Analysis 2. Computing H12 in a populaton

H12 is a haplotype based scan for selection. For a window of a given number of SNPs, we identify each unique haplotype, and calculate the frequency of each. H12 is the sum of the first and second most common haplotypes, it is therefore bound between 0 and 1. The number of SNPs to use in the window is extremely important, it should be small enough so that we have good resolution, but large enough so that the probability of two haplotypes being Identical by state, but not Identical by descent is very small. For most *Anopheles* populations, we find 300 SNPs to be a good compromise.

As this is a scan, it is data, not hypothesis driven, and we compute H12 over a whole chromosome.

We use data from the ag1000 genomes project phase 2 AR1 data release. As we need phased haplotypes to compute H12, we cannot use the VCFs, we use data that has already been phased, it is in hdf5 format.

This analysis will also be carried out on the lstm cluster.

NOTES: - The commands below can't be copied and pasted verbatim, you will need to change parts of them. - hdf5 files are much easier to work with than VCF, which need to be parsed into numeric format.

```
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 h12_analysis
cd h12_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)

Next we need to run the script. The H12 script is written in python, but we will call it using bash. Copy the template for the script, as follows.

```
cp scripts/run_h12.sh my_run_h12.sh
```

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

```
nano my_run_h12.sh
```

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

```
bash my_run_h12.sh
```

Expect this to take about 5 minutes.

If it finishes successfully. Look at the file you have created in `output/`!

Extra tasks

- Open the `compute_h12.py` script, see if you can understand what is happening.
- Download the results file from the cluster

- Plot these data in R

~ nick harding