Extending *TreeMix* to microsatellite data

Joseph K. Pickrell[1,⋆], Jonathan K. Pritchard[2,3,⋆]

[1] Department of Genetics, Harvard Medical School, Boston, MA, USA

[2] Department of Human Genetics and

[3] Howard Hughes Medical Institute, University of Chicago, Chicago, IL, USA

⋆ E-mail: joseph_pickrell@hms.harvard.edu, pritch@uchicago.edu

October 1, 2012

We recently introduced the *TreeMix* model for inferring the set of population splits and mixtures in the history of a set of populations [Pickrell and Pritchard, 2012]. In the original model, the data were assumed to be biallelic SNPs, and the covariance matrix of allele frequencies is a sufficient statistic for the data. In this note, we extend the model to handle microsatellites.

Methods

Notation will be consistent with that in Pickrell and Pritchard [2012] where possible. Consider a single microsatellite locus with mean length l_A in an ancestral population. Length can be measured in terms of numbers of repeats or in terms of absolute length; the former is preferable. Now consider a population B. We model the mean length of the microsatellite in population B as:

$$L_B \sim N(l_A, c_B) \tag{1}$$

where c_B is a factor related to the amount of genetic drift that occurred between the ancestral population and population B. If we assume a mutation model where increments in microsatellite length are drawn from a distribution with mean 0 and variation σ_m^2, then $c_B \approx \mu t \sigma_m^2$, where t is the number of generations that have passed between the ancestral population and A, μ is the mutation rate at the locus [Slatkin, 1995]. In a stepwise mutational model, $\sigma_m^2 = 1$, and so $c_B \approx \mu t$ [Goldstein et al., 1995]. In reality, the distribution of mutation lengths is not symmetric, in that shorter microsatellites tend to increase in length and longer microsatellites tend to decrease in length [Sun et al., 2012]. Thus, the interpretation of the c parameters in terms of time should be taken as only a rough approximation.

We can now proceed with the same math as in the original *TreeMix* model, where instead of allele frequencies in populations we have mean microsatellite lengths. We define $\mathbf{V}$ as the variance-covariance matrix of mean allele lengths implied by a graph. The expected sample covariance matrix $\mathbf{W}$ is a simple transformation of $\mathbf{V}$, as in Pickrell and Pritchard [2012]. Assume that we have genotyped n microsatellites in each of m populations, and let the estimated mean microsatellite length at microsatellite k in population i be $\hat{X}_{ik}$. The observed sample covariance matrix $\hat{\mathbf{W}}$ is defined as:

$$\hat{\mathbf{W}}_{\mathbf{ij}} = \frac{\sum_{k=1}^{n}[(\hat{X}_{ik} - \hat{\mu}_k)(\hat{X}_{jk} - \hat{\mu}_k)]}{n} \tag{2}$$

where $\hat{\mu}_k = \frac{1}{m}\sum_{i=1}^{m} \hat{X}_{ik}$. This is a biased estimate of the covariance matrix because of finite sample sizes. In particular:

$$\hat{X}_{ik} \sim N(X_{ik}, \frac{\sigma_{ik}^2}{N_i}) \tag{3}$$

where N_i is the number of haplotypes sampled from population i and σ_{ik}^2 is the variance in allele lengths at microsatellite k in population i. In the SNP model, we use the bias correction shown in Eq. 4 in the Supplementary Material in Pickrell and Pritchard [2012]. For application to

microsatellites, we replace this with:

$$B_i = \frac{\bar{\sigma_i^2}}{\bar{N}_i} \tag{4}$$

where $\bar{\sigma_i^2}$ is the mean variance in microsatellite lengths in population i and $\bar{N}_i$ is the mean number of haplotypes (averaged across microsatellites) in population i. The bias-corrected covariance matrix is then Eq. 8 from the Supplementary Material in Pickrell and Pritchard [2012], with the bias (Eq. 7 in the same Supplement) calculated as above.

We then use the same *TreeMix* algorithm to search for tree/graph that maximizes the composite likelihood describing the fit of $\mathbf{W}$ to $\hat{\mathbf{W}}$.

Application to human data

We applied this model to a set of African populations reported in Tishkoff et al. [2009]. We restricted ourselves to the 848 microsatellites reported in this paper, and considered only the African populations and a handful of representative non-African populations. Only the absolute lengths of the microsatellites were reported (rather than the lengths in terms of numbers of repeats), and most microsatellites did not strictly follow a periodic size distribution (that is, even if the empirical distribution of sizes at each repeat seemed to show a periodic pattern, there were many individuals that fell at intermediate allele lengths). We thus used the absolute lengths in terms of basepairs as input to *TreeMix*, and treated each microsatellite as independent. The inferred population tree is shown in Figure 1, and the residual fit from this tree is shown in Figure 2. We then allowed *TreeMix* to add ten migration events to this tree. The resulting graph is shown in Figure 3. Note that there are large standard errors on the branch lengths, meaning the precise branching order of the populations in this tree is uncertain.

A number of known migration events are seen in Figure 3; these include Khoisan-related admixture in the !Xhosa, agriculturalist-related admixture in the Mbuti, Biaka, and Baka, and west Eurasian admixture in the Fulani. The extent of admixture in Africa (see e.g. Tishkoff et al. [2009]) means that we are likely missing large numbers of historically important events.

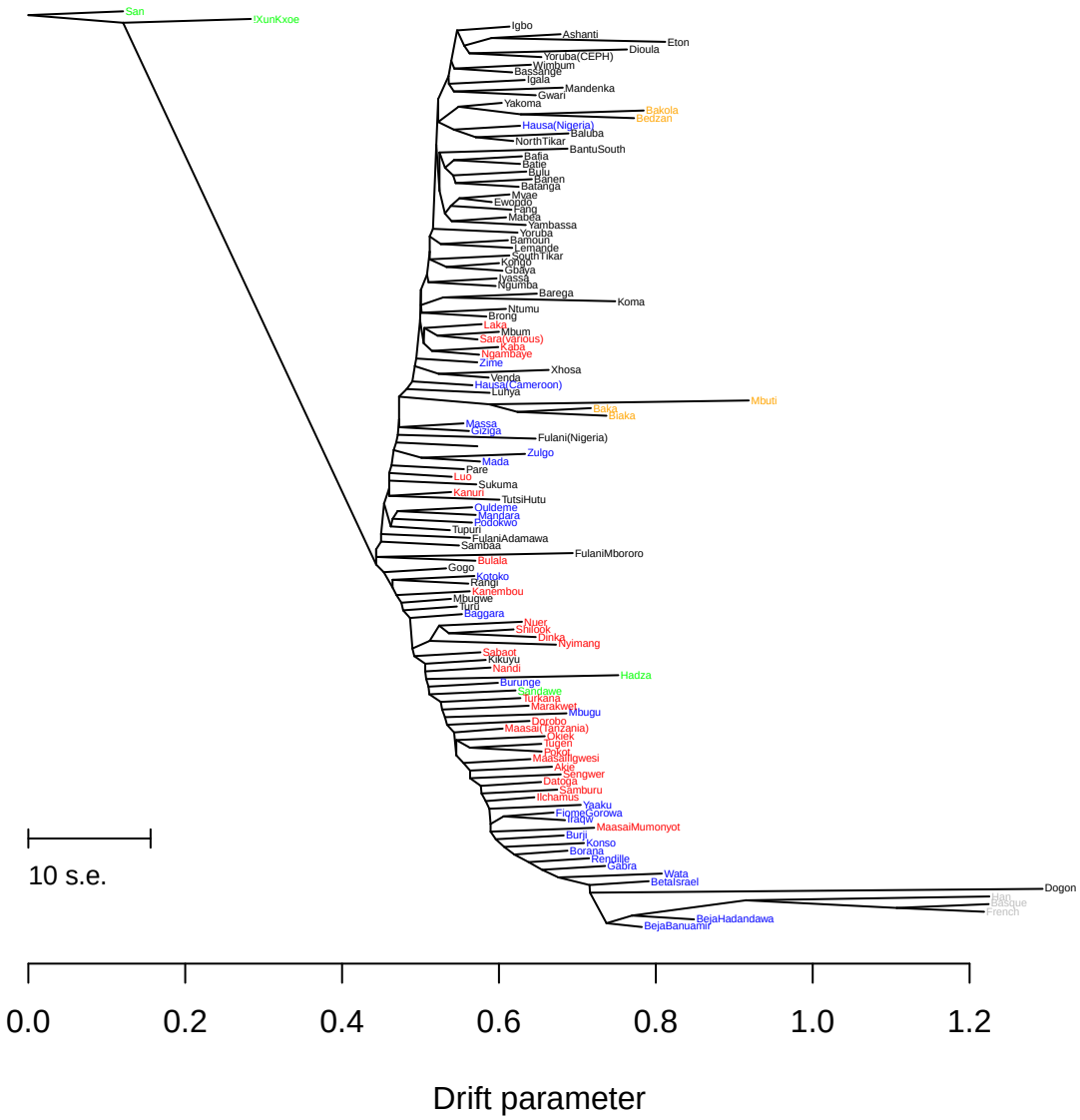

Figure 1**: Tree of African populations.** Populations are colored according to language/geographic groups. Green: Khoisan, black: Niger-Congo, orange: central African hunter-gatherers, green: Nilo-Saharan, blue: Afro-Asiatic, grey: non-African.

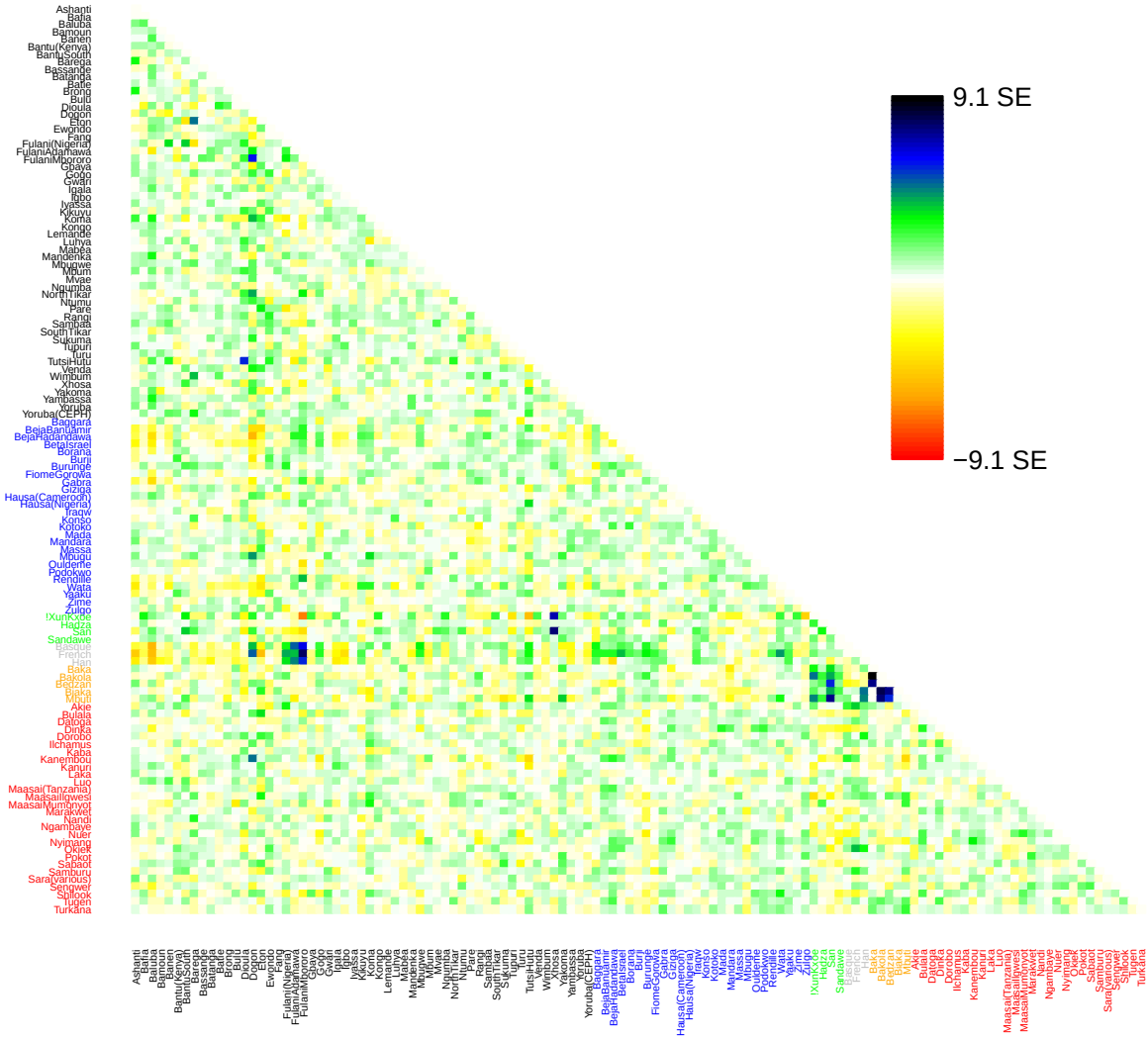

Figure 2: **Residuals from tree of African populations.** Shown are the residuals from the tree is Figure 1. Colors are the same as in Figure 1.

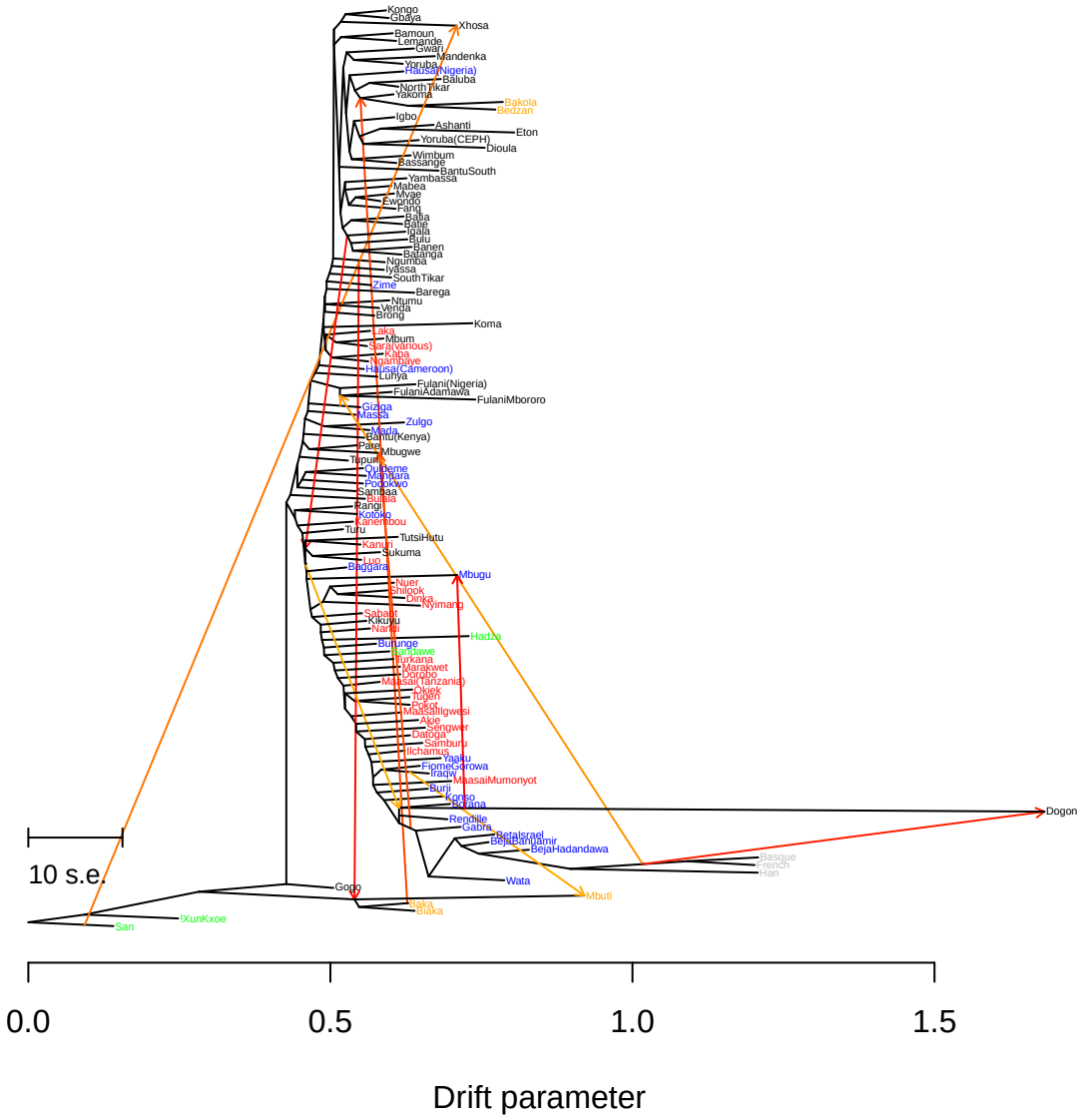

Figure 3**: Graph of African populations.** Shown in the graph of African populations allowing for ten admixture edges. Colors are as in Figure 1.

References

Goldstein, D., Linares, A., Cavalli-Sforza, L., and Feldman, M., 1995. An evaluation of genetic distances for use with microsatellite loci. *Genetics*, **139**(1):463–471.

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

Slatkin, M., 1995. A measure of population subdivision based on microsatellite allele frequencies. *Genetics*, **139**(1):457–62.

Sun, J. X., Helgason, A., Masson, G., Ebenesersdóttir, S. S., Li, H., Mallick, S., Gnerre, S., Patterson, N., Kong, A., Reich, D., *et al.*, 2012. A direct characterization of human mutation based on microsatellites. *Nat Genet*, .

Tishkoff, S. A., Reed, F. A., Friedlaender, F. R., Ehret, C., Ranciaro, A., Froment, A., Hirbo, J. B., Awomoyi, A. A., Bodo, J.-M., Doumbo, O., *et al.*, 2009. The genetic structure and history of Africans and African Americans. *Science*, **324**(5930):1035–44.