Detecting directional selection (selective sweeps)

```
mkdir ~/session_9
cd ~/session_9
```

Download files for this practical from http://wasabiapp.org/vbox/data/session_9

```
wget http://wasabiapp.org/vbox/data/session_9/finland.italy.fvb6.vcf.gz

wget http://wasabiapp.org/vbox/data/session_9/finland.vesca

wget http://wasabiapp.org/vbox/data/session_9/italy.vesca

wget http://wasabiapp.org/vbox/data/session_9/input_sweed_italy_fvb6.txt

wget http://wasabiapp.org/vbox/data/session_9/Plot_CLR_TajimaD_Fst.R
```

SweeD (Sweep Detector) is used to locate selective sweeps along the 6th chromosome in 17 Italian strawberry samples

Calculate Composite likelihood ratio (CLR) statistics using unfolded sites (ancestral allele predicted based the outgroup sequence, Japanese strawberry, Fragaria iinumae)

```
SweeD -name run_name -input input_sweed_italy_fvb6.txt -grid 3900 -noSeparator
```

Rename output file based on grid size

Test different grid sizes (for example, grid size 3900 means that you try to locate a sweep using 10 kb windows, given the length of the 6th chromosome is around 39 Mb), greater number of grids makes analysis more accurate but also slower.

We use vcftools to calculate different summary statistics along the 6th chromosome

Calculate Tajima's D in 10kb windows for Italian samples

```
vcftools --gzvcf finland.italy.fvb6.vcf.gz --keep italy.vesca --TajimaD 10000
```

Calculate Fst between Finnish and Italian strawberries at non-overlapping 10kb windows

```
vcftools --gzvcf finland.italy.fvb6.vcf.gz --weir-fst-pop italy.vesca --weir-fst-pop
finland.vesca --fst-window-size 10000
```

Try different window sizes (eg. 30kb, 100kb)

Name output files accordingly

Look results using R

This needs two R packages that are not included in the VM. They can be installed within R or using the apt-get commands:

```
sudo apt-get install r-cran-ggplot2
sudo apt-get install r-cran-gridextra
```

Start R Studio

Open the script (Plot_output_CLR_TajimaD_Fst.R)

Replace input files with the file names you gave for output files (Check your output file path)

Open the script

Draw CLR, Tajima's D and Fst along the chromosome 6

Save figures

Discussion

Where in the chromosome you find potential sites for recent directional selection? Mark those regions

Explain why

How extent the swept genomic regions are? What can you conclude based on that?

What other population genetic forces could explain observed patterns of variation?