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.[1] 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?[2]

A little over 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,[3] the basic idea is fairly simple. I'll give an outline of the math in a moment, but let's do this in words first. Suppose we have genetic data on a series of individuals at several to many (or very many) loci. If two individuals are part of the same population, we expect them to be more similar to one another than they are to individuals in other populations. So if we "cluster" individuals that are "genetically similar" to one another, those clusters should correspond to populations. Rather than pre-defining the populations, we will have allowed the data to tell us what the populations are.[4] We haven't even required *a priori* that individuals be grouped according to their geographic proximity. Instead, we can examine the clusters we find and see if they make any sense geographically.

Now for an outline of the math. Label the data we have for each individual x_i. Suppose that all individuals belong to one of K populations[5] and let the genotype frequencies in pop-

[1] To be a little more precise (and more than a little pedantic), we have to *assume* that the sample locations we decide to treat as populations are discrete, well-mixed populations that are distinct from others.

[2] Would I be asking this question if the answer were "No?"

[3] OK, to be fair. They get *very* hairy.

[4] I'm playing fast and loose with words here. The data haven't actually told us what the *populations* are. They've told us what *clusters* are found in the data.

[5] Remember the peculiar thing I mentioned about population geneticists earlier? We like to imagine we

ulation k be represented by γ_k. Then the likelihood that individual i comes from population k is just

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

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.[6] Remember, though, that we've arrived at the assignment by *assuming* that there are K populations and that the genotype frequencies are in Hardy-Weinberg in all of those populations.[7] Since we don't know K, we have to find a way of estimating it. Different choices of K may lead to different patterns of individual assignment, which complicates our interpretation of the results.[8] We'll discuss both of these challenges in a simple, but real, data set to illustrate the principles.

Applying assignment to understand invasions

To see a simple example of how `Structure` can be used, we'll use it to assess whether cultivated genotypes of Japanese barberry, *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

know something even when we don't. In this case, I'm imagining we know that there are K populations 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.

[6] 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 how many populations there are. Then we have to find a way of estimating allele frequencies (and genotype frequencies) in populations when we don't even know which populations individuals in our sample belong in. Estimating the genotype frequencies is straightforward, because we assume the genotype frequencies at every locus are in Hardy-Weinberg. Think about that for a bit. It means we really shouldn't be using `Structure` if we think that populations are inbred.

[7] Did you read the last footnote?

[8] This is an example of the "no free lunch" principle. You don't get something for nothing. Here we gained the ability to have the data tell us what the populations are, but we made interpreting the results more difficult.

K	Mean L(K)
2	-2553.2
3	**-2331.9**
4	-2402.9
5	-2476.3

Table 1: Mean log probability of the data for $K = 2, 3, 4, 5$ in the *Berberis thunbergii* data (adapted from [3]).

barberry data, $K = 3$ is the obvious choice.[9]

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 vulgaris* 'Atropurpurea', a different modern cultivar.[10] 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* $\times$*ottawensis* is a hybrid cultivar whose parents are *Berberis thunbergii* and *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.

[9] As 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.

[10] I find it very confusing that *Berberis thunbergii* var. *atropurpurea* and *Berberis vulgaris* 'Atropurpurea' both have "atropurpurea" associated with their names, but that's the way life is.

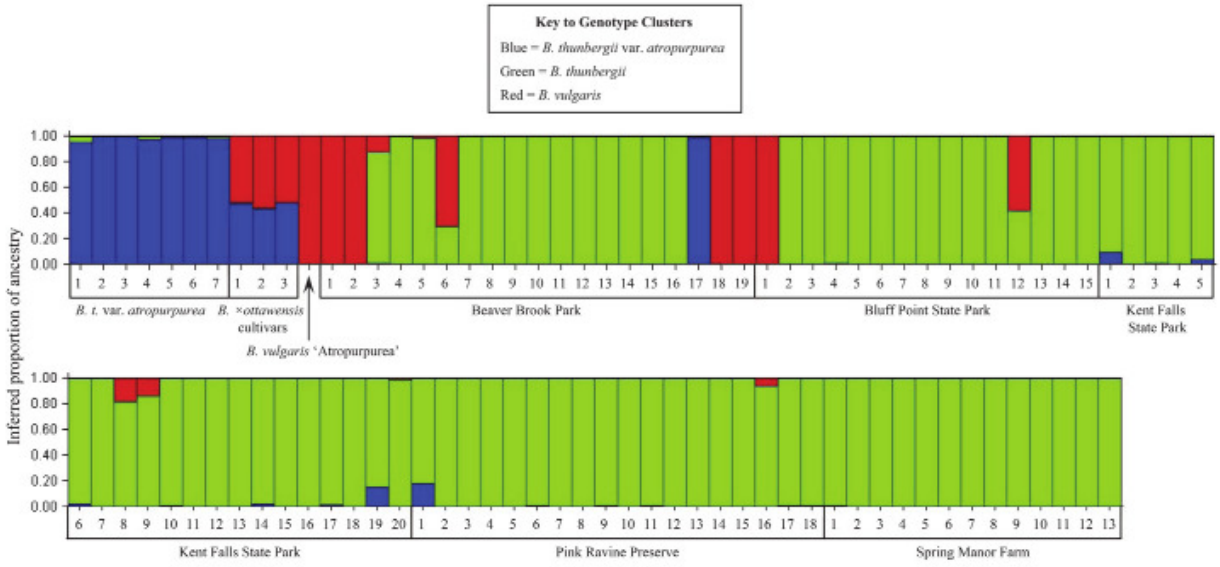

Figure 1: Analysis of AFLP data from *Berberis thunbergii* [3].

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).

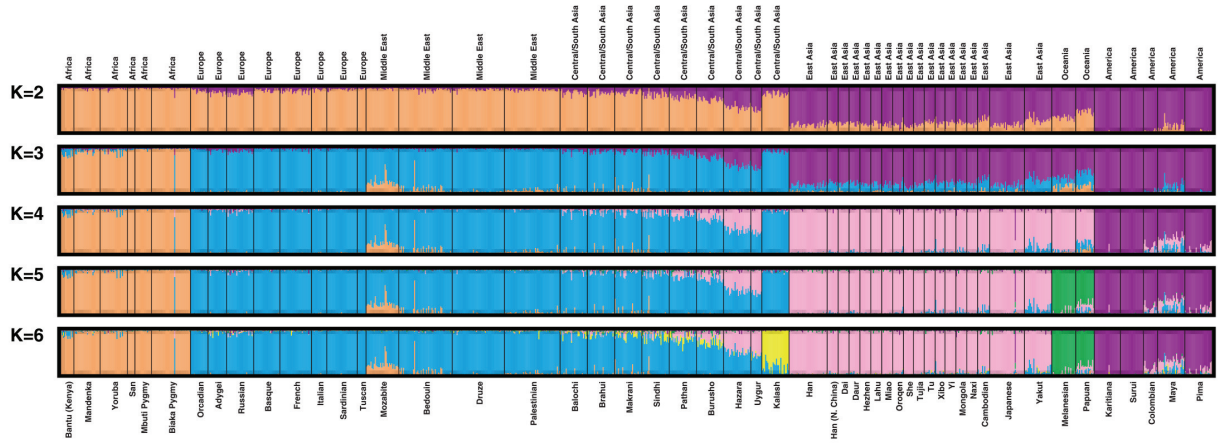

Figure 2: `Structure` analysis of microsatellite diversity in the Human Genome Diversity Cell Line Panel (from [9]).

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.[11] (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,[12], 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

[11] 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.

[12] See [7] for details

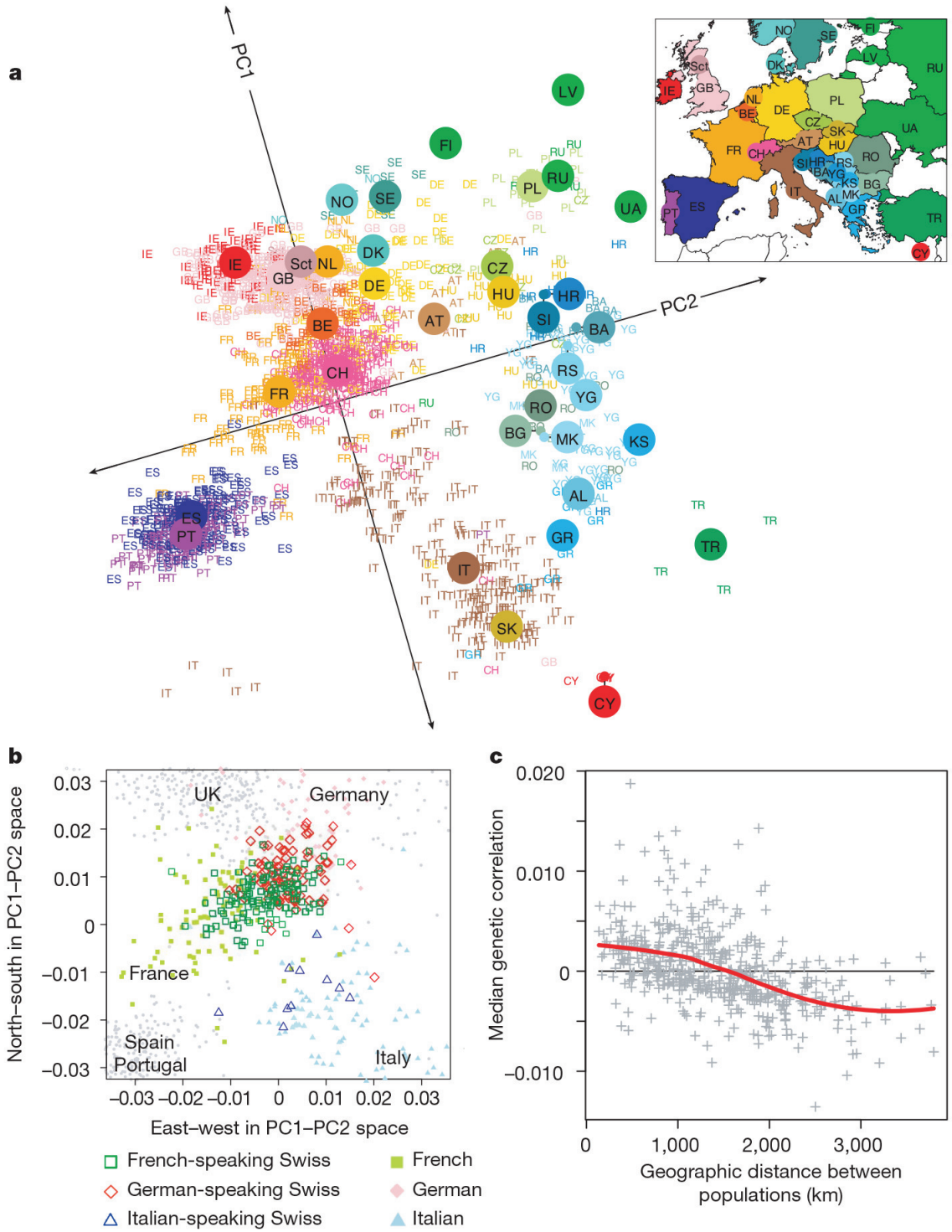

Figure 3: Principal components analysis of genetic diversity in Europe corresponds with geography (from [6]). Panel b is a close-up view of the area around Switzerland (CH).

genetic principles, and since computers are getting so fast (especially when you have a computational cluster available) that I worry less about how long it takes to run an analysis on large datasets.[13] That being said, Gopalan et al. [1] released `teraStructure` about five years ago, which can analyze data sets consisting of 10,000 individuals scored at a million SNPs in less than 10 hours. I haven't tried it myself, because I haven't had a large data set to try it on, but you should keep it in mind if you collect SNP data on a large number of loci. 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 Albrecthsen [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

[1] P. Gopalan, W. Hao, D. M. Blei, and J. D. Storey. Scaling probabilistic models of genetic variation to millions of humans. *Nat Genet*, 48(12):1587–1590, 2016.

[2] Thibaut Jombart, Sébastien Devillard, and François Balloux. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics*, 11(1):94, 2010.

[3] J D Lubell, M H Brand, J M Lehrer, and K E Holsinger. Detecting the influence of ornamental *Berberis thunbergii* var. *atropurpurea* in invasive populations of *Berberis thunbergii* (Berberidaceae) using AFLP. *American Journal of Botany*, 95(6):700–705, 2008.

[4] Jonas Meisner and Anders Albrechtsen. Inferring population structure and admixture proportions in low-depth ngs data. *Genetics*, 210(2):719–731, 2018.

[5] N. Mitchell and K. E. Holsinger. Cryptic natural hybridization between two species of protea. *South African Journal of Botany*, 118:306–314, 2018.

[13] I also remember that a very long time ago when systematists were complaining that likelihood analyses of their data sets were taking a couple of weeks, Joe Felsenstein was reported to have said "Why are you complaining that your analysis is taking a couple of weeks when you spent a couple of years collecting the data?"

[6] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, Matthew Stephens, and Carlos D Bustamante. Genes mirror geography within Europe. *Nature*, 456(7218):98–101, 2008.

[7] John Novembre and Matthew Stephens. Interpreting principal component analyses of spatial population genetic variation. *Nat Genet*, 40(5):646–649, 2008.

[8] Jonathan Pritchard, Matthew Stephens, and Peter Donnelly. Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, 155(2):945–959, 2000.

[9] Noah A Rosenberg, Jonathan K Pritchard, James L Weber, Howard M Cann, Kenneth K Kidd, Lev A Zhivotovsky, and Marcus W Feldman. Genetic structure of human populations. *Science*, 298(5602):2381–2385, 2002.

Creative Commons License