Analyzing the Genetic Structure of Populations: Individual Assignment

Introduction

Although $F$-statistics are widely used and very informative, they suffer from one fundamental limitation: We have to know what the populations are before we can estimate them. They are based on a conceptual model in which organisms occur in discrete populations, populations that are both (1) well mixed within themselves (so that we can regard our sample of individuals as a random sample from within each population) and (2) clearly distinct from others. What if we want to use the genetic data itself to help us figure out what the populations actually are? Can we do that?\footnote{Would I be asking this question if the answer were “No?”}

Almost 20 years ago a different approach to the analysis of genetic structure began to emerge: analysis of individual assignment. Although the implementation details get a little hairy,\footnote{OK, to be fair. They get very 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\footnote{Remember the peculiar thing I mentioned about population geneticists earlier? We like to imagine we know something even when we don’t. In this case, I’m imagining we know $K$ even though we don’t. If we knew $K$, we’d probably already know which individual belonged in which population. We’ll get to the problem of determining what $K$ is later.} and let the genotype frequencies in population $k$ be represented by $\gamma_k$. 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 $\gamma_k$, 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.\footnote{You can find details in [8]. If you think about that equation a bit, you can begin to see why the details get very hairy. First, we’re trying to get the data to tell us what the populations are, so we don’t even know...}
### Table 1: Mean log probability of the data for $K = 2, 3, 4, 5$ in the *Berberis thunbergii* data (adapted from [3]).

<table>
<thead>
<tr>
<th>$K$</th>
<th>Mean L(K)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-2553.2</td>
</tr>
<tr>
<td>3</td>
<td><strong>-2331.9</strong></td>
</tr>
<tr>
<td>4</td>
<td>-2402.9</td>
</tr>
<tr>
<td>5</td>
<td>-2476.3</td>
</tr>
</tbody>
</table>

### Applying assignment to understand invasions

We’ll use *Structure* to assess whether cultivated genotypes of *Berberis thunbergii* contribute to ongoing invasions in Connecticut and Massachusetts [3]. The first problem is to determine what $K$ to use, because $K$ doesn’t necessarily have to equal the number of populations we sample from. Some populations may not be distinct from one another. There are a couple of ways to estimate $K$. The most straightforward is to run the analysis for a range of plausible values, repeat it 10-20 times for each value, calculate the mean “log probability of the data” for each value of $K$, and pick the value of $K$ that is the biggest, i.e., the least negative (Table 1). For the barberry data, $K = 3$ is the obvious choice.

Having determined that the data support $K = 3$, 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. Look at the labels beneath the figure. You’ll see that with the exception of individual 17 from Beaver Brook Park, all the of the individuals that are solid blue are members of the cultivated *Berberis thunbergii* var. *atropurpurea*. The solid red bar corresponds to *Berberis thunbergii* 'Atropurpurea', a different modern cultivar. You’ll notice that individuals 1, 2, 18, and 19 from Beaver Brook Park and individual 1 from Bluff Point State Park fall into the same genotypic cluster as this cultivar. *Berberis × ottawensis* is a hybrid cultivar whose parents are *Berberis thunbergii* and

---

5As part of her dissertation, Nora Mitchell used *Structure* to study a hybrid zone between two species of *Protea* [5]. Nora was interested in determining the extent to which individuals reflected ancestry from one of the two species involved. She set $K = 2$ to separate individuals as cleanly into two categories as possible and used the individual assignment score as an index of hybridity. There wasn’t any reason to attempt to infer $K$ from the data.
Berberis vulgaris, so it makes sense that individuals of this cultivar would be half blue and half red. The solid green bars are feral individuals from long-established populations. Notice that the cultivars are distinct from all but a few of the individuals in the long-established feral populations, suggesting that contemporary cultivars are doing relatively little to maintain the invasion in areas where it is already established.

Genetic diversity in human populations

A much more interesting application of Structure appeared a shortly after Structure was introduced. The Human Genome Diversity Cell Line Panel (HGDP-CEPH) consisted at the time of data from 1056 individuals in 52 geographic populations. Each individual was genotyped at 377 autosomal loci. If those populations are grouped into 5 broad geographical regions (Africa, Europe, the Middle East, and Central/South Asia, East Asia, Oceania, and the Americas), we find that about 93% of genetic variation is found within local populations and only about 4% is a result of allele frequency differences between regions [9]. You might wonder why Europe, the Middle East, and Central/South Asia were grouped together for that analysis. The reason becomes clearer when you look at a Structure analysis of the data (Figure 2).
Figure 2: **Structure** analysis of microsatellite diversity in the Human Genome Diversity Cell Line Panel (from [9]).

**A non-Bayesian look at individual-based analysis of genetic structure**

*Structure* has a lot of nice features, but you’ll discover a couple of things about it if you begin to use it seriously: (1) It often isn’t obvious what the “right” $K$ is.\(^6\) (2) It requires a *lot* of computational resources, especially with datasets that include a few thousand SNPs, as is becoming increasingly common. An alternative is to use principal component analysis directly on genotypes. There are technical details associated with estimating the principal components and interpreting them that we won’t discuss,\(^7\), but the results can be pretty striking. Figure 3 shows the results of a PCA on data derived from 3192 Europeans at 500,568 SNP loci. The correspondence between the position of individuals in PCA space and geographical space is remarkable.

**Other approaches**

Jombart et al. [2] describe a related method known as discriminant analysis of principal components. They also provide an R package, *dapc*, that implements the method. I prefer *Structure* because its approach to individual assignment is based directly on population

\(^6\)In fact, it’s not clear that there is such a thing as the “right” $K$. If you’re interested in hearing more about that, feel free to ask.

\(^7\)See [7] for details
Figure 3: Principal components analysis of genetic diversity in Europe corresponds with geography (from [6]). Panel b is a close-up view of the area around Switzerland (CH).
genetic principles, and as of a couple of years ago, I don’t have to worry so much about how long it takes to run an analysis on large datasets. A few months ago Gopalan et al. [1] released teraStructure, which can analyze data sets consisting of 10,000 individuals scored at a million SNPs in less than 10 hours. Here are a couple more alternatives to consider that I haven’t investigated yet:

- sNMF estimates individual admixture coefficients. It is reportedly 10-30 faster than the likelihood based ADMIXTURE, which is itself faster than Structure. sNMF is part of the R package LEA.

- Meisner and Albrechtsen [4] present both a principal components method and an admixture method that accounts for sequencing errors inherent in low-coverage next generation DNA sequencing data.

References


**Creative Commons License**

These notes are licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/ or send a letter to Creative Commons, 559 Nathan Abbott Way, Stanford, California 94305, USA.