1 Introduction

You're going to be using STRUCTURE and PCA to analyse a couple data sets.

2 Getting stuff installed

You can use any program you want (even write your own) to do these analyses, but I'll guide you through doing a PCA in R and using the base STRUCTURE GUI.

Things to install:

1. Download and install R if you don't have it.
2. Rstudio is a nice IDE for R.
3. Grab STRUCTURE here. (I'll be talking about the GUI version)

3 The input files

The course git repo has sample input files in the `structure/files` directory. You should open `tester.txt` in a text editor to look at the format. The first row lists the names of all the polymorphisms in our dataset. Then each following row lists data for an individual including it's name, sampling location, and then its genotype as a 0/1/2/-1. The genotype is the count of derived alleles where a -1 indicates missing data

This format is needed by STRUCTURE, you'll need to turn SLiM output into this format later.

4 PCA in R

You'll need to first install the `factoextra` package in R. In R studio you can use the package browser for this (usually a tab on the right side), or simply type the following into the console:

```
install.packages(factoextra)
```

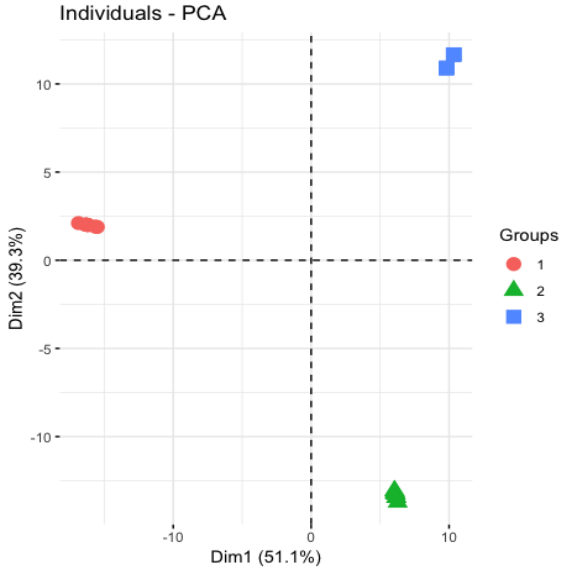
Next make a new R script and write the following code:

```
library(factoextra)
d = read.table("PATH_TO_YOUR_FILE", skip=1) #reads in our input file
vars=d[,3:ncol(d)] #pulls out the genotype columns
res.pca = prcomp(vars) #does the actual PCA
pair = c(1,2) #picks the PCA axes
#this pulls out the sampling locations:
sources = d[,2]
fviz_pca_ind(res.pca,
```

```
            habillage = sources, #if you uncomment this option then the
            #points will be labeled by sampling location
            label = "none",
            axes=pair, invisible = "quali",pointsize=4
)
```

You need to replace `PATH_TO_YOUR_FILE` with the path to the sample data set in the course git repo at `structure/files/tester.txt`.

The critical lines here are the `prcomp` command which actually does the PCA, and the final line which produces a plot like this:



You'll notice that this data set is not very exciting, there are 3 groups which correspond perfectly to their location label in the input file. If you comment out the `habillage` line (or just delete the parameter) the color will not be plotted. You'll need to do this for later datasets where the location column is not provided. (Note: you should also correct the vars row to start from column 2 not 3.)

5 Running Structure

This website has a few nice gifs showing the steps we are about to do. We are only doing steps 1.1-1.4 from that source.

5.1 Setting up the data

1. Open STRUCTURE.
2. Select File >New Project

3. Fill in the 3 fields, select the `texter.txt` file we used above as the data file.
4. Press next
5. Input the number of individuals in the file (this one has 30).
6. We are using diploid data, but STRUCTURE is silly, so put a 1 in the 'ploidy' field.
7. Input the number of loci (492 in this data).
8. Input '-1' as our missing data value.
9. Press next
10. Check the box next to 'Row of marker names'. This indicates to structure that the first line will be the list of the names of all the sites in our dataset.
11. Press next
12. Check the first two boxes: 'Individual ID for each indivdiual' and 'Putative population origin for each individual'. This tells STRUCTURE that the first two columns in data rows are a sample name an population number. (The population has to be an ID number, it cant be alphanumeric.) In future runs when you don't have location data you will not need to check this box.
13. Press finished
14. Press proceed

This will open the project view and should show you a grid will all of your data. The marker names will be highlighted in blue, and the sample names and population ID should be under columns called 'label' and 'Pop ID'.

5.2 Complete a run

1. From the top bar select Parameter Set >New
2. Put in a burn-in period of 3
3. Put in a number of MCMC Reps to 1000
4. Press OK
5. Name the parameter set 'p1' and press OK
6. From the top bar select Project >Start a Job
7. Select the parameter set 'p1' from the list

8. Set K from 2 to 5

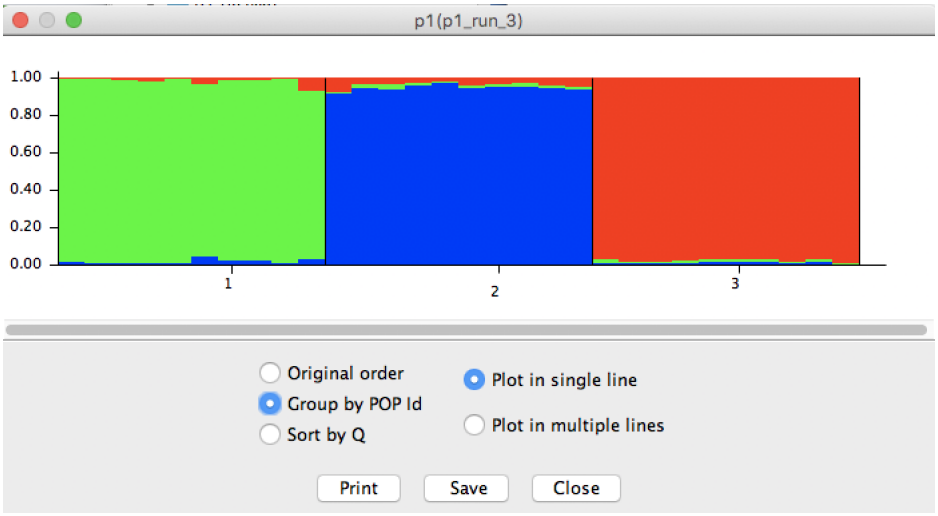
9. Press Start

This will potentially take several minutes to run, and STRUCTURE will notify you whenever each of the sets is completed.

5.3 Analysis

If you press View >'Simulation Summary' a table will pop up showing the results of all the runs, you can look at the `Ln P(D)` column (log of the Probability of the Data) to get a feel for which run best fit the data. In my run K=4 was the most likely (e.g. the most negative number), but we'd need to do some model fitting to actually tell which is best. Don't bet everything on these numbers on their own. Note also that because this is a probabilistic algorithm your exact numbers wont be the same if you re-run it.

If you select one of the runs listed in the left pane a new window will show up which summarizes data for that run. Go to 'Bar plot' >'Show' in this new window to show the STRUCTURE plot, you may want to press the 'Sort by Pop ID' or 'Sort by Q' to get a slightly different look at the data:



You can save an image of your plot in this window.

6 The assignment

You will need to analyze several data sets in these two methods below. You should produce at least one PCA and STRUCTURE plot for each one and write a brief analysis about each one. You will submit a PDF of your completed analysis to your personal github folder. Also include any associated scripts written for analysis.

Feel free to work together on these analyses, though each person should turn in their own report and their own simulation for 6.2 below.

6.1 Three populations with migration

Check out the `structure/simulations/three_pop_split_migration.slim` file. This short simulation creates 3 populations that split off from each other in sequence and then later come back into contact. The callbacks like:

```
p1.setMigrationRates(p2, 0.005)
```

turn on migration from one population to another (in this case 0.5% of the p2 population will be migrants from population p1 each generation). Note that these callbacks are unidirectional, if you want migration both ways you need to include two calls. You can check the SLiM manual for more info on this.

Analyse the data in `structure/files/three_pop_migration.txt` using the methods described above. Explain the results based on your understanding of PCA, STRUCTURE, and what is happening in the simulation.

6.2 Write your own simulation

You should write your own simulation to analyse data from. This can be arbitrarily complex but it should include at least 2 population splits and migration between at least two populations. You may want to look at some of the SLiM recipes for producing more complicated or interesting population histories.

Sample individuals from at least 3 of your populations and create output in the format that structure uses. You can convert the SLiM output directly into the structure format using Eidos or some external programming language of your choice.

Analyse the output from your simulation as above and discuss the results. If the results of STRUCTURE do not agree with your model you should discuss why that might be. Consider what assumptions of STRUCTURE your model may be breaking.

6.3 Some mysteries

You should now analyse the files names mystery1.txt, mystery2.txt, and mystery3.txt. Note that you should check the files before you start analysis, not all of them include the population label column, nor do they have the same sample

sizes. You may also need to tweak the settings for the run parameters we used above, you should do this based on the run results. You can use some of the other plots (e.g. likelihood, Fst) from the individual runs to see if you need to run for more than 1000 iterations, and based on the run summary you can decide if you should expand the search range of valid K's.

Questions you should try to answer about these simulations:

1. How many populations were there?
2. Were there differences in population sizes?
3. What patterns of migration were present?