Genesis and population analysis

Tree-guided Bayesian inference of population structures

Yu Zhang*

Department of Statistics, The Pennsylvania State University, State College, PA, USA

Received on December 5, 2007; revised on February 3, 2008; accepted on February 18, 2008

Associate Editor: Keith Crandall

ABSTRACT

Motivation: Inferring population structures using genetic data sampled from a group of individuals is a challenging task. Many methods either consider a fixed population number or ignore the correlation between populations. As a result, they can lose sensitivity and specificity in detecting subtle stratifications. In addition, when a large number of genetic markers are used, many existing algorithms perform rather inefficiently.

Result: We propose a new Bayesian method to infer population structures using multiple unlinked single nucleotide polymorphisms (SNPs). Our approach explicitly considers the population correlation through a tree hierarchy, and treat the population number as a random variable. Using both simulated and real datasets of worldwide samples, we demonstrate that an incorporated tree can consistently improve the power in detecting subtle population stratifications. A tree-based model often involves a large number of unknown parameters, and the corresponding estimation procedure can be highly inefficient. We further implement a partition method to analytically integrate out all nuisance parameters in the tree. As a result, our method can analyze large SNP datasets with significantly improved convergence rate.

Availability: http://www.stat.psu.edu/~yuzhang/tips.tar
Contact: yuzhang@stat.psu.edu
Supplementary information: Supplementary data are available at Bioinformatics online.

1 INTRODUCTION

Population-based case-control association study has been routinely used in detecting genes responsible for human complex diseases. The most commonly genotyped markers from the sampled individuals are single nucleotide polymorphisms (SNPs). The rough locations of disease genes can be revealed by comparing the SNP allele distributions between cases and controls. Population-based association studies, however, can be complicated by the presence of population structures (Freedman et al., 2004; Lander and Shork, 1994; Rosenberg and Nordborg, 2006; Thomas and Witte, 2002). Biased selection of individuals or unknown heterogeneity between cases and controls can cause population structures. Allele distributions across populations can vary due to mutation, drift and selection. When analyzing structured individuals in a case-control sample, spurious associations will appear. It is thus important to detect populations presented in a sample and adjust for the population effects. Although correction methods without inferring populations have been developed (Devlin and Roeder, 1999; Price et al., 2006; Thomas and Witte, 2002), their performance are contingent on the assumption of structural homogeneity or linearity, and their parameter estimation uncertainties can also increase spurious associations (Kimmel, 2007; Marchini et al., 2004). Structured corrections based on explicitly detected populations, on the other hand, are more tolerant to the homogeneity assumption. The estimation uncertainties can also be addressed using Bayesian methods (Pritchard et al., 2000). However, structured correction methods are more contingent on the power of detecting subtle structures in the sample. It is thus important to further improve the sensitivity and specificity of existing population detection algorithms.

Several Bayesian algorithms have been developed to identify either a fixed or an unknown number of populations using multi-locus genotype data, among which STRUCTURE (Falush et al., 2003; Pritchard et al., 2000) is probably one of the most popular methods. STRUCTURE requires the user to specify a number of populations to be detected in a sample. Markov chain Monte Carlo (MCMC) is then used to estimate a posterior distribution of individuals’ origins. Both population allele frequencies and population prevalence are treated as unknown variables, and are estimated from the data a posteriori. To infer the true population number, STRUCTURE compares the posterior likelihoods estimated from a variety of population numbers. Other Bayesian approaches that directly estimate an unknown number of populations have also been developed (Corander et al., 2003; Dawson and Belkhir, 2001; Francois et al., 2006; Pella and Masuda, 2006). Many of them, however, assume independence among populations and thus may lose sensitivity compared to STRUCTURE (Falush et al., 2003), except for the method by Francois et al. (2006), which uses a spatial model to account for the continuity in allele differentiation across geographic regions. Non-Bayesian approaches have also been developed, including expectation-maximization algorithms (Kohler and Bickeboller, 2005; Satten et al., 2001) and distance-based approaches (Purcell et al., 2007; Sridhar et al., 2007). Compared to Bayesian methods, they are computationally much faster, but they generally lack of rigorous measurements in predicting the unknown population numbers. In addition, uncertainties in the estimated population origins and missing data are more naturally handled in a Bayesian framework.

We introduce a new Bayesian method, tree-guided inference of population structures (TIPS), to detect an unknown number of
populations using multiple unlinked SNP markers. Our method attempts to unify two unique advantages of existing methods. First, the method incorporates a tree hierarchy to capture the correlation among multiple populations, such that closely related populations are allowed to share a more recent common ancestor than distantly related populations. By simultaneously estimating an unknown number of populations, the population origins, and the tree structures, our method can improve the sensitivity in detecting subtle population structures often occurring at the tips of a population tree. Most existing methods do not consider tree models. For example, STRUCTURE (Falush et al., 2003) assumes completely independent drifting of populations from a single common ancestor without modeling fine structures among subpopulations. Many other methods do not even consider the correlation between populations (Corander et al., 2003; Dawson and Belkhir, 2001; Francois et al., 2006; Pella and Masuda, 2006), where they only estimate trees as a visualization tool at the post-analysis stage. Using both simulated and real datasets, we demonstrate that a tree-guided population inference not only can improve the power in detecting subtle stratifications, but also is computationally feasible even for genome-wide SNP datasets.

A tree-based model often involves a huge number of nuisance parameters, such as allele frequencies of the current and ancestral populations. These nuisance parameters are strongly correlated with the unknown population origins and tree structures. Estimating a large number of these correlated parameters is thus highly inefficient. For example, STRUCTURE may take weeks to converge when analyzing a genome-wide SNP dataset, because the ancestral allele frequency of each marker needs to be estimated iteratively. To alleviate computational issues, researchers often compromise on using a subset of markers in their analysis, such as a few hundreds of ancestral informative markers (AIMs). Using subsets of markers is clearly statistically inefficient, and arbitrary marker selection can bias the inference results (Enoch et al., 2006). To achieve computational efficiency, we utilize a partition model to assign marker into groups, such that markers in each group describe a unique dependency structure among populations in a tree. After assigning markers into groups, we can obtain a marginalized likelihood function with all nuisance parameters integrated out. As a result, MCMC sampling from the marginalized likelihood function will converge significantly faster to the true posterior distribution than from the full model. Marginalizing nuisance parameters has been used by other methods (Corander et al., 2003; Pella and Masuda, 2006). However, their models only assume independent population clusters with fixed hyperparameters, and thus lose valuable information about subtle population structures. Our Bayesian approach outputs a joint posterior distribution of the population number, the individuals’ origins, and the tree structure, all of which are estimated via MCMC sampling.

2 METHODS

2.1 Notations

Consider a sample of $N$ individuals with $L$ unlinked SNP markers genotyped from each individual. Let $G = (g_1, g_2, \ldots, g_L)$ denote the observed genotypes with $g = (g_1, \ldots, g_N)^T$. We assume that the sample contains at most $K_{\text{max}}$ populations, and let $O = (o_1, \ldots, o)$ denote individuals’ population origins. The actual number of populations $K$ in the sample is then determined by $O$. We further assume Hardy–Weinberg Equilibrium (HWE) for diploid individuals and that populations are not admixed.

A bifurcating tree is introduced to capture the correlation among multiple populations. Let $T = (H, B)$ denote a tree, where $H$ denotes the tree hierarchy with leaves uniquely representing $K$ populations, and $B$ denotes the branch lengths of the tree. For model convenience, we define the branch length as the proportion of segregating markers between two population groups connected by the branch. A segregating marker is assumed to take different allele frequencies in two populations, while a non-segregating marker is assumed to take common allele frequencies. As an example shown in Figure 1, a tree contains four labeled leafs representing $K = 4$ populations. Let $p_i \in [0, 1]$ denote the proportion of segregating markers within populations descended from node $u$. We define the branch length between $u$ and $v$ as $|p_u - p_v|$. Our branch lengths satisfy the additivity property.

It is not straightforward to assign priors to $p_i$ because a node is partially ordered under the tree constraint. Instead, suppose $v$ is the parent node of $u$, we let $q_v = p_u / p_v \in [0, 1]$ and define the branch length variable as $B = \{q_v\}$. $B$ is a $(K-1)$-dimensional vector with each $q_v$ associated with an internal node of the tree, and $q_1 = p_1$ at the root. We can easily assign priors to $\{q_v\}$ because they represent relative proportions and are independent with each other. It can be shown that $p_v = q_1 q_2 \cdots q_v$, where the subscripts $(r, n_1, \ldots, u)$ represent the nodes visited by the shortest path from the root $r$ to the node $u$. As a result, the branch length between two adjacent nodes $v$ and $u$ is given by $|p_v - p_u| = q_v q_{1u} \cdots q_{1u} (1 - q_v)$, where the additivity property still holds true.

2.2 Population partition model

One typical approach to model the correlation among populations in a tree is by introducing a hierarchical prior on the allele frequencies, such as the Balding–Nichols model (Balding and Nichols, 1995). Estimating allele frequency for all genotyped markers, however, can be extremely inefficient as they are strongly correlated with the unknown population origins. We use an alternative partition approach. The idea is to assign markers into groups that uniquely correspond to some partitions of $K$ populations as non-overlapping clusters. For each population partition, we assume that markers in the associated group are segregating across clusters, but share common allele frequencies within clusters. As a result, populations are correlated through non-segregating markers and all allele frequencies can be integrated out. For a mixture of Chinese, Japanese and European individuals, markers may be assigned to three groups: markers in group $1$ are treated as segregating in all three
Fig. 2. Given a tree of three populations [C, J, E], there are three possible partitions. Indicators are shown at every internal node of the tree to determine the partition. Non-distinguishable populations are shown in common brackets. Distinguishable populations are shown in separate brackets. Additional partitions of these populations are possible by modifying the tree structure.

We determine marker groups and the corresponding population partitions from the tree as follows: let $J = (J_1, \ldots, J_m)$ denote the membership of L markers, where $J_i$ of marker i is a $(K-1)$-dimensional vector of indicators; for each marker, we assign an indicator $\delta_j$ to each of $(K-1)$ internal nodes $v$ of the tree; $\delta_j = 1$ indicates that populations descended from $v$ segregate at this marker, and 0 otherwise; if $u$ is descended from $v$ and $\delta_u = 1$, then we let $\delta_v = 1$ because segregation under $u$ indicates segregation under its ancestor $v$. As a result, the $K$ current populations can be uniquely partitioned into non-overlapping clusters by the value of $J = (\delta_{v_1}, \ldots, \delta_{v_{K-1}})$, where populations connected by nodes with $\delta_j = 0$ are assigned into common clusters. The model assumption is that a marker does not segregate populations within clusters but has different allele frequencies between clusters. Different values of $J_i$ determine different population partitions. An example is shown in Figure 2.

Intuitively, our approach distinguishes the correlation among populations through a dynamic partitioning model. Closely related populations in a tree can be more frequently assigned to common clusters and thus share fewer segregating markers than distantly related populations. The advantage of using this approach is that we can explicitly integrate out all allele frequency parameters from the likelihood function. The population origins ($O$) is given by modifying the tree structure.

We can write $P(G | J, T, O)$ as a nonlinear function of $\delta_1, \ldots, \delta_{K-1}$, which can perform much better than the full model in terms of its MCMC convergence (Liu, 2001). Let $\Phi(\gamma, v)$ denote the likelihood of genotypes at marker i among individuals descended from node v. We can write $\Phi(\gamma, v) = \prod_{j=1}^{L} \Gamma(\gamma_0 + \gamma_1) / \Gamma(\gamma_0) / \Gamma(\gamma_1 + \gamma_0)$, where $\gamma_0$ denotes the allele frequency and $\gamma_1$ the number of alleles.

The population origins ($O$) are strongly correlated with the marker memberships ($J$). A Gibbs sampler that iteratively samples $O$ and $J$ can be easily trapped in a local mode. This is similar to the problems when we simultaneously estimate population origins and marker allele frequencies. We therefore integrate out $J$ to obtain a collapsed model $P(G | T, O)$, which can perform much better than the full model in terms of its MCMC convergence (Liu, 2001).

We use MCMC sampling to use posterior samples from $P(G | T, O)$. The MCMC algorithm works as follows:

1. Randomly select an individual i and update $O_i$.
2. Randomly reconnect a branch of the tree to a new location.
3. Randomly update $q_k$ of a node by $q_k + \epsilon$, $\epsilon \sim N(0, 0.01)$.
4. Randomly split a leaf and the associated population into two, and sample a new branch length variable from $[0, 1]$. 

$P(T | O)$: Given K populations, we assign a uniform prior to the tree $T = (H, B)$. Since there are $(2K - 3)! / (2^K - 2)!$ rooted bifurcating trees with $K \geq 1$ labeled leaves (Felsenstein, 1978), we let $P(H | O) = (2^K - 2)! / (2K - 3)!$ for $K \geq 1$, and $P(H | O) = 1$ for $K = 1$. Non-uniform priors that favor unbalanced trees can also be used here. We further assign a uniform prior $U(0,1)$ to each $q_k$ in the branch length variable $B$. The joint prior distribution of a tree is then $P(T | O) = P(H | O) P(B | O)$.

$P(O | T, O)$: Given a tree structure, we assign markers into groups, where the group membership $J = (\delta_1, \ldots, \delta_{K-1})$ is a vector of indicators. We assign a prior distribution to $\{J_i\}$ through a recursive function

$$f(v, i) = (1 - q_i) \Gamma(1 - \delta_v) + q_i \Gamma(v_1, i) \Gamma(v_2, i) \delta_v$$

Here, $f(v, i)$ is defined on each node $v$ for each marker $i$, and $v_1, v_2$ denote the two children nodes of $v$ in the tree. We calculate $f(v, i)$ from the bottom of the tree to the root, and the prior $P(J_i | T, O)$ is given by $f(\text{root}, i)$. At any leaf node, we have $\delta_{\text{leaf}} = q_{\text{leaf}} = 0$ and thus $f(\text{leaf}, i) = 1$. For two populations, the tree has only one internal node (the root), so $P(J_i | T, O) = (1 - q_{\text{root}})(1 - \delta_v + \delta_1) + q_{\text{root}} \delta_v$. If the indicator is 0 (the marker is not segregating), then $P(J_i | T, O) = 1 - q_{\text{root}}$, otherwise $P(J_i | T, O) = q_{\text{root}}$. We always have $P(J_i | T, O) = 1$.

We finally compute the conditional likelihood $P(G | J, T, O)$. Let $\Omega_k$ denote a partition of K populations by $J$, at marker i. Let $p_{\alpha k}$ denote the probability of allele $\alpha$ in the kth cluster in $\Omega_k$, and let $n_{\alpha k}$ denote the corresponding allele counts. The likelihood of genotypes can be written as a multinomial, and we integrate out $p_{\alpha k}$ using a conjugate prior $D(\alpha)$ with $\alpha_k = 0.5$ for allele $k$ to obtain

$$P(G | J, T, O) = \prod_{i=1}^{L} \prod_{\alpha_k} \left[ \Gamma(n_{\alpha k} + \alpha_k) / \Gamma(\alpha_k) / \Gamma(n_{\alpha k} + \alpha_k) \right] \prod_{j=1}^{L} \Gamma(\gamma_0 + \gamma_1) / \Gamma(\gamma_0) / \Gamma(\gamma_1 + \gamma_0)$$

2.4 The collapsed model

The population origins ($O$) are strongly correlated with the marker memberships ($J$). A Gibbs sampler that iteratively samples $O$ and $J$ can be easily trapped in a local mode. This is similar to the problems when we simultaneously estimate population origins and marker allele frequencies. We therefore integrate out $J$ to obtain a collapsed model $P(G | T, O)$, which can perform much better than the full model in terms of its MCMC convergence (Liu, 2001).

Let $\Phi(\gamma, v)$ denote the likelihood of genotypes at marker i among individuals descended from node v. We can write $\Phi(\gamma, v) = \prod_{j=1}^{L} \Gamma(\gamma_0 + \gamma_1) / \Gamma(\gamma_0) / \Gamma(\gamma_1 + \gamma_0)$, where $\gamma_0$ denotes the allele frequency and $\gamma_1$ the number of alleles.

We integrate out $J_i$ using a recursive function

$$h(v, i) = (1 - q_i) \Phi(\gamma, v_1, i) \Phi(\gamma, v_2, i)$$

where $v_1, v_2$ denote the two children nodes of $v$. $h(v, i)$ is calculated from the bottom of the tree to the root, and the likelihood of genotypes of marker i is given by $P(G_i | T, O) = h(\text{root}, i)$. The final marginal likelihood model is

$$P(G, T, O) = \prod_{i=1}^{L} P(G_i | T, O) P(T) P(O)$$

2.5 MCMC sampling

We use MCMC sampling to use posterior samples from $P(G, T, O)$. The MCMC algorithm works as follows:

1. Randomly select an individual i and update $O_i$.
2. Randomly reconnect a branch of the tree to a new location.
3. Randomly update $q_k$ of a node by $q_k + \epsilon$, $\epsilon \sim N(0, 0.01)$.
4. Randomly split a leaf and the associated population into two, and sample a new branch length variable from $[0, 1]$. 

967
followed by 50 samplings on each dataset, and let diploid individuals each. We ran TIPS with 50 burn-ins in supplementary data, and let $Q$ switched labels (see supplementary online). Analysis. We occasionally observed label switching in our MCMC exchange their labels during MCMC sampling without affecting its dimension varies with the population number. However, we add or duals to but assumed no admixture effects. To predict the population number, we call the restricted population tree (RPT). We also ran star tree with equal branch lengths and a single ancestor, which nodes except for the root. As a result, the fixed tree resembles a $K$.

3 RESULTS

3.1 Simulation study

We first simulated genotypes of $N$ diploid individuals at $L$ SNP markers with a chosen population number $K$ and a tree $T$. We used the Balding–Nichols model (Balding and Nichols, 1995) in our simulation and assumed HWE. First, we generate the ancestral minor allele frequency (MAF) at each marker at a parent node, we sample the MAF at a children node from Beta$(p(1 - c)/c, (1 - p)(1 - c)/c)$, where $c$ measures the population divergence (Marchini et al. 2004). After generating the MAFs of all markers in $K$ populations, we determine a population prevalence $\pi = \pi_1, \ldots, \pi_K$ from $D(\psi)$ with $\psi_1 = 50/K$. We then randomly assign individuals to $K$ populations and generate their genotypes.

We choose $K = 1, 2, 3, 4, 8$ with the corresponding trees given in supplementary data, and let $L = 500, 1000, 2000$, respectively. For each $K$ and $L$, 20 datasets were simulated with 400 diploid individuals each. We ran TIPS with 50 burn-ins followed by 50 samplings on each dataset, and let $K_{max} = 5$ for $K = 1, 2, 3, 4$ and $K_{max} = 10$ for $K = 8$. To check the effect of the incorporated tree structure, we further ran our method with a fixed tree, where the branch length variables $q_u = 1$ for all nodes except for the root. As a result, the fixed tree resembles a star tree with equal branch lengths and a single ancestor, which we call the restricted population tree (RPT). We also ran STRUCTURE on the same datasets with 5000 burn-ins and 5000 samplings, where we allowed populations to be correlated but assumed no admixture effects. To predict the population number, we ran STRUCTURE using $K - 1$, $K$ and $K + 1$ clusters, respectively, and selected the best result with the maximized marginal likelihood. Although empirical method exists that may better select population numbers in STRUCTURE (Evanno et al., 2005), we want to avoid additional complexity in our comparison, and the conclusions should not change dramatically. Finally, we calculated the error rate of the predicted individual origins using pairwise comparison, which measures the proportion of individual pairs whose relationships are mistakenly predicted by the algorithm. That is, the probability of two individuals with different origins being clustered together, and vice versa.

As shown in Table 1, when the number of markers were relatively small (e.g. $L = 500$), TIPS performed less accurately than STRUCTURE. This is expected because TIPS uses a simplified model on trees as a compromise to obtain the analytical marginal likelihood function. Our model can lose some power compared to the appreciated Balding–Nichols model (Balding and Nichols, 1995), which is assumed by STRUCTURE and used in our simulation. However, when large enough markers were used (e.g. $L = 1000, 2000$), TIPS consistently outperformed STRUCTURE. Running STRUCTURE for longer iterations can slightly improve its prediction accuracy when $K \leq 3$, but at the cost of significantly increased computation time. When $K > 3$, STRUCTURE performed less well than TIPS. We manually checked the results and found that two closely related populations were often grouped by STRUCTURE into a single cluster. This issue remained after either increasing the MCMC iterations or genotyping additional markers (e.g. using 20 000 iterations or 4000 markers, data not shown), which probably indicates that a completely independent population drift model assumed in STRUCTURE is inadequate. Finally, the restricted tree model RPT performed similarly to TIPS when $K = 1$ and $K = 2$, in which case a tree is unnecessary. However, RPT performed much worse when multiple populations were present in the data. When $K = 3$, RPT lost power because the branch lengths were assumed equal.

We further calculated the average population numbers predicted by each method. As shown in Table 2, TIPS tends to underestimate the population number when the marker

| Table 1. Error rates of predicted population origins for 400 individuals |
|-----------------------------|----------------------|----------------------|----------------------|
|                             | $L^* = 500$          | $L = 1000$           | $L = 2000$           |
| Error rate                  | STR$^c$  | TIP$^d$  | RPT$^e$  | STR  | TIP    | RPT    | STR  | TIP    | RPT    |
| $K^* = 1$                   | 0.000   | 0.035   | 0.000   | 0.000 | 0.000  | 0.000  | 0.225| 0.028  | 0.000  |
| 2                           | 0.045   | 0.067   | 0.068   | 0.021 | 0.006  | 0.006  | 0.036| 0.000  | 0.000  |
| 3                           | 0.051   | 0.131   | 0.272   | 0.011 | 0.008  | 0.135  | 0.004| 0.000  | 0.134  |
| 4                           | 0.095   | 0.181   | 0.243   | 0.095 | 0.044  | 0.249  | 0.086| 0.002  | 0.244  |
| 8                           | 0.062   | 0.046   | 0.088   | 0.043 | 0.021  | 0.078  | 0.081| 0.007  | 0.087  |

Best results in each category are highlighted.

$^a$number of SNPs.
$^b$population number.
$^c$STRUCTURE.
$^d$our method.
$^e$the restricted tree model.
Table 2. Estimated number of populations

<table>
<thead>
<tr>
<th>Error rate</th>
<th>L = 500</th>
<th>L = 1000</th>
<th>L = 2000</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>STR</td>
<td>TIP</td>
<td>RPT</td>
</tr>
<tr>
<td>K = 1</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td>2</td>
<td>2.0</td>
<td>2.0</td>
<td>2.0</td>
</tr>
<tr>
<td>3</td>
<td>3.1</td>
<td>2.8</td>
<td>2.2</td>
</tr>
<tr>
<td>4</td>
<td>3.9</td>
<td>2.5</td>
<td>2.0</td>
</tr>
<tr>
<td>8</td>
<td>6.7</td>
<td>6.5</td>
<td>5.0</td>
</tr>
</tbody>
</table>

Most accurate estimate in each category is highlighted.

a number of SNPs.

b population number.

c STRUCTURE.

d our method.

e the restricted tree model.

number is relatively small. However, the estimation became increasingly accurate as the number of markers increases.

On the other hand, STRUCTURE tends to overestimate the population number when K ≤ 3, but underestimate the population number when K > 3. The underestimation remained after either increasing the MCMC iterations or genotyping more markers, as STRUCTURE frequently grouped closely related populations together. Again, RPT performed the worst among the three approaches. Overall, the result demonstrates that simultaneously estimating a tree structure in the model can significantly increase the sensitivity in detecting multiple subtle populations. In addition, the complexity of estimating trees did not appear to create artificial populations.

3.2 MCMC convergence

After integrating out the allele frequency parameters from the tree, the number of markers should not influence the convergence of our MCMC algorithm. As shown in Figure 3, the log likelihood trace of TIPS stabilized after only tens of iterations on the simulated datasets (with all variables updated once per iteration), regardless of the number of markers used (ranging from 500 to 2000). The autocorrelation between our MCMC samples was also small, indicating little dependence between the MCMC samples. As a result, we only need to draw a relatively small number of posterior samples to make reliable inference on populations. In comparison, we have to increase the iterations of STRUCTURE from the default 4000 to 10 000 in order to obtain accurate results on the simulated datasets.

3.3 Application to hapmap dataset

To test our method on genome-wide SNP datasets, we applied TIPS to a set of 83 534 randomly selected autosomal SNPs genotyped from 45 Chinese and 44 Japanese individuals from the International HapMap project (The International HapMap Consortium 2003, 2005). The dataset was obtained from PLINK package (Purcell et al., 2007). After removing non-polymorphic SNPs and SNPs with missing genotypes, the dataset has 71 252 SNPs for each individual across the genome. Allowing a maximum of five populations, TIPS grouped 45 Chinese individuals and one Japanese individual (JPT257) into one cluster, and the remaining 43 Japanese individuals into another cluster. We further tested TIPS with different starting values and allowed at most 2–7 populations, respectively. In all runs, we observed the same results where JPT257 was consistently grouped into the Chinese cluster, indicating that the identified Chinese and Japanese populations probably corresponded to the global mode of our likelihood function. In terms of MCMC convergence, our algorithm converged within about 20 iterations, and overall it took 3 h on a Pentium 1.6 GHz PC to finish 100 iterations on this dataset.

We further ran STRUCTURE and PLINK on the same dataset. We let STRUCTURE detect exactly two populations using 10 000 burn-ins and 10 000 samplings. From three independent runs, the error rates were 33%, 45% and 40%, respectively, which were similar to a random assignment of individuals into two clusters. We found that the Markov chains of STRUCTURE were always trapped in some local modes, and the chains were not able to escape within a practical running period. This issue was expected because the allele frequencies of 71 252 markers estimated by STRUCTURE are strongly correlated with the unknown population origins. Although 20 000 iterations were not long enough for STRUCTURE to converge, it already took 18 h to finish each run. We further ran PLINK on the same dataset, which requires the user to specify a minimum number of populations in the dataset. With a minimum of one population, PLINK grouped all individuals into a single cluster. With a minimum of two populations, PLINK correctly identified the two populations, where JPT257 was also assigned to the Chinese cluster.
3.4 Application to HGDP-CEPH microsatellite dataset

We further applied TIPS to a microsatellite dataset genotyped from the Human Genome Diversity Project–Centre d’Etude du Polymorphism Humain (HGDP-CEPH) worldwide sample of populations (Rosenberg et al., 2002). The dataset contains 783 microsatellite markers of 1048 individuals sampled from 53 populations worldwide. This dataset was originally analyzed by Rosenberg et al. (2002) using STRUCTURE on a subset of 377 microsatellites. They identified six main population clusters corresponding to the main geographical regions of the globe, including Africa, Europe, South and West of Himalaya, Asia, Oceania and America. By applying STRUCTURE within each of the six main clusters, additional subpopulations were further identified.

Since TIPS is designed for SNPs, we first converted microsatellite alleles into dummy SNP alleles, e.g., a 16-allele microsatellite marker can be converted to 15 dummy SNPs, disregarding the numerical ordering of microsatellite alleles. The converted dataset contains 8563 dummy SNPs. To identify as many as 53 populations from this dataset is extremely challenging, if not impossible. We therefore implemented an adaptive search strategy as follows: (1) run TIPS with at most eight populations for several iterations; (2) randomly split the leaf nodes of the reconstructed tree and the associated populations into two, and run TIPS on the new configuration for several iterations and (3) check if the current likelihood is larger than that before splitting the tree; if yes, return to step 2 and continue splitting the tree, otherwise let the Markov chain run for several iterations and start to collect posterior samples.

As shown in Figure 4, TIPS identified 23 population clusters from this dataset. The six main population clusters of the globe were detected by our method and they formed distinct clades in the estimated tree. We also identified many subpopulations within each of the six major population clusters, and they corresponded well to their known origins, such as Yakut and Mongola in East Asia, Kalash in Central South Asia. We observed near perfect separation of populations in Africa, Oceania and America, but only weak separation in Europe and Middle East. In addition, the reconstructed population tree appeared to be roughly consistent with the geographic distributions of the known populations and their migration histories. For example, the African populations formed the outmost clade in the tree in support of the out of Africa model (Stringer and McKie, 1996). Overall, our result was concordant with the previous analysis. On the other hand, we realize that TIPS does not directly handle microsatellite markers, and thus the result should be interpreted with caution.

Interestingly, when we ran STRUCTURE on the converted dataset of 1048 individuals with the same adaptive search strategy, only the six main population clusters and two additional subpopulations in America were identified. This indicates that STRUCTURE may miss subtle stratifications at the tips of a tree when strong stratifications coexist in the data. In comparison, our method can consistently predict individuals’ origins regardless of whether the entire sample or a subset of sample is used. For example, after removing some populations from the data, we still correctly predicted the origins of the remaining individuals (data not shown). The inferred trees also resembled some substructures of the complete tree in Figure 4. We conclude that the consistent results produced by our method were due to the incorporated tree model.

4 DISCUSSION

We introduced TIPS to infer an unknown number of populations. Using both simulated and real datasets, we showed that TIPS performs favorably than STRUCTURE when enough markers are provided. Our method addresses the correlation among populations through a tree structure, such that both distinct and subtle populations can be simultaneously detected. It was shown that modeling correlation among populations through their ancestors can significantly increase the accuracy in predicting population clusters (Falush et al., 2003). In this article, we demonstrate that a tree structure can further increase the prediction accuracy, particularly when both distinct and subtle stratifications are present in the sample.

A practical concern for increasing the model complexity, such as modeling trees, is the computational load. Models like Balding-Nichols require estimating a huge number of nuisance parameters for all genotyped markers (such as ancestral allele frequencies). When applied to trees, the number of nuisance parameters is further multiplied by the number of internal nodes in the tree. As a result, the estimation procedure can be computationally very expensive, not to mention the strong correlation among parameters. We therefore choose to use an alternative approach. We assign markers into groups, with each group representing a unique partition of $K$ populations into non-overlapping clusters. We assume that the markers in the corresponding group take different allele frequencies between population clusters, but share common allele frequencies...
within clusters. Population partitions are based on the tree, such that closely related populations in the tree are more frequently grouped into common clusters than distantly related populations. By doing this, populations are allowed to be correlated through non-segregating markers, but all nuisance parameters can be integrated out. We thus achieve a reasonable MCMC convergence even if hundreds of thousands of markers are analyzed. We demonstrated this point using a genome-wide SNP dataset with 71,252 SNPs. Although the model collapsing strategy has been used by others, such as Pella and Masuda (2007), they did not address the correlation between populations. On the other hand, our method loses some accuracy compared to STRUCTURE when the number of markers is small. However, the loss diminishes very quickly as we increase the marker size. Since the model complexity is invariant to the marker size, our method can explore the space of various population numbers and tree structures more easily. Finally, the autocorrelation between our posterior samples is small, and thus a relatively small number of posterior samples will be sufficient to make reliable inference.

The current TIPS model has some limitations. First, TIPS may perform less well than STRUCTURE when the marker size is relatively small. Our method is more suitable for analyzing datasets with a large number of markers, e.g. at least one thousand randomly selected markers. Second, we define the tree branch length as the proportion of segregating markers between populations. As estimated from data, it can be affected by the sample size and thus complicates the interpretation of the tree. Although in our simulation, the estimated trees did agree very well to the simulated trees, it is important to realize that our tree model is a simplification from the reality, and thus interpretation of trees should be carried out with caution. For possible extensions to the current model, we currently assume a uniform prior on trees, which is the least informative prior. In practice, it may be desirable to use unbalanced priors such that trees better resembling population evolution can be preferred. In addition, a potentially better approach to analyze densely distributed markers is to use haplotypes rather than treating them independently. Similar extension should be made to handle microsatellites, as using dummy SNPs may lose valuable information about population structures. Recent migration of populations has created admixture of genetic variants in individuals’ genome in some populations. To account for admixture effect, we need to define population origins on the genome level instead of on individual’s level. A hidden Markov model can be used for the transition across populations along individual’s genome. Finally, when applying TIPS to population-based case-control datasets, we should use two population prevalence parameters, one for cases and one for controls; otherwise the inferred origins can be biased (Kohler and Bickeboller, 2005).

ACKNOWLEDGEMENTS
The author thanks Dr James L. Rosenberg and two anonymous reviewers for their invaluable comments on the manuscript.

Conflict of Interest: none declared.

REFERENCES