Subject Section

Ohana, a tool set for population genetic analyses of admixture components

Jade Yu Cheng¹,²,³*, Thomas Mailund¹ and Rasmus Nielsen²,³

¹Bioinformatics Research Centre, Aarhus University, Aarhus 8000 Denmark.
²Departments of Integrative Biology and Statistics, University of California, Berkeley, Berkeley, CA 94720, USA.
³Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, Øster Voldgade 5-7, Copenhagen 1350 Denmark.

* To whom correspondence should be addressed.
Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Structure methods are highly used population genetic methods for classifying individuals in a sample fractionally into discrete ancestry components. Contribution: We introduce a new optimization algorithm of the classical Structure model in a maximum likelihood framework. Using analyses of real data we show that the new optimization algorithm finds higher likelihood values than the state-of-the-art method in the same computational time. We also present a new method for estimating population trees from ancestry components using a Gaussian approximation. Using coalescence simulations modeling populations evolving in a tree-like fashion, we explore the adequacy of the Structure model and the Gaussian assumption for identifying ancestry components correctly and for inferring the correct tree. In most cases, ancestry components are inferred correctly, although sample sizes and times since admixture can influence the inferences. Similarly, the popular Gaussian approximation tends to perform poorly when branch lengths are long, although the tree topology is correctly inferred in all scenarios explored. The new methods are implemented together with appropriate visualization tools in the computer package Ohana. Availability: Ohana is publicly available at https://github.com/jade-cheng/ohana. Besides its source code and installation instructions, we also provide example workflows in the project wiki site. Contact: jade.cheng@birc.au.dk

1 Introduction

To quantify population structure, researchers often use methods based on the Structure model (Pritchard et al., 2000). The basic assumption in this model is that individuals belong to a set of $K$ discrete groups, each with unique allele frequencies and obeying Hardy-Weinberg Equilibrium, although the latter assumption can be relaxed (Gao et al., 2007). Furthermore, individuals are allowed to have fractional memberships of each group. The groups are often termed ‘ancestry components’ and are sometimes interpreted to represent ancestral populations. This interpretation may be correct in some scenarios, for example when analyzing balanced samples of recently admixed individuals from otherwise highly divergent groups. However, if basic model assumptions are violated, for example if populations truly are not discrete units, the interpretation is more unclear. Nonetheless, inferences under the Structure model have proven highly popular for quantifying population genetic variation and for exploring the basic structure and divisions of genetic diversity in a sample.

STRUCTURE (Pritchard et al., 2000), FRAPPE (Tang et al., 2005), and ADMIXTURE (Alexander et al., 2009) are arguably the three most commonly used programs that apply the Structure model. STRUCTURE uses a Bayesian approach and relies on a Markov Chain Monte Carlo (MCMC) algorithm to sample jointly the posterior distribution of allele frequencies and fractional group memberships. FRAPPE uses a maximum likelihood approach and relies on a Markov Chain Monte Carlo (MCMC) algorithm to sample jointly the posterior distribution of allele frequencies and fractional group memberships. FRAPPE uses a maximum likelihood approach and optimizes the likelihood for both allele frequencies and fractional group memberships using an expectation-maximization (EM) algorithm. ADMIXTURE uses the same model and statistical framework as FRAPPE but uses a faster optimization algorithm. ADMIXTURE executes a two-stage process, first taking a few fast EM steps and then executing a sequential quadratic programming (QP) algorithm. ADMIXTURE uses a pivoting algorithm to solve each QP
problem and applies a quasi-Newton acceleration to each iteration. This acceleration does not respect parameter bounds. ADMIXTURE projects an illegal update to the nearest feasible point, and the acceleration step contributes only when it results in a better likelihood; otherwise the original QP update is used.

The interpretation of parameter estimates under the structure model is somewhat contentious (Royal et al., 2010; Weiss and Long, 2009). It is not clear exactly what the groups, or ancestry components, represent, but in the most simple interpretation we can think of them as estimates of some idealized ancestral populations. If a researcher has inferred the existence of $K$ ancestral populations and knows the fractional memberships of each individual in these populations, a next question would be to explore their evolutionary history. The estimated allele frequencies can provide information about this.

The first approaches for using allele frequencies to estimate population histories dates back to the seminal work by Edwards and Cavalli-Sforza (Cavalli-Sforza et al, 1967). They used Gaussian models for the joint distribution of allele frequencies of multiple populations to estimate genetic distances and to infer population trees. The use of Gaussian models to approximate genetic drift has recently had a resurgence after the availability of large Single Nucleotide Polymorphism (SNP) data sets. It is used in numerous methods and studies, including tests of the availability of large Single Nucleotide Polymorphism (SNP) data (Cavalli-Sforza et al, 1994).

However, a Gaussian model for the allele frequencies corresponds to an assumption of a Brownian motion process. This is a limitation of the approach implemented in Ohana and in other approaches that use Brownian motion models to approximate the Wright-Fisher diffusion.

### 2 Methods

Ohana’s qgas program infers admixture using genotype observations stored in the ped format from Plink (Purcell et al., 2007) or genotype likelihoods in the bgf format from beagle (Browning et al., 2007). Ohana’s nemeo program infers population covariances, and Ohana’s convert program facilitates different stages of the analysis by providing file conversions and fast approximations. The source code, installation instructions, and example workflows are available on GitHub at https://github.com/zheng-ohana.

#### 2.1 Statistical Models

The likelihood model using genotype observations is given by

$$
\ln L^g(Q, P) = \sum_{i,j} \ln \left( \sum_k q_{ik} A_j^k + (1 - q_{ik}) B_j^k \right),
$$

where $K$ is the number of ancestry components, $I$ is the number of individuals, and $J$ is the number of polymorphic sites. This is the same as the model used in STRUCTURE (Pritchard et al., 2000), FRAPPE (Tang et al., 2005), ADMIXTURE (Alexander et al., 2009), and SPA (Yang et al., 2012). Using the model in NGSADMIX (Skotte et al., 2013), qgas can also work on genotype likelihoods. In that case the likelihood model is given by

$$
\ln L^p(Q, P) = \sum_{i,j} \ln \left( \sum_k q_{ik} A_j^k + (1 - q_{ik}) B_j^k \right),
$$

where

$$
A_j^k = \sum_{aA} g_{aj}^k A_a A_a,
$$

and

$$
B_j^k = \sum_{aa} g_{aj}^k A_a A_a
$$

are the probabilities of observing the sequence data at the $i$th individual’s $j$th marker, conditioned on genotypes $AA$, $Aa$ or $aa$, respectively. This representation assumes markers with two alleles, although it could easily be generalized to multiple alleles. The advantage of working on genotype likelihoods instead of called genotypes is that genotype likelihoods incorporate the uncertainty regarding genotype calls inherent in much NGS data, and this makes it more applicable to low- or medium-coverage data (see e.g., Skotte et al., 2013).

To infer population histories, Ohana models the joint distribution of allele frequencies across all ancestry components as a multivariate Gaussian similar to TREEMIX (Pickrell et al., 2012) and Bayescan (Gunter et al., 2013). The covariance matrix $\Omega$ of dimension $K \times K$ is assumed to be constant among all sites, and the process has a mean $\mu_j$ at site $j$. The joint distribution of allele frequencies is then given by

$$
P(f_j | \Omega, \mu_j) \sim \mathcal{N}(\mu_j, \mu_j (1 - \mu_j) \Omega).
$$

This system is under-determined (see e.g., Felsenstein, 2004 chapter 23), i.e. multiple covariance matrices induce the same probability distribution on the allele frequencies. Similar to Felsenstein’s restricted maximum likelihood approach (Felsenstein, 1981), we therefore root the tree in one of the observations corresponding to conditioning on the allele frequencies in one of the populations when calculating the joint distribution.
of allele frequencies in the other populations. We emphasize that the
rooting is arbitrary but that it does not imply any assumptions of this
population actually being ancestral (for time reversible models). We
then obtain a new covariance matrix $\Sigma'$, which has size $(K - 1) \times (K - 1)$
and a joint density of the form

$$
\ln[p_2(F)] = \ln \left( \prod_{i=1}^{K-1} \frac{1}{\sqrt{2\pi}\sigma_i} \exp \left( -\frac{1}{2\sigma_i^2} (f_i - f_i')^2 \right) \right)
$$

where $f_i = f_i - f_0$,

$$
\frac{\partial}{\partial \sigma_i} \ln[p_2(F)] = \sum_{i=1}^{K-1} \left( f_i - f_i' \right) \frac{1}{\sigma_i} (f_i - f_i')
$$

2.2 Parameter Inference

2.2.1 Inference for individual ancestries

To estimate $Q$ and $F$, we use Newton’s approach. In general, we
can approximate a function $F(x)$ with its second order Taylor expansion.
We proceed to second-order approximation by solving $\Delta \Sigma$. In our
problem, $\Delta Q$ and $\Delta F$ are constrained by $\sum m q_k \Delta q_k +
\Delta \Psi_1 = 0$, $\sum k \Delta q_k = 0$, and $i, j, k, j'$. The analytical forms of the differential for
ln $[p_2(Q, F)]$ are presented below:

$$
\frac{\partial^2 \ln[p_2(Q, F)]}{\partial q_i \partial q_j} = \frac{1}{\sum_{k=1}^{K-1} (\sum_{m=1}^{Q} \gamma_{km} f_{km} + \sum_{j'=1}^{K-1} q_{j'} \omega_{ij} + \sum_{j'=1}^{K-1} q_{j'} \Phi_{ij}) (f_{ij} - f_{ij}')}{\sum_{k=1}^{K-1} (\sum_{m=1}^{Q} \gamma_{km} f_{km} + \sum_{j'=1}^{K-1} q_{j'} \omega_{ij} + \sum_{j'=1}^{K-1} q_{j'} \Phi_{ij})}
$$

The analytical forms of the differential for ln $[p_2(Q, F)]$ can also
be found below. For both ln $[p_2(Q, F)]$ and ln $[p_2(Q, F)]$, most
off-diagonal values of the Hessians diminish. Leveraging this block
structure, we convert the problem from manipulating huge matrices into
manipulating sequences of small matrices of size $K$.

$$
\frac{\partial^2 \ln[p_2(Q, F)]}{\partial q_i \partial q_j} = \frac{1}{\sum_{k=1}^{K-1} (\sum_{m=1}^{Q} \gamma_{km} f_{km} + \sum_{j'=1}^{K-1} q_{j'} \omega_{ij} + \sum_{j'=1}^{K-1} q_{j'} \Phi_{ij}) (f_{ij} - f_{ij}')}{\sum_{k=1}^{K-1} (\sum_{m=1}^{Q} \gamma_{km} f_{km} + \sum_{j'=1}^{K-1} q_{j'} \omega_{ij} + \sum_{j'=1}^{K-1} q_{j'} \Phi_{ij})}
$$

2.2.2 Inference for population covariances

To optimize the likelihood model defined in the last equation of section
2.1, we use a black-box style of optimizer, the Nelder-Mead (NM)
simplex method (Nelder & Mead, 1965). We use sample covariances,
$S_{ij} = \sum (x_i - \bar{x}_i)(x_j - \bar{x}_j)$, as the initial starting point for the
NM optimizer, and we use Cholesky decomposition (Cholesky, 1910) to
determine the positive semi-definiteness and to compute matrix inverses
and determinants. The nemecpo program in Ohana performs this analysis.
High-level pseudo-code of this algorithm appears in SI Algorithm 2.

2.3 Estimation of phylogenetic trees

With the estimated covariance matrix in hand, we can construct a
phylogenetic tree. We use the Neighbor-Joining (NJ) method for this,
taking advantage of the NJ theorem (Saitou and Masatoshii, 1987), which
states that when a distance matrix is compatible with a phylogenetic
tree, this tree will be accurately reconstructed by the NJ method. To do so, we
first transform the covariance matrix to a distance matrix by observing the
distance between two populations is given by $D_{ij} = \|p_i - p_j\| = \sqrt{\sum_{j=1}^{Q} (p_{ij} - p_{ij})^2}$.

Notice that there is a one-to-one correspondence between the
covariance matrix and distances. These distances are then fed to the NJ
algorithm. Ohana’s convert program performs all of these steps and in
addition, provides an option to render the tree as SVG.
3.3 Model limitations

There are at least three reasons why tree estimation using a Gaussian model based on estimated allele frequencies may face challenges. First, the allele frequencies are treated as observed data, but they are truly estimates. This has the potential for introducing a variety of biases. Second, the use of a Brownian motion model to approximate genetic drift is inaccurate near the boundaries and for long divergence times, likely leading to underestimates of the lengths of long branches. Third, due to differences in sample sizes
for different populations, the Structure model may not identify groups that correspond to natural units of a tree, even when the populations truly have evolved in a tree-like fashion.

We explore some of these issues in the following simulation study (Figure 4) by simulating trees with different divergence times: short, medium, and long. For very short divergence times (Figure 4-a), the covariance matrix was estimated poorly because of the small differences in allele frequencies across populations. This in turn leads to reduced accuracy in the estimation of the tree. While the topology is recovered correctly, the lengths of the external branches are overestimated. This likely happens because the Structure model tends to maximize allele frequency differences for finite sample sizes, i.e. the estimated difference in allele frequencies between pairs of populations tends to be larger than the true difference. This is an issue that can be mitigated with larger sample sizes and tends to be a problem only when branch lengths are very small. Nonetheless, it will likely affect many real data analyses.

In the medium-length divergence scenario (Figure 4-b), neither of these problems affect the inferences, and the estimates of the branch lengths are therefore quite close to the true values. In all three divergence scenarios, the tree topologies were always estimated accurately.

<table>
<thead>
<tr>
<th>K</th>
<th>Dataset #1</th>
<th>Dataset #2</th>
<th>Dataset #3</th>
<th>Dataset #4</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-1967733</td>
<td>-1967733</td>
<td>-385358</td>
<td>-385356</td>
</tr>
<tr>
<td></td>
<td>-1957667</td>
<td>-1956799</td>
<td>-379987</td>
<td>-379988</td>
</tr>
<tr>
<td>3</td>
<td>-1946218</td>
<td>-1946244</td>
<td>-378859</td>
<td>-378860</td>
</tr>
<tr>
<td></td>
<td>-1935775</td>
<td>-1936205</td>
<td>-377735</td>
<td>-377736</td>
</tr>
<tr>
<td>4</td>
<td>-1925636</td>
<td>-1925877</td>
<td>-376558</td>
<td>-376564</td>
</tr>
<tr>
<td></td>
<td>-1915552</td>
<td>-1915743</td>
<td>-375585</td>
<td>-375586</td>
</tr>
<tr>
<td>5</td>
<td>-1905430</td>
<td>-1905638</td>
<td>-374622</td>
<td>-374541</td>
</tr>
<tr>
<td></td>
<td>-1895372</td>
<td>-1895879</td>
<td>-373524</td>
<td>-373607</td>
</tr>
<tr>
<td>6</td>
<td>-1885306</td>
<td>-1885466</td>
<td>-372558</td>
<td>-372562</td>
</tr>
<tr>
<td></td>
<td>-1875503</td>
<td>-1875853</td>
<td>-371543</td>
<td>-371517</td>
</tr>
<tr>
<td>7</td>
<td>-1865492</td>
<td>-1865965</td>
<td>-370609</td>
<td>-370715</td>
</tr>
<tr>
<td></td>
<td>-1855502</td>
<td>-1856262</td>
<td>-369753</td>
<td>-369851</td>
</tr>
<tr>
<td>8</td>
<td>-1845732</td>
<td>-1846490</td>
<td>-368890</td>
<td>-368912</td>
</tr>
<tr>
<td></td>
<td>-1836315</td>
<td>-1836775</td>
<td>-368109</td>
<td>-368082</td>
</tr>
</tbody>
</table>

Table 1. A table of the highest log likelihoods achieved from ADMIXTURE and the qpas program in Ohana for a range of K values. For each data set, each program, and each value of K, we executed 100 times using random seeds (0, 1, ..., 99) and chose the highest value found in any run. This mimics the procedure often used for real data analysis. In the vast majority of cases, the qpas program in Ohana found significantly higher likelihood values than ADMIXTURE. Dataset #1 is a compilation of Europeans containing 17,507 markers and 118 individuals. Dataset #2 is the benchmark dataset used in ADMIXTURE (Alexander et al., 2009) containing 324 CEU, YRI, MEX, and ASW individuals and 13,922 markers. Dataset #3 is a compilation of 171 Han Chinese samples and 9,822 markers. Dataset #4 is a worldwide population of 60 individuals and 4,695 markers.

3.4 Other simulation scenarios

We also evaluated the performance of the method under several other simulation scenarios, and the results are presented in SI Section 2 to 5. A few noteworthy observations include: (1) In more than one simulation scenario with ancient admixture, the population was not inferred to be admixed but received a unique admixture component, SI Section 2 Figure 4 and Section 3 Figure 5. The probability of inferring admixture likely depends on the amount of drift since admixture. In the context of much human data showing evidence of ancient admixture, it might be worthwhile in future studies to explore how much drift after admixture is required to erase the signal of admixture. (2) When K is smaller than the true number of ancestry components, populations with few individuals represented in the sample tend to be (wrongly) inferred as admixed, SI Section 5 Figure 7. There is a clear dependence on sample size in inferences of admixture components in the Structure model. Similarly, the outgroup tends to be identified as the first admixture component that splits from the rest of the individuals, only when the outgroup is well-represented in the sample in terms of the number of individuals.

3.5 Real data analysis

To illustrate the method, we apply it to the panel of global human data described in the Methods section (Figure 5), using a range of K values. The topologies of the trees largely mimic what is already known about human ancestry (e.g., (Reich et al., 2012)), i.e. using a root in Africa, Asians and Native Americans cluster together, the European and middle Eastern groups cluster together, etc. In addition to Yorubans having a long branch because this group is an outgroup to the rest, we also notice a relatively long branch leading to Native Americans, reflecting the increased drift...
in this group due to the bottleneck into the Americas and possibly small population sizes thereafter.

4 Discussion

In this paper, we introduced a new implementation of the Structure model in a maximum likelihood framework. We compared the new optimization algorithm to the one implemented in the hitherto fastest program, ADMIXTURE. The qpas program in our software, Ohana, generally outperformed ADMIXTURE by obtaining estimates with higher likelihood values in similar computational time.

In addition, we presented a new approach for estimating trees for ancestry components. Using coalescence simulations, we showed that when the trees are interpreted as reflecting true population trees, external branch lengths tend to be overestimated for small divergence times.

However, for long divergence times, the use of a Gaussian model and its inaccuracy in approximating genetic drift cause branch length estimates to be downward biased. Nonetheless, the estimates of tree topology appear reasonably robust. The tree estimation and visualization tool should be of use to other researchers as an additional possible component of a Structure model analysis of the data. The tree is a visualization of the covariance structure of the admixture components, and it may as such be useful even if a strict interpretation of a evolutionary tree may not be warranted. There might be several reason why such an interpretation may not be appropriate, most of all because the true nature of the evolution of the ancestry components may not be well-described by a tree. Ancestry components are constructions that may or may not reflect true ancestral populations.

Acknowledgements

This work is funded by the Danish Council of Independent Research Sapere Aude grant 12-125062; Conflict of Interest: none declared.

References


Alexander, David H., and Kenneth Lange. “Enhancements to the ADMIXTURE algorithm for individual ancestry estimation.” BMC
short Title

**Fig. 4.** A simulation study for different divergence times. We simulated 140 individuals in 7 groups, 20 individuals per group. The first 6 groups were un-admixed. The last group was an equal mixture of the first 3 groups. We illustrate the simulated demography on the top. We simulated 6 divergence scenarios, 2 short shown in (a), 2 medium shown in (b), and 2 long shown in (c). From the shortest to the longest divergence scenario (top to bottom), the split times \( t_0, t_1, t_2, t_3 \) in generation were: (10, 20, 30, 40), (100, 200, 300, 400), (1000, 2000, 3000, 4000), (10000, 20000, 30000, 40000). We simulated 6 divergence scenarios, 2 short shown in (a), 2 medium shown in (b), and 2 long shown in (c). From the shortest to the longest divergence scenario (top to bottom), the split times \( t_0, t_1, t_2, t_3 \) in generation were: (10, 20, 30, 40), (100, 200, 300, 400), (1000, 2000, 3000, 4000), (10000, 20000, 30000, 40000). We simulated 6 divergence scenarios, 2 short shown in (a), 2 medium shown in (b), and 2 long shown in (c). From the shortest to the longest divergence scenario (top to bottom), the split times \( t_0, t_1, t_2, t_3 \) in generation were: (10, 20, 30, 40), (100, 200, 300, 400), (1000, 2000, 3000, 4000), (10000, 20000, 30000, 40000).

**Fig. 5.** Analysis of human global data. We used a data set compiled from the HGDP project containing 86 individuals from 8 populations, 10 per population. We filtered markers using Plink (Purcell et al. 2007) with options –indep 50 5 2 –geno 0.0 –maf 0.05. A total of 125,787 markers survived the filtration and were used for the analysis. For each \( K \) value, we dispatched 32 executions with random seeds from 0 to 31. We report only results from the execution that reached the best likelihood for each \( K \). The plots show individual admixture proportions and population trees for several different values of \( K \). The map combines the admixture results and geographical records of the HGDP samples. Each slice of each pie chart shows the sum of one component estimated in samples collected at that region. (a, b, c, and d) show the admixture and tree estimates for \( K = 2, 4, 6, 8 \), respectively.


Purcell, Shaun, Benjamin Neale, Kateh Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses." The American Journal of Human Genetics 81, no. 3 (2007): 559-575.


