

ANALYZING THE GENETIC STRUCTURE OF POPULATIONS: INDIVIDUAL ASSIGNMENT

Introduction

In the last 5-6 years a different approach to the analysis of genetic structure has emerged: analysis of individual assignment. Although the implementation details get a little hairy, the basic idea is fairly simple. Suppose we have genetic data on a series of individuals. Label the data we have for each individual x_i . Suppose that all individuals belong to one of K populations and let the genotype frequencies in population k be represented by γ_k . Then the likelihood that individual i comes from population k is just

$$P(i|k) = \frac{P(x_i|\gamma_k)}{\sum_k P(x_i|\gamma_k)} .$$

So if we can specify prior probabilities for γ_k , we can use Bayesian methods to estimate the posterior probability that individual i belongs to population k , and we can associate that assignment with some measure of its reliability.

Applying assignment to the *Isotoma petraea* data

We can use **Structure** to do this analysis for us. For our first analysis we'll assume $K = 8$. Notice that K doesn't necessarily have to equal the number of populations we sample from, because some populations may not be distinct from one another. We'll come back to this point in just a moment. The results of the analysis are displayed in Figure 1. Each vertical bar corresponds to an individual in the sample, and the proportion of each bar that is of a particular color tells us the posterior probability that the individual belongs to the cluster with that color.

Figure 1 may not look terribly informative, but actually it is. It tells us that if we specify that there are eight populations from which individuals could be drawn, all individuals are equally likely to have been drawn from any of those eight populations. The problem here is that assignment procedures don't really work unless we have information from several

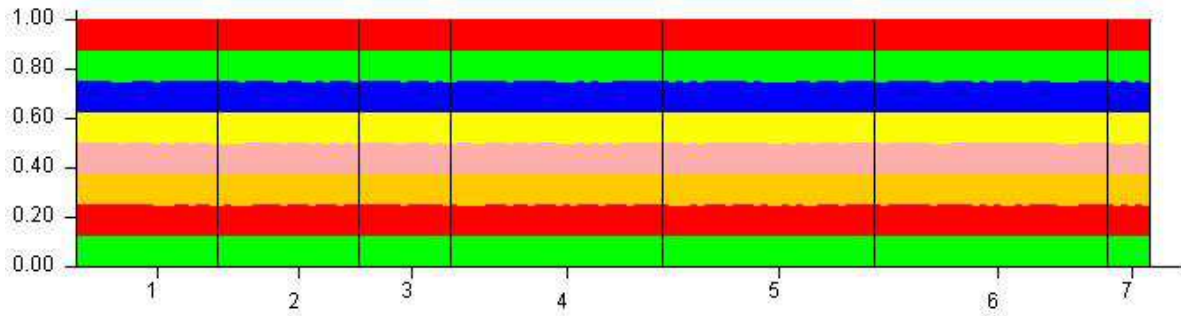


Figure 1: Analysis of *GOT* data from *Isotoma petraea*.

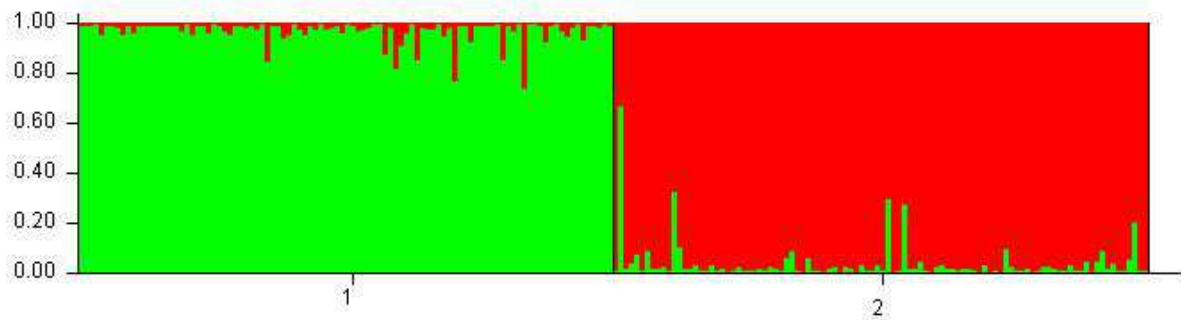


Figure 2: Initial analysis of sample data distributed with **Structure**.

independently inherited loci. Let's try the sample data distributed with **Structure**. The **PopID** column suggests that there are two different populations, so let's try $K = 2$ (Figure 2).

That looks a lot more informative. We can see that there are two clearly distinct populations, but there are also a few individuals that appear to be at least partially in the "wrong" population. As we'll discuss later, these individuals could be migrants. But all of this assumes that $K = 2$. What if there aren't really 2 populations here. Maybe there's only one or there are really three. How can we tell? Funny you should ask.

We can use **Structure** to try values of K from 1 to 3¹ When we do that we can compare the line called "Estimated Ln Prob of Data" in the printout. The K that has the least negative value (i.e., the number that has the smallest magnitude) for "Estimated Ln Prob of Data" is our best guess for K . In these data, the smallest value is -3969, which we get with $K = 2$. That means that Figure 2 provides the best description of the genetic structure for

¹Or any other for that number

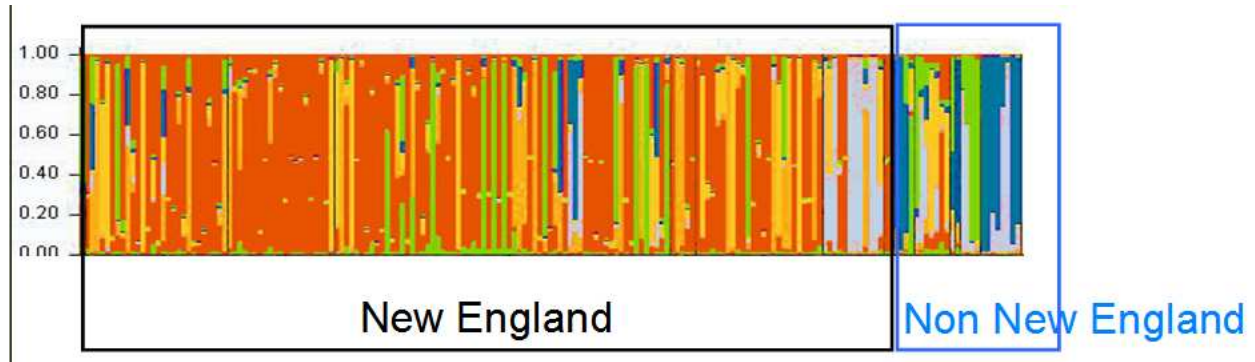


Figure 3: Structure analysis of AFLP variation in *Desmodium cuspidatum* (Skogen unpublished).

these data.

Figure 3 provides a real example using Krissa Skogen's data on *Desmodium cuspidatum*. You can see that New England and non New England populations appear to be quite different from one another.

Creative Commons License

These notes are licensed under the Creative Commons Attribution-NonCommercial-ShareAlike License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-sa/2.5/> or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.