Lisa N. Barrow
9 October 2015
Practical Inference in Pop Gen and Phylogenetics

Application of TreeMix to simulated datasets with known migration events

Introduction

Inferring population history from genetic data is an important, but challenging endeavor in biology. Clustering methods (e.g., Structure—Pritchard et al. 2000) can be useful for inferring population membership and admixture, but they do not provide information about many demographic parameters of interest. Alternative models are needed to estimate parameters such as population divergence, gene flow, and changes in population size.

One approach involves modeling populations under different demographic scenarios to estimate parameter values of interest (e.g., ∂a∂I—Gutenkunst et al. 2009; IMa2—Hey 2010), but these methods are often limited to a small number of populations. Another approach that can accommodate many populations involves depicting population relationships as a bifurcating tree, but gene flow among populations is a major violation of tree model assumptions (Leaché et al. 2014). Pickrell and Pritchard (2012) address these shortcomings by developing a statistical model implemented in the software TreeMix. The method enables the estimation of population trees to model divergences, or splits, and also infers gene flow, or mixture, between diverged populations.

Methods

Genetic Data

The data used in this study were simulated by P. Beerli. Two simulated datasets containing diploid individuals were used to explore the utility of TreeMix for population scenarios with varying levels of gene flow. The genetic data for each scenario consisted of bi-allelic SNPs (single nucleotide polymorphisms) from 1,000 loci that were 500 base pairs in length. Multiple SNPs were recorded per locus resulting in final datasets of 10,987 SNPs (first scenario described below) and 23,708 loci (second scenario).

Four-population Scenario

The first scenario included four populations with multiple individuals sampled per population (pop1 = 40, pop2 = 20, pop3 = 10, pop4 = 5). The model was a linear stepping-stone with bi-directional migration between the first two populations, and the direction of migration from pop4→pop3→pop2→pop1. TreeMix was used to first build the maximum likelihood tree. To account for linkage disequilibrium (LD), i.e., the fact that nearby SNPs are not independent, SNPs were grouped together in windows, or bins of specified length. Based on the locus length and the recommendation from the TreeMix manual that the window far exceed the known extent of LD, a window size of 500 was used. Migration events were then added to the tree one at a time, adding from one to four mixture events. Standard errors of migration rates were calculated with the "-se" option, and the "-global" option was used to do a round of global rearrangements after adding all populations. The tree graph and residuals were visualized using R (R Development Core Team 2015).

The three-population and four-population tests introduced by Keinan et al. (2007) and detailed in Reich et al. (2009) were used to further evaluate the data. The three-population test is of the form f_3(A;B,C) and tests the "treeness" of the data, such that a significantly negative value indicates the population A is admixed. The four-population test is of the form f_4(A,B;C,D), and a significantly non-zero value indicates gene flow in the tree. All possible trees for the four-population scenario were evaluated.

Eight-population Scenario

The second scenario consisted of 8 populations with 10 individuals per population. The model was a linear stepping-stone model in a single direction, with a 10-fold reduction in migration in comparison to the first scenario. The same parameter settings as above were used to build the maximum likelihood tree. Migration events were added to the tree from one to seven migration events, and the residuals were visualized.

Results and Discussion

Four-population Scenario

The best maximum likelihood tree suggests pop1 and pop2 are most closely related to each other, and pop3 and pop4 are most closely related to each other (Fig. 1). The residuals show, however, that there is high standard error between some population pairs, suggesting these may be candidates for admixture events. For example, the residual between pop2 and pop3 is greater than 0, although the standard error is still low (less than 0.1; Fig. 1). When migration events (up to four events) are added to the tree, only a single migration edge is found and modeled. This mixture event occurs from pop3 to pop2, with a migration weight of 0.5% (Fig. 2). The residuals suggest pop1 and pop4 may also be more closely related to each other in the data than in the best-fit tree.

The three-population test did not result in significant Z-scores for any of the population combinations. The most negative value was for the tree in which pop2 is admixed from pop1 and pop3, which is consistent with the modeled scenario. A lack of negative Z-scores in the three-population test does not necessarily mean there is no admixture. According to Reich et al. (2009), it could reflect the fact that there has been substantial genetic drift in the groups since mixture.

The four-population test suggested the best tree topology was (pop1,pop2;pop3;pop4), but it is not entirely consistent with the data (non-significant Z-score). This makes sense because these dichotomous tree structures are probably not a suitable way to model the step-wise population scenario.

In short, although the best ML tree structure is partially consistent with the step-wise scenario used to generate the data, i.e., neighboring populations are more closely related, the evidence for mixture events is fairly weak. The population scenario being considered here may not be a suitable model to test in TreeMix. It may be interesting to compare other methods that model migration (e.g. Migrate-n—Beerli and Palczewski 2010; IMa2—Hey 2010) using the same simulated dataset to determine whether this particular scenario is difficult to elucidate.

Eight-population Scenario

The best ML tree resembles a ladder that is somewhat consistent with the stepwise scenario used to simulate the data (Fig. 3). The residuals indicate some candidate pairs for admixture events including pop2 and pop3, pop3 and pop4, pop4 and pop5, and pop5 and pop6.

However, there is also high error between pop1 and pop7, and pop1 and pop8, which are not entirely consistent with the simulated data.

The first added migration edge goes from pop8 to the node between pop1 and pop2 with a weight of 17.9% (Fig. 4). The second migration edge is from pop2 to pop3 and has a weight of 46.3% (Fig. 5). In this tree with two modeled migration edges, the first migration edge increased in weight to 23.2%. The other migration edges (Fig. 6, 7) are not consistent with the simulated data and the stepping-stone population model.

This second scenario may also not be an appropriate model that TreeMix is designed to test, but it is informative to evaluate different scenarios since the underlying history of empirical data from natural systems can never be known with certainty. Additional simulations that evaluate the conditions under which TreeMix performs well, including shallow or old population divergences, single admixture events versus constant migration, the magnitude of migration, and numbers of mixture events would be useful. TreeMix also has a function that can incorporate known migration events, which could be used in future simulation work.

Table 1. Three-population test for "treeness". A significantly negative value (Z << -2) of the f_3 statistic implies that the first population is admixed. Described in detail in Reich et al. (2009).

Populations Used	**f_3 statistic**	**SE**	**Z-score**
pop1;pop2,pop3	0.0114797	0.00169613	6.76816
pop2;pop1,pop3	-0.000690503	0.00148198	-0.465934
pop3;pop1,pop2	0.160799	0.00514016	31.2828
pop1;pop2,pop4	0.0108277	0.00171415	6.31665
pop2;pop1,pop4	-3.85E-05	0.00150882	-0.0255215
pop4;pop1,pop2	0.198249	0.00592033	33.4861
pop1;pop3,pop4	0.155434	0.00589145	26.383
pop3;pop1,pop4	0.0168439	0.00196486	8.57257
pop4;pop1,pop3	0.0536418	0.00285921	18.7611
pop2;pop3,pop4	0.143916	0.00551487	26.096
pop3;pop2,pop4	0.0161919	0.00178342	9.07917
pop4;pop2,pop3	0.0542938	0.00283503	19.1511

Table 2. Four-population test for "treeness" introduced by Keinan et al. (2007) and detailed in Reich et al. (2009). If the f_4 statistic is consistent with 0 (significant when |Z|>>2), this suggests the best tree topology. Significantly non-zero values indicate gene flow in the tree.

Population Used	**f_4 statistic**	**SE**	**Z-score**
pop1,pop2;pop3,pop4	-0.000651996	0.000410947	-1.58657
pop1,pop3;pop2,pop4	0.143955	0.00556332	25.8757
pop1,pop4;pop2,pop3	0.144607	0.00549083	26.336

Figure 1. (Top) Maximum likelihood tree for four-population scenario. The drift parameter reflects the amount of genetic drift that has occurred between populations. (Bottom) Residual fit from the maximum likelihood tree. The colors represent the Standard Error as the gradient depicted in the top left (light colors are close to 0; dark colors have error of greater magnitude.

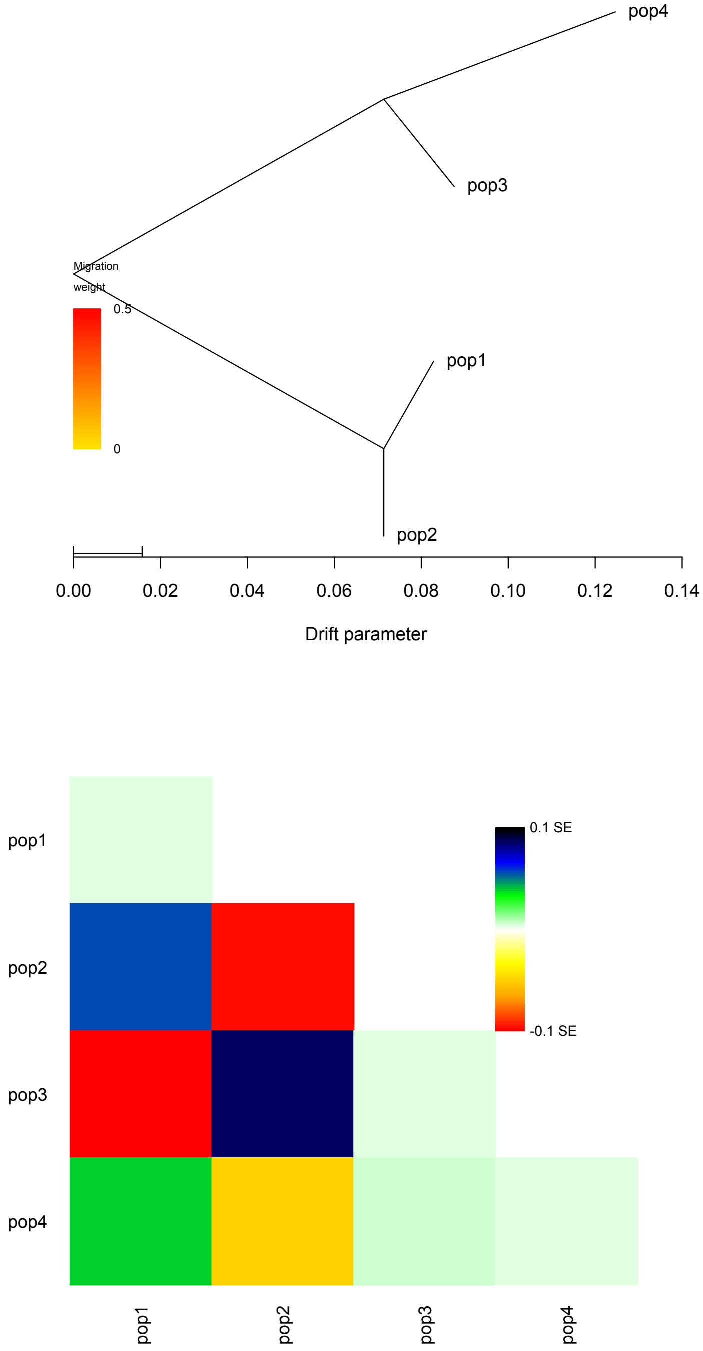


Figure 2. (Top) Maximum likelihood tree with added migration for four-population scenario. (Bottom) Residual fit from the maximum likelihood tree with migration.

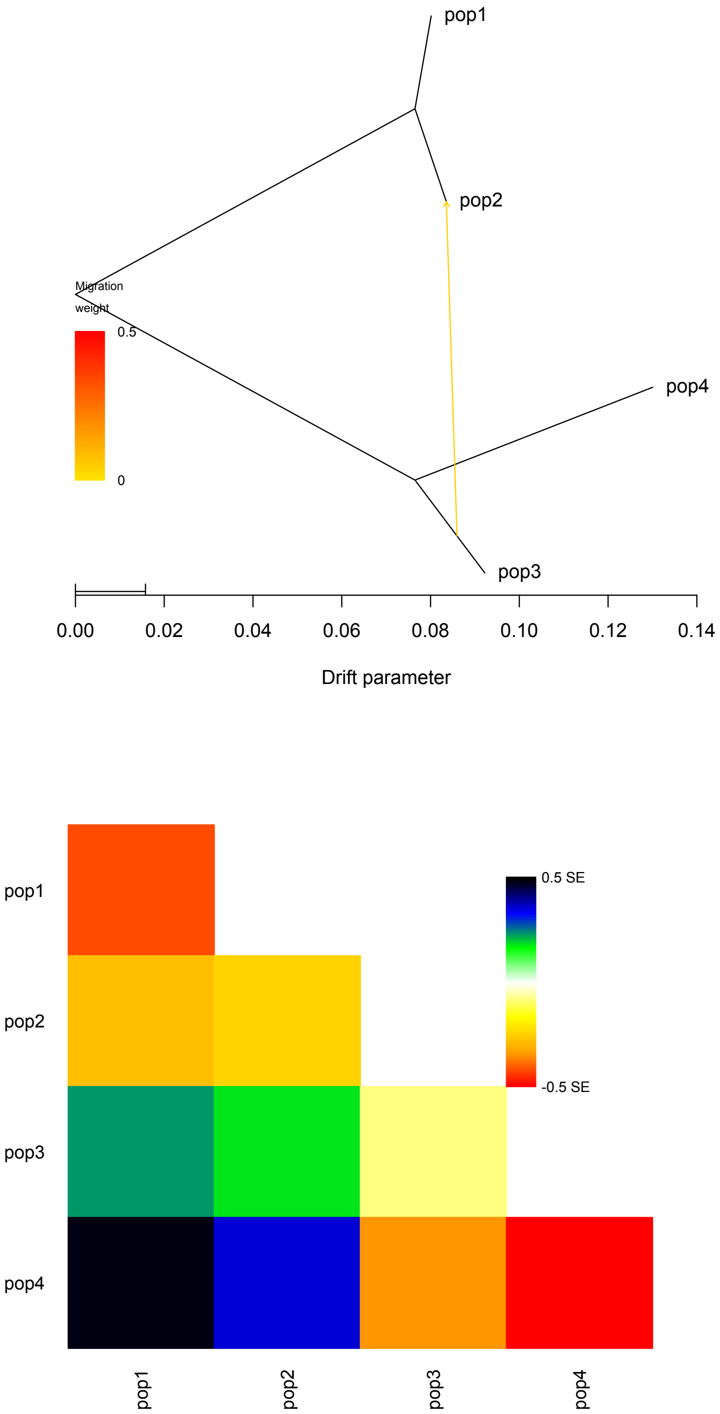


Figure 3. (Top) Maximum likelihood tree for eight-population scenario. (Bottom) Residual fit from the maximum likelihood tree.

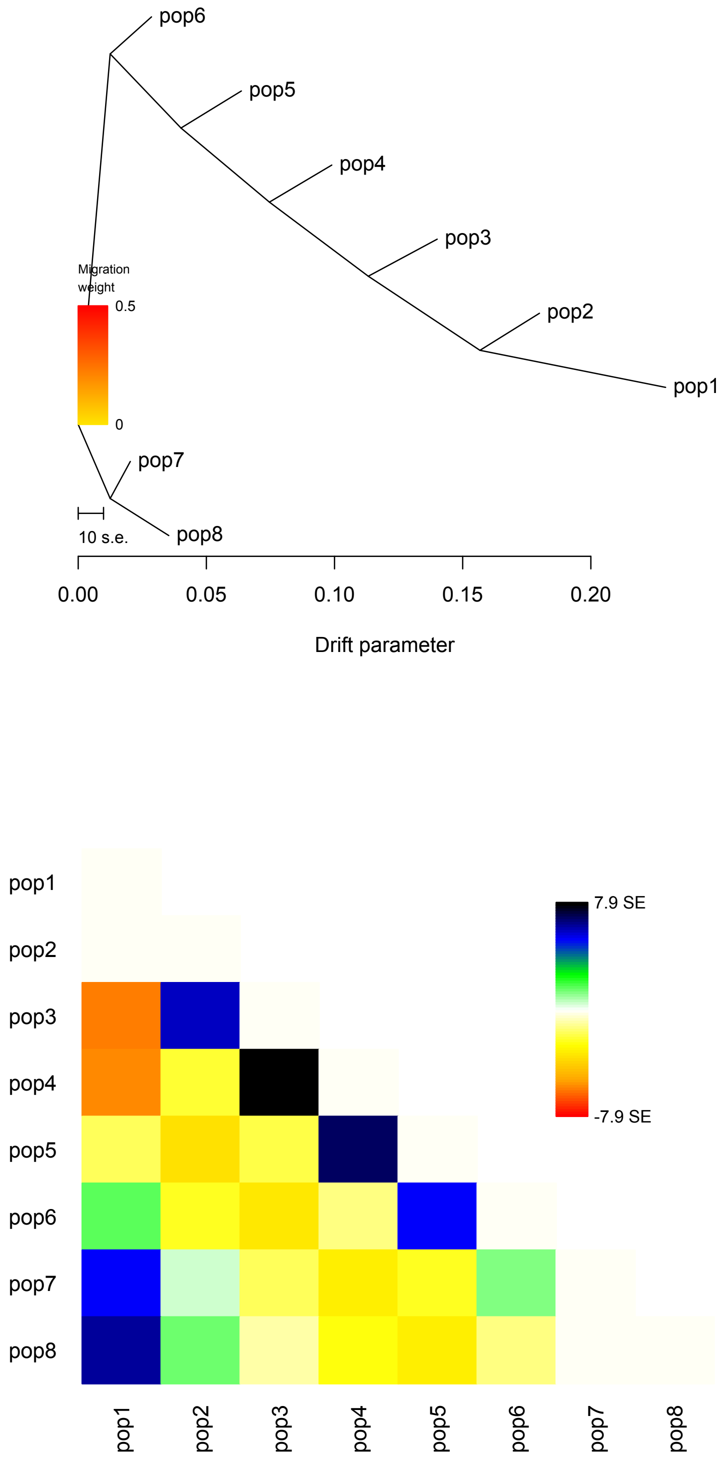


Figure 4. (Top) Maximum likelihood tree with one migration event for eight-population scenario. (Bottom) Residual fit from the maximum likelihood tree.

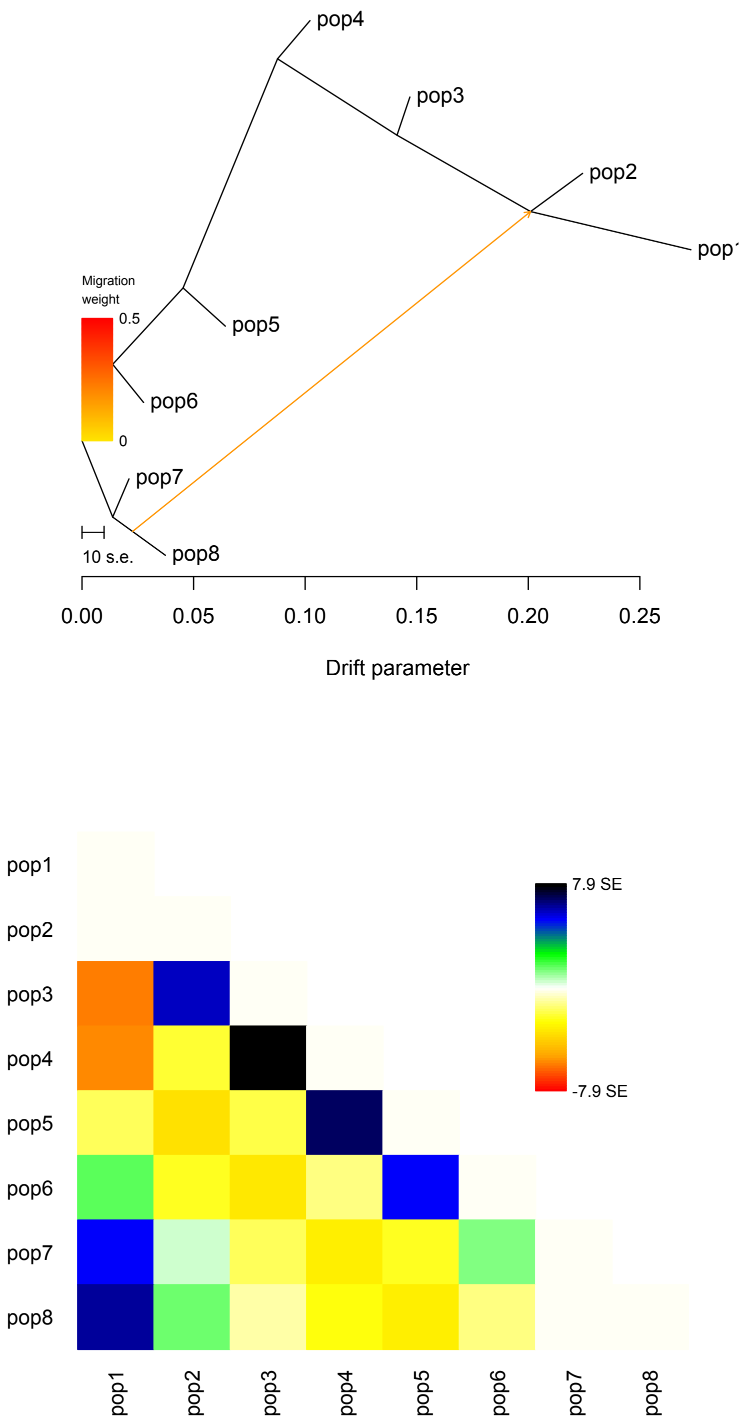


Figure 5. (Top) Maximum likelihood tree with two migration events for eight-population scenario. (Bottom) Residual fit from the maximum likelihood tree.

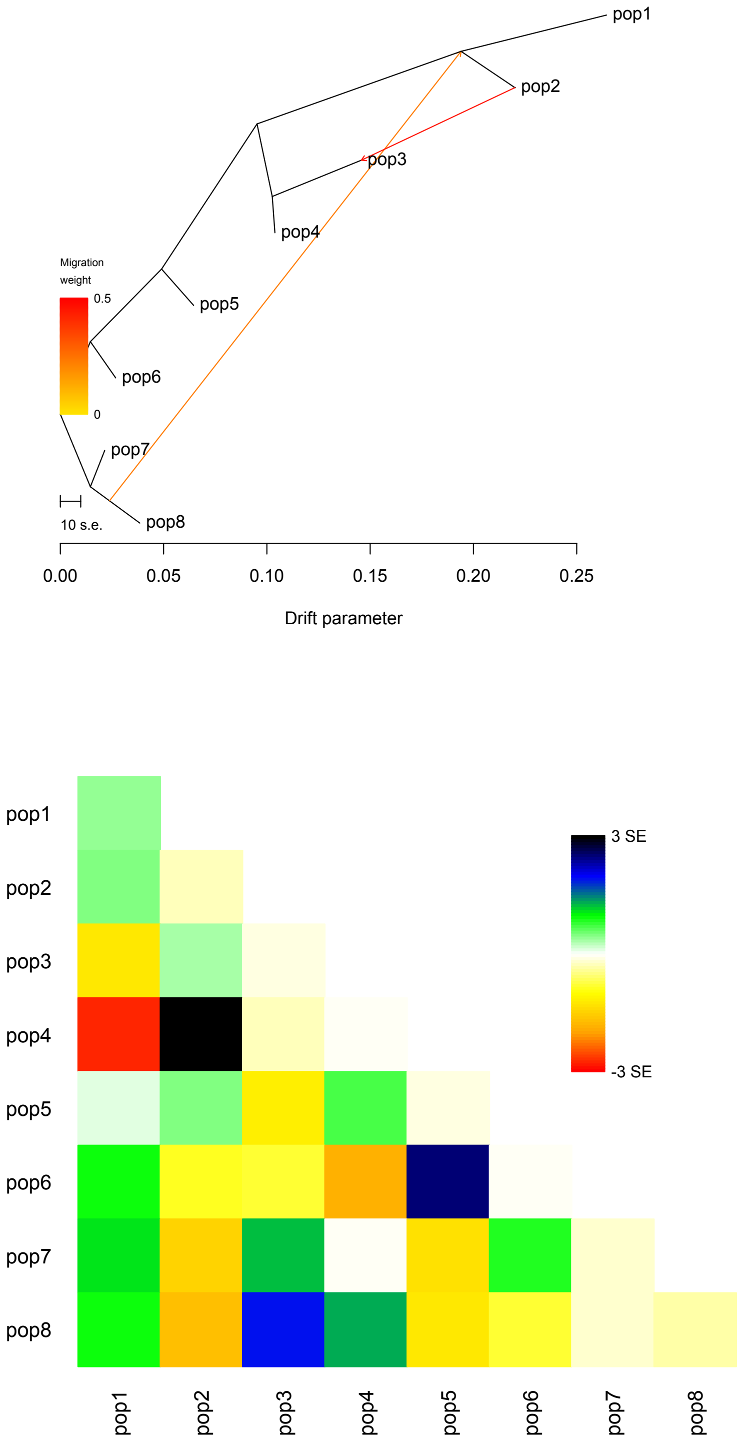


Figure 6. (Top) Maximum likelihood tree with three migration events for eight-population scenario. (Bottom) Residual fit from the maximum likelihood tree.

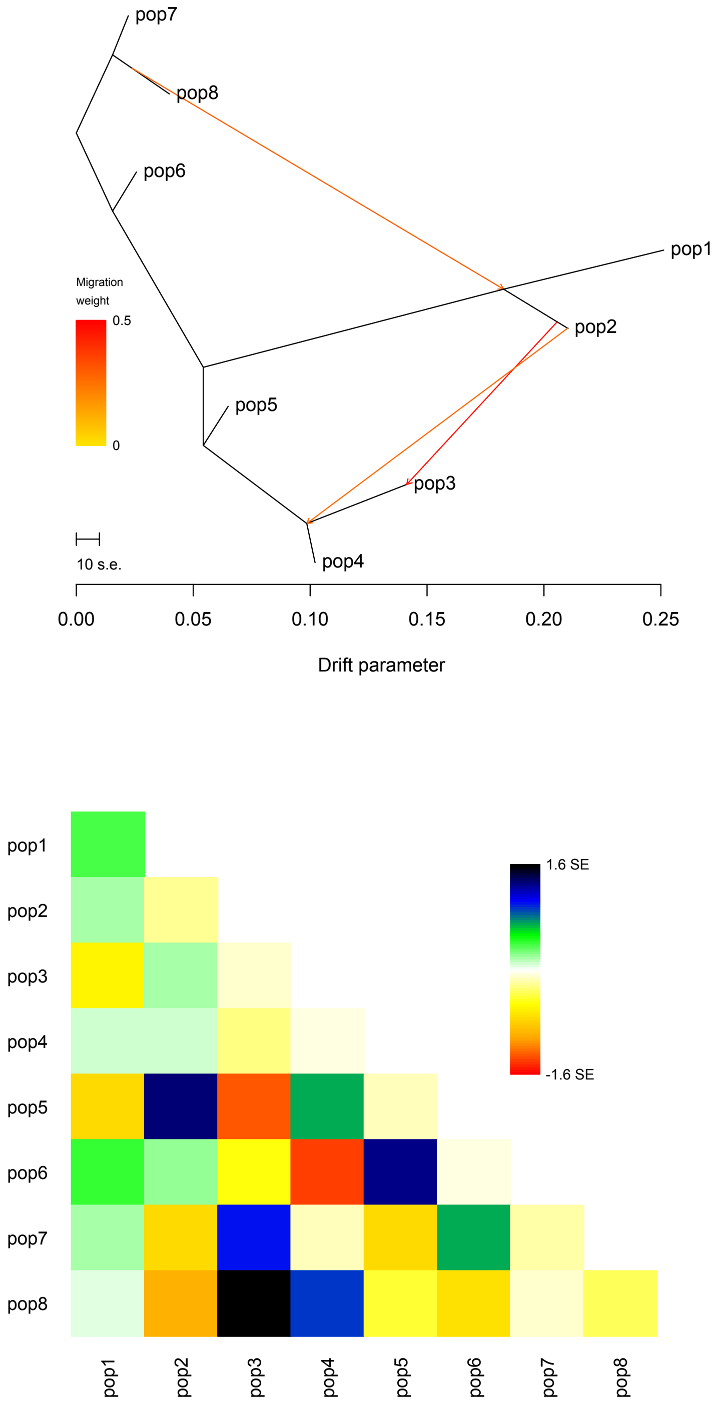
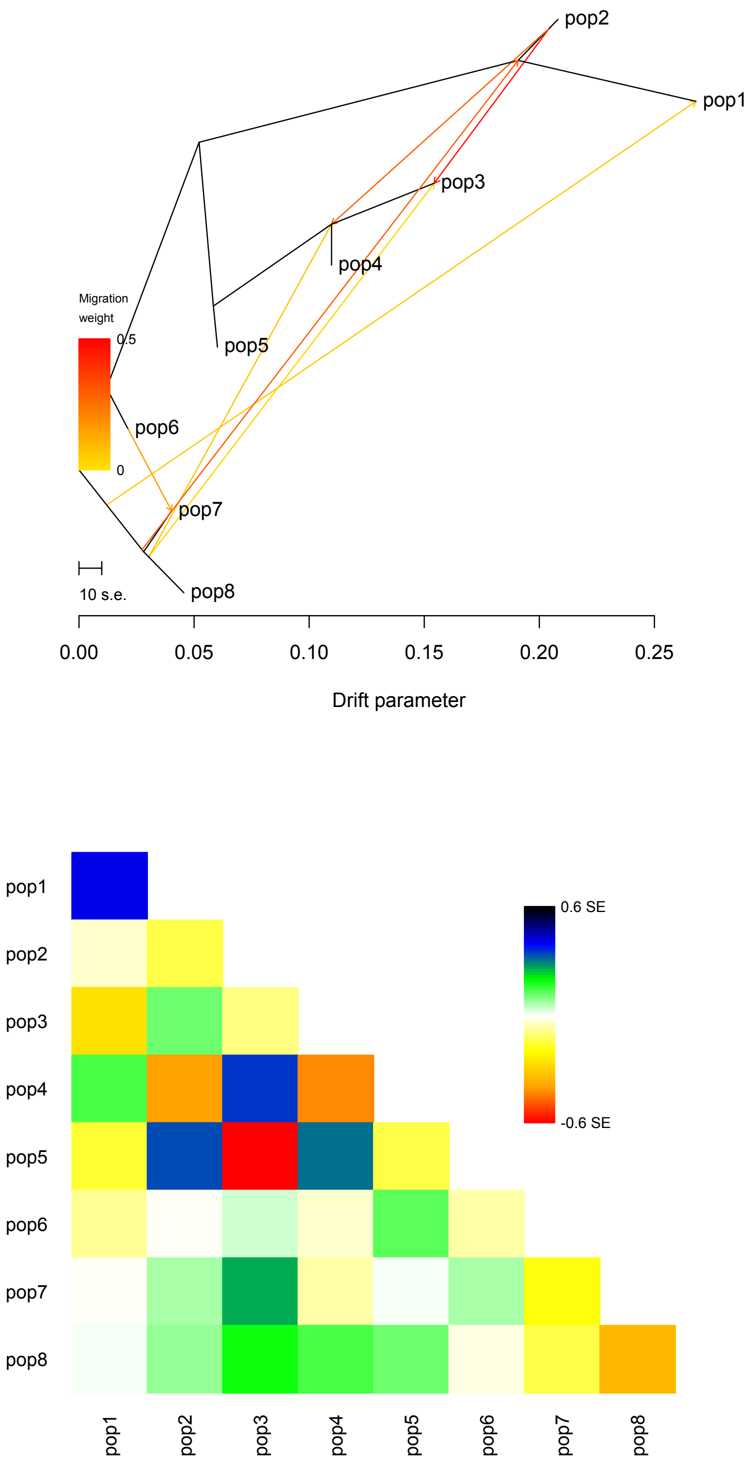


Figure 7. (Top) Maximum likelihood tree with seven migration events for eight-population scenario. (Bottom) Residual fit from the maximum likelihood tree.



References

Beerli P., Palczewski M. 2010. Unified framework to evaluate panmixia and migration direction among multiple sampling locations. Genetics, 185:313–26.

Gutenkunst R.N., Hernandez R.D., Williamson S.H., Bustamante C.D. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP data. PLoS Genetics 5:e1000695.

Hey J. 2010. Isolation with migration models for more than two populations. Molecular Biology and Evolution, 27:905–920.

Keinan A., Mullikin J.C., Patterson N., Reich D. 2007. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nature Genetics, 39(10):1251–5.

Leaché A.D., Harris R.B., Rannala B., Yang Z. 2014. The influence of gene flow on species tree estimation: a simulation study. Systematic Biology, 63:17–30.

Pickrell J.K., Pritchard, J.K. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics e1002967.

Pritchard J.K., Stephens M., Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics, 155:945–959.

Reich D., Thangaraj K., Patterson N., Price A. L., Singh L. 2009. Reconstructing Indian population history. Nature, 461(7263):489–94.