Population Genetics Project #2

Rachel Prunier and I collaborated with several colleagues on an analysis of single nucleotide polymorphisms in *Protea repens* the most widespread member of the genus.[1] The paper describing our work has just been accepted for publication in *American Journal of Botany.* You'll find a link to it from the lecture detail page.[2] Here's the basic story:

- We sample 663 individuals from 19 different populations of *Protea repens.* The populations are broadly representative of the geographical and environmental distribution that *P. repens* inhabits.
- We used genotyping by sequencing to identify 2006 polymorphic single nucleotide polymorphisms. We'll talk more about this later in the course, but for all you need to know is that SNPs are almost always bi-allelic, i.e., each locus typically only has two alternative alleles.
- We estimated F_{ST} at individual loci and used a relatively complicated Bayesian approach to identify "outlier" loci, those that have substantially larger estimates of F_{ST} than would be expected by chance given the background variation among loci. We identified 109 outliers in the paper.

I am providing a subset of the data we analyzed. It consists of a slightly larger subset of outliers than we identified in the paper. Specifically, it includes 64 loci that didn't quite meet the stringent criterion we used in the paper. It also excludes one individual we included in the paper.[3] So the data set you will analyze include 662 individuals from 19 populations with SNP genotypes at 173 loci.

Using these data answer the following questions:

[1] Rachel was one of my PhD students. She graduated in 2010.

[2] If I've set things up correctly, you will only be able to download a copy of the paper if you are within the UConn.Edu domain, either physically on campus or using the University's VPN. I haven't included all of the appendices referred to in the paper because I don't think you'll need them. If you think you do, let me know, and I'll upload them too.

[3] It happens that this individual was not scored at any of the 173 outlier loci included in this data set.

1. What is are the estimates of F_{IS} and F_{ST} you obtain from using Weir & Cockerham's approach to estimating F-statistics?

2. What are the estimates of F_{IS} and F_{ST} you obtain from using my Bayesian approach to estimating F-statistics? How different are these estimates from the Weir & Cockerham estimates?

3. Is there evidence for inbreeding within populations in *Protea repens*?

4. How similar to the genetic structure estimated from non-outlier loci is the estimated genetic structure derived from individual assignment in this data set? What do the similarities (or differences) you see suggest to you about the evolutionary processes affecting population differentiation at these loci. **Important note**: Send your `Structure` output as a ZIP file to Nora no later than midnight Thursday, 9 February. She will use `Structure harvester` to compile the results and return them to you by Friday night.

Hints

- You can estimate Weir & Cockerham statistics using `adegenet`.

- Population labels in Prunier et al. correspond to the population numbers in the `Structure` file as shown in Table **??**.

- If you have a Mac, you will probably need to fiddle with System Preferences to get `Structure` to run. When you click on the `Structure` icon, you may receive an error that says "Structure.app is damaged and can't be opened. You should move it to the Trash." Don't panic. `Structure` is fine. I just downloaded it on Saturday and tested it. You will, however, need to change a setting in your System Preferences. There are instructions at `https://www.tekrevue.com/tip/gatekeeper-macos-sierra/` on how to do it. Here's a quick summary:

  - If you are running Sierra, open a Terminal window and type the following command: `sudo spctl --master-disable`. You'll be asked for your password. It's the same password that you use when you log on to your Mac (if you have accounts enabled) or when you install new software. If you're running a version of macOS earlier than Sierra, you can skip this step.
  - Now open System Preferences, go to Security & Privacy and highlight the "General" tab. Click on the "Anywhere" radio button under "Allow apps downloaded

Population	Index
VanRhynsdorp	19
Cederberg	6
Banghoek	3
Riverlands	15
Ceres	7
Kleinmond	10
Montagu	13
Riviersonderend	16
Bredasdorp	5
DeHoop	14
Anysberg	2
Klein Swartberg	11
Garcia's Pass	8
Swartberg	17
Uniondale	18
Baviaanskloof	4
Kareedouw	9
Loerie Dam	12
Alicedale	1

Table 1: Population names and `Structure` indices for the *Protea repens* data set.

from:". You may need to click the lock in the lower left corner and enter your password.

Now before all of you Windoze users starting laughing about how convoluted this is, you should know that the reason for it is that Apple will only allow software downloaded from the App Store or from identified developers to run on macOS. That makes it far less likely that a malicious program can be installed and run on a Mac than on a Windoze machine.

Mac users: Given what I just told our Windoze friends, after you've finished your work with `Structure`, you may want to go back into System Preferences and change the radio button to "App Store" or "App Store and identified developers."

- If you specify `DIC=TRUE` when calling `analyze.data()` and use `print.summary()` to examine the posterior distribution, you will get posterior estimates for all of the allele frequencies in addition to estimates for f and θ. They take up a lot of screen space, so if you cut and paste your results in your report to Nora, leave these lines out.