STRUCTURE EXERCISE

Table of contents

Background and aim

We all know, approximately, who African-Americans are. Although numbers are controversial, a large number of black-skinned people were brought as slaves from the west coast of Africa to the US, most in the 17th and 18th centuries. Slavery was abolished during the 19th century but in the Southern US legally enforced segregation continued until the 1960s. Despite integration and a genetic contribution from American settlers of European origin, and to a lesser extent also Native Americans, the population has to a large extent kept its distinct identity to the present. In this practical we will use genetic data to investigate their ancestry, doing our analysis using the software STRUCTURE.

Get STRUCTURE running on your computer

STRUCTURE is available from http://pritchardlab.stanford.edu/structure.html. Click on 'Download Structure 2.3.4', and choose the package with a graphical front end for whatever operating system you are using. The STRUCTURE documentation is available for download or in online form, which may be useful if any problems crop up.
NOTE: STRUCTURE is also installed on our cloud system, but only the command line version is fully functional!

Using genetic data to distinguish people from different continents

First use STRUCTURE to do something easy, which is to distinguish Africans from white Americans based on their genes. Run STRUCTURE and run the wizard to create a new project (click 'New Project ...' in the 'File' menu). In the four panels of the project wizard, enter the following information:

Panel 1

Choose a convenient project name, directory, and AfAmdata1.txt as the data file. On the cloud, this file can be found in `~/wpsg_2016/activities/STRUCTURE/`.

Panel 2

Here you specify the size of the data matrix, as well as how missing data is coded in the input file.
Individuals: 200
Ploidy: 2
Number of loci: 247
Missing data: 0

Panel 3

In the next two panels, you specify the format of the input file. Here, the rows included in the input file are specified.
Row of marker names: yes
Row of recessive alleles: no
Map distances between loci: yes
Phase information: no
Data file stores data for individuals in a single line: yes

Panel 4

Finally, the columns contained in the input file are specified.
Individual ID for each individual: yes
Putative population origin for each individual: yes
USEPOPINFO selection flag: no
Sampling location information: no
Phenotype information: no
Other extra columns: no

After you click 'Finish' and 'Proceed', the program should ask if the data are phased. Click 'No'.

If everything is correct, the front end should load the data at this point.
The 247 microsatellite loci are listed in their order on each of the chromosomes. The first line gives the name of the locus, while the second indicates its distance in Centimorgans from the previous marker. Each new chromosome is marked by a map distance of -1. One Morgan represents a stretch of chromosome in which one genetic crossover event per human generation occurs on average. A Centimorgan is a hundredth of this in physical units; one Centimorgan typically corresponds to about one million base pairs of DNA.

For each of the 200 individuals the data is stored in a single line. The first entry is a label identifying the individual (note that some of the Nigerian individuals have identical labels; they are in fact unrelated members of the same family). The second entry is a label indicating which population they are from. Here 2 is white American and 3 is Nigerian. There then follows data from each of microsatellite loci. At each locus each individual has two alleles. These are listed one after another, with the shorter one first.

In order to run STRUCTURE, you'll first have to define a new parameter set by clicking the 'New Parameter Set' button. A window will open where in the first panel you'll have to specify the run duration for the MCMC chain. Sensible values for run length are a burnin of 500 iterations followed by a further 1000 MCMC iterations. These values are a lot shorter than would be used to get really accurate answers, but will be relatively quick to run. These numbers can be varied depending on patience and the speed of the machine.

When running STRUCTURE, there are many different options. For the purpose of this practical, the default parameters will be appropriate for most purposes. However, we will try varying the ancestry model. Three ancestry models are available (in the second panel of the window for parameter set specification):

1) No admixture model

This model assumes that each individual receives all of its ancestry from only one of *K* populations. The model output is then the probability that the individual comes from each population.

2) Admixture model

This model assumes that each individual receives a proportion of ancestry q_k from population *k*. The output is then the estimated proportion of ancestry from each of the *K* populations. This model assumes that each locus has the same probability of coming from the *K* populations, independent of its neighbours.

3) Linkage model

The linkage model is the same as the admixture model but ancestry comes in chunks, so that neighbouring loci are more likely to come from the same population.

First, choose the model with no admixture. Close the window for parameter set specification by clicking 'OK', and give the new parameter set a name. Run STRUCTURE by clicking 'Start a Job' in the 'Project' menu. Select the parameter set you just defined and test both *K*=2 populations and *K*=3 populations. After the run has completed, repeat the same steps for the admixture model (without using sampling locations as a prior). What happens? What do these results mean?

Now make a new project for the second dataset (AfAmdata2.txt). This dataset contains an additional 150 African Americans (thus make sure to specify 350 as the number of individuals), labelled as population 1, but is otherwise identical to the first. Perform the same runs as for the first dataset. What happens and what do these results mean? Which model is more appropriate for this dataset: the admixture model or the no admixture model?

Finally, we will investigate linkage. Because this analysis uses a hidden Markov model (HMM) rather than considering each locus independently, the linkage model is slower than the other two models. Specify the parameter set for the linkage model with default settings and only run it with *K*=2 (set both 'from' to 2 and 'to' to 2). The value of *r* gives an estimate of the average size of the chunks that are inherited as unbroken units from the two source populations (Europeans and Africans), here measured in Centimorgans.
How do the results compare with those under the admixture model?

What values are of *r* are obtained?

How big, on average are the chunks of chromosome inherited from the two source populations?

(Harder question) What do the distribution of ancestry proportions and the value of *r* tell us about the dates at which European genes entered the African-American gene pool?

(General question) What additional data would improve our knowledge of the ancestry of African Americans?

Running STRUCTURE on the command line

If for some reason the installation of the STRUCTURE version with graphical front end fails on your machine, or if you have time left and want to learn how to run STRUCTURE through the command line, please use SSH to log in to your instance (or open a terminal on your remote desktop) and continue with the exercise below.

If the command line version of STRUCTURE is properly installed (as it should be on your instance), you can simply start the program by typing

```
structure
```

If you do that, you'll see an error message saying that the program couldn't open a file called 'mainparams'. That's because the command line version of STRUCTURE expects by default that this file is present and contains all the settings that you would otherwise specify through the graphical front end. In addition to file 'mainparams', a second file called 'extraparams' is also read if present, but this second file is optional and only needed for some settings. Templates for the two files come with the STRUCTURE installation download, and can be found in `~/wpsg_2016/activities/STRUCTURE/`, together with the input files for the analysis of Afro-American genetic variation, `AfAmdata1.txt` and `AfAmdata2.txt`.

Either create a new analysis directory, or navigate to the existing activity directory for this exercise:

```
cd ~/wpsg_2016/activities/STRUCTURE/
```

If you type `ls`, you'll see that the 'mainparams', 'extraparams', and data files are present. If you start STRUCTURE from this directory…

```
structure
```

…you'll see that the software now reads the 'mainparams' and 'extraparams' files, but it complains that it is unable to open a file called 'infile'. To get STRUCTURE to run, you'll have to change the name of the input file specified in the file 'mainparams'. Try to find out which other settings you need to change in order to run the no admixture model, with *K*=2, as described above?

[toggle title_open="Hide Answer" title_closed="Show Answer" hide="yes" border="yes" style="default" excerpt_length="0" read_more_text="Read More" read_less_text="Read Less" include_excerpt_html="no"]In the file 'mainparams', the following changes are needed:

```
#define BURNIN 500
#define NUMREPS 1000
#define INFILE AfAmdata1.txt
#define NUMINDS 200
#define NUMLOCI 247
#define MISSING 0
#define ONEROWPERIND 1
#define MAPDISTANCES 1
```

In the file 'extraparams', two lines need to be changed:

```
#define NOADMIX 1
#define ADMBURNIN 250
```

Obviously, if you were to run the linkage model instead of the no admixture model, you would leave this setting unchanged…

```
#define NOADMIX 0
```

…and you would specify this line instead:

```
#define LINKAGE 1
```

[/toggle]

After changing these settings, start STRUCTURE again, which should then run the analysis just like the graphical front end in the analysis described above.

Survey

Please let us know how you got along with the exercise in this very short opinion poll.

JOIN OUR NEWSLETTER

Enter your e-mail address

SUBSCRIBE

KEEP UPDATED