**MCB 525, Fall 2016**

**2.5: Analysis of population structure in STRUCTURE**

Using a combined file of filtered genotypes from 96 samples + 6 unknowns, we will analyze the population structure of these samples. STRUCTURE is one of the most commonly used software used to infer population structure, and many tutorials exist to guide parameter choice and data interpretation.[[1]](#footnote-1) For now, we will run the program using mainly default parameters.

**Objectives:**

* Run and interpret STRUCTURE analyses
* Identify population of origin of unknown samples based on genotype

**Protocol:**

1. We will first convert the combined and filtered genotypes to a format suitable for input to structure. This script will actually output 2 files, one for structure and one for R, which we will need tomorrow.

$ ../scripts/TabToStructuRe.pl MCB525\_genos.tab genos4structure.txt genos4R.txt

1. The file is a bit unwieldy, but investigate the structure of the structure file with head –n or less. The first row of the file lists names of each locus. After that there are two lines for each sample, (since *Aiptasia* is diploid). The first column is the sample name, the second is the population ID, and every column following that lists either a 0 (if the allele matches the reference), a 1 (for the alternate allele), or -9 (for missing data). For example:

Locus1 Locus2 Locus3

Anemone1 1 1 -9

Anemone1 0 1 -9

Anemone2 0 0 0

Anemone2 0 0 0

Anemone3 1 -9 1

Anemone3 0 -9 0

1. Download the genos4structure.txt file to your desktop using the SSH Filter Transfer Client
2. Open STRUCTURE on your desktop
3. Start a new project (File > New Project). Name the project and choose where to save it. Finally, choose genos4structure.txt as the input data file.
4. On step 2 of 4 of the Project Wizard, click “Show data format” to get some information about the file. You should see the file has a header line (1 line with X number of columns – X is the number of loci) and Y lines with X + 2 additional columns (Y/2 is the number of samples, since there are two lines for each diploid sample). Enter in the appropriate values for Number of Individuals[Y/2], Ploidy [2], Number of loci [X], and Missing data value [-9].
5. On step 3 of 4 of the Project Wizard, check only the top box for “Row of marker names”.
6. On step 4 of 4, check the top two boxes for “Individual ID for each individual” and “Putative population origin for each individual”.
7. Click finish and proceed. You should see the data come up in a nice table.
8. Now to create a new parameter set to run the MCMC simulation (Parameter Set 🡪 New).
9. There are tabs to set the Run Length, Ancestry Model, Allele Frequency Model, and Advanced options. Set Burn in at 4000 and MCMC reps after burn-in to 10000.
10. We will use the admixture model with correlated frequencies, etc. (the default options).
11. Click OK and name the Parameter Set.
12. To run the simulation, (Project > Start a Job). Highlight your parameter set and enter 2 to 4 for K, and 1 iteration. Then click Start.

K is a parameter that describes the number of ancestral populations to assume.

1. Now you should be up and running! When the run is finished, open the results folder under the folder for your parameter set in the left panel.
2. To get a bar plot of each individual’s admixture proportion, Q, highlight the results output of a parameter run, and in the middle panel, choose (Barplot > Show). There are several options for viewing the barplot and other graphs that you can explore.

In these plots, each bar represents one individual, and each color represents one of K populations. The proportion of the genome originating from each inferred population is shown in distinct colors.

For a publishable analysis, we would need to compare many different simulations with variable parameters and longer MCMC repetitions. But for now, identify the bar that corresponds to your ‘unknown’ sample.

~~~

1. *From which population(s) is your unknown sample genetically similar?*
2. *What are its admixture proportions?*

1. Porras-Hurtado, L., Ruiz, Y., Santos, C., Phillips, C., Carracedo, Á., & Lareu, M. V. (2013). An overview of *STRUCTURE*: applications, parameter settings, and supporting software. *Frontiers in Genetics*, *4*, 98. http://doi.org/10.3389/fgene.2013.00098 [↑](#footnote-ref-1)