The MDL approach should also in principle 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 protocols for simulating sequence data generated from a human-like population history, 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 U.S. National Science Foundation IIS award #0612099 and by NIH T32 training grant T32 EB009403 as part of the HHMI-NIBIB Interfaces Initiative.

## References

1. Adams, E.N.: N-trees as nestings: Complexity, similarity, and consensus. Journal of Classification 3(2), 299–317 (1986) 10.1007/BF01894192
2. Cann, R.L., Stoneking, M., Wilson, A.C.: Mitochondrial DNA and human evolution. Nature 325(6099), 31–36 (1987) 10.1038/325031a0
3. Chu, Y.J., Liu, T.H.: On the shortest arborescence of a directed graph. Science Sinica 14, 1396–1400 (1965)
4. Grnwald, P.D., Myung, I.J., Pitt, M.A.: Advances in Minimum Description Length: Theory and Applications. The MIT Press, Cambridge (2005)
5. Hammer, M.F., Spurdle, A.B., Karafet, T., Bonner, M.R., Wood, E.T., Novelletto, A., Malaspina, P., Mitchell, R.J., Horai, S., Jenkins, T., Zegura, S.L.: The geographic distribution of human Y chromosome variation. Genetics 145(3), 787–805 (1997)
6. He, M., Gitschier, J., Zerjal, T., de Knijff, P., Tyler-Smith, C., Xue, Y.: Geographical affinities of the HapMap samples. PLoS ONE 4(3), e4684, 03 (2009)
7. International HapMap Consortium. A second generation human haplotype map of over 3.1 million snps. Nature 449(7164), 851–861 (October 2007)
8. Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, R.J., Vanliere, J.M., Fung, H.C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., Bras, J.M., Schymick, J.C., Hernandez, D.G., Traynor, B.J., Simon-Sanchez, J., Matarin, M., Britton, A., van de Leemput, J., Rafferty, I., Bucan, M., Cann, H.M., Hardy, J.A., Rosenberg, N.A., Singleton, A.B.: Genotype, haplotype and copy-number variation in worldwide human populations. Nature 451(7181), 998–1003 (2008)
9. Jorde, L.B., Bamshad, M.J., Watkins, W.S., Zenger, R., Fraley, A.E., Krakowiak, P.A., Carpenter, K.D., Soodyall, H., Jenkins, T., Rogers, A.R.: Origins and affinities of modern humans: a comparison of mitochondrial and nuclear genetic data. American Journal of Human Genetics 57, 523–538 (1995)

10. Kayser, M., Krawczak, M., Excoffier, L., Dieltjes, P., Corach, D., Pascali, V., Gehrig, C., Bernini, L.F., Jespersen, J., Bakker, E., Roewer, L., de Knijff, P.: An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. American Journal of Human Genetics 68(4), 990–1018 (2001)
11. Margush, T., Mcmorris, F.R.: Consensus n-trees. Bulletin of Mathematical Biology 43, 239–244 (1981)
12. Meila, M.: Comparing clusterings–an information based distance. Journal of Multivariate Analysis 98(5), 873–895 (2007) doi: 10.1016/j.jmva.2006.11.013
13. Nei, M., Kumar, S.: Molecular Evolution and Phylogenetics. Oxford University Press, Oxford (2000)
14. Nei, M., Roychoudhury, A.K.: Genetic relationship and evolution of human races. Evolutionary Biology 14, 1–59 (1982)
15. Patterson, N., Price, A.L., Reich, D.: Population structure and eigenanalysis. PLoS Genetics 2(12), e190+ (2006)
16. Pritchard, J.K., Stephens, M., Donnelly, P.: Inference of population structure using multilocus genotype data. Genetics 155(2), 945–959 (2000)
17. Reich, D., Thangaraj, K., Patterson, N., Price, A.L., Singh, L.: Reconstructing indian population history. Nature 461(7263), 489–494 (2009) 10.1038/nature08365
18. Shriver, M.D., Kittles, R.A.: Genetic ancestry and the search for personalized genetic histories. Nature Reviews Genetics 5, 611–618 (2004)
19. Sohn, K.A., Xing, E.P.: Spectrum: joint bayesian inference of population structure and recombination events. Bioinformatics 23(13), i479–i489 (2007)
20. Sridhar, S., Lam, F., Blelloch, G., Ravi, R., Schwartz, R.: Direct maximum parsimony phylogeny reconstruction from genotype data. BMC Bioinformatics 8(1), 472 (2007)
21. Tang, H., Coram, M., Wang, P., Zhu, X., Risch, N.: Reconstructing genetic ancestry blocks in admixed individuals. The American Journal of Human Genetics 79(1), 1–12 (2006) doi: 10.1086/504302
22. Thomas, D.C., Witte, J.S.: Point: Population stratification: A problem for case-control studies of candidate-gene associations? Cancer Epidemiol Biomarkers Prev. 11(6), 505–512 (2002)
23. Tishkoff, S.A., Dietzsch, E., Speed, W., Pakstis, A.J., Kidd, J.R., Cheung, K., Bonn-Tamir, B., Santachiara-Benerecetti, A.S., Moral, P., Krings, M., Pbo, S., Watson, E., Risch, N., Jenkins, T., Kidd, K.K.: Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science 271(5254), 1380–1387 (1996)
24. Tishkoff, S.A., Verrelli, B.C.: Patterns of human genetic diversity: Implications for human evolutionary history and disease. Annual Review of Genomics and Human Genetics 4(1), 293–340 (2003)
25. Tishkoff, S.A., Williams, S.M.: Genetic analysis of African populations: human evolution and complex disease. Nat. Rev. Genet. 3(8), 611–621 (2002) 10.1038/nrg865