Selection / linux workshop

~ Nick Harding

Talk on selection scans.

After the course you can download this from (here)[]

Analysis 1. Computing fixation index (Fst)

Fst is a measure of genetic differentiation between two populations. For this reason it can be a useful indicator of selection (as well as other things!).

As this is a scan, it is data, not hypothesis driven, and we compute Fst over a whole chromosome.

We use data from the ag1000 genomes project phase 2 data release. We use a VCF file, this is ok as we do not need phased haplotypes to compute Fst.

This analysis is carried out in windows, what might be an appropriate window size?

This analysis will be carried out on the lstm cluster.

NOTES: - The VCF file we are using has been downsampled to 10% to allow the analysis to be done in a short time. - The commands below can't be copied and pasted verbatim, you will need to change parts of them.

```
ssh yourusername@opteron.lstmed.ac.uk
```

Once you have logged into the cluster, we are going to create a directory in your home for this analysis. Good organisation of your analyses is *very* important.

```
cd # make sure you are in your home directory.
mkdir fst_analysis
cd fst_analysis/
mkdir output
```

Now we create some *symlinks*, these are shortcuts to the data we will be using. They are not strictly required, but make finding files much easier.

```
ln -s /GAARD/selection/samples/
ln -s /GAARD/selection/scripts
ln -s /home/elucas/galaxy_stuff/phase2.AR1
```

Now we have created 3 symlinks to:

- *samples*. This links to txt files describing which samples in the project are in which population.
- *scripts*. This links to scripts we will use to compute Fst.
- *phase2.AR1*. This links to the ag1000G data.

NOTE: These are just links to the data. Deleting your link will not delete the data itself.

This is a table describing the files in the `samples/` directory.

File	Population
phase2.ar1.AOcol.txt	Angola coluzzii
phase2.ar1.CIcol.txt	Cote d'Ivoire coluzzii
phase2.ar1.GAgam.txt	Gabon gambiae
phase2.ar1.GM.txt	The Gambia
phase2.ar1.GQgam.txt	Equatorial Guinea gambiae
phase2.ar1.UGgam.txt	Uganda gambiae
phase2.ar1.BFcol.txt	Burkina Faso coluzzii
phase2.ar1.CMgam.txt	Cameroon gambiae
phase2.ar1.GHcol.txt	Ghana coluzzii
phase2.ar1.GNcol.txt	Guinea coluzzii
phase2.ar1.GW.txt	Guinea Bissau
phase2.ar1.BFgam.txt	Burkina Faso gambiae
phase2.ar1.FRgam.txt	French Mayotte - gambiae
phase2.ar1.GHgam.txt	Ghana coluzzii
phase2.ar1.GNgam.txt	Guinea gambiae
phase2.ar1.KE.txt	Kenya (kilifi)

Task: Use the `wc` command to count how many samples are in each population. The `wc` command counts the numer of lines, words, and characters in a file.

Try something like:

```
wc samples/phase2.ar1.CMgam.txt
```

What does `wc -l` do?

Next we need to run the script. The Fst script is written in python, but we will call it using bash. Copy the template for the script, as follows.

```
cp scripts/run_fst.sh my_run_fst.sh
```

Open the script for editing using `nano`, and follow the instructions in the file.

```
nano my_run_fst.sh
```

When you have edited you file, you can run the script using:

```
bash my_run_fst.sh
```

Look at the file you have created in `output/`!

If you get this far, we can pull this file off the cluster for you to plot these data.