Computational Genomic Approaches in Evolution Population Genetics

Talk 2: Genealogies under selection and a method for detecting selection

Outline

The aim of the talk is to give a brief introduction to selection and pcadapt.

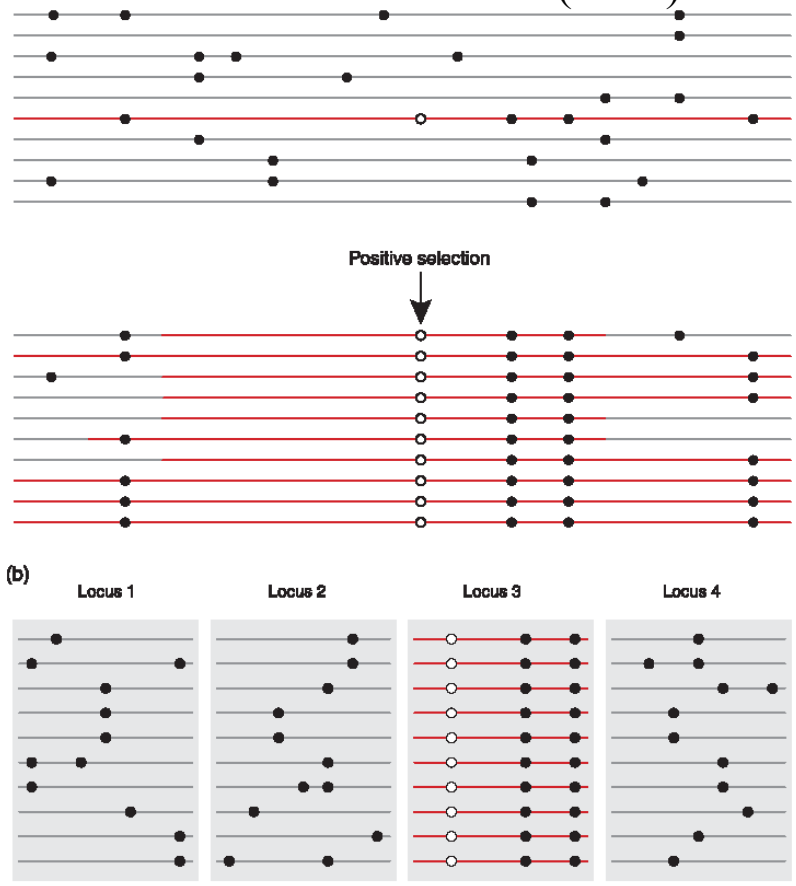
1. Population Structure
2. Selection
3. Principal component methods in population genetics
4. The pcadapt method

Directional selection and selective sweeps

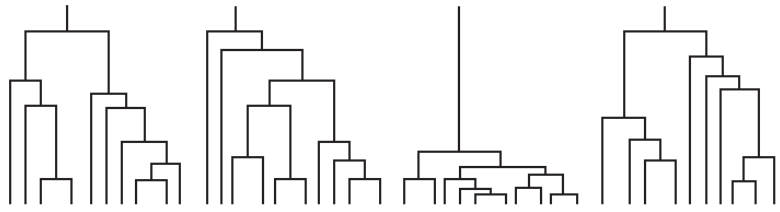
Modified from Storz (2005)

Newly advantageous mutation

‘Sweeps away variation’



In comparison with other loci there is a shorter gene genealogy and reduced variation.



Genealogies under selection

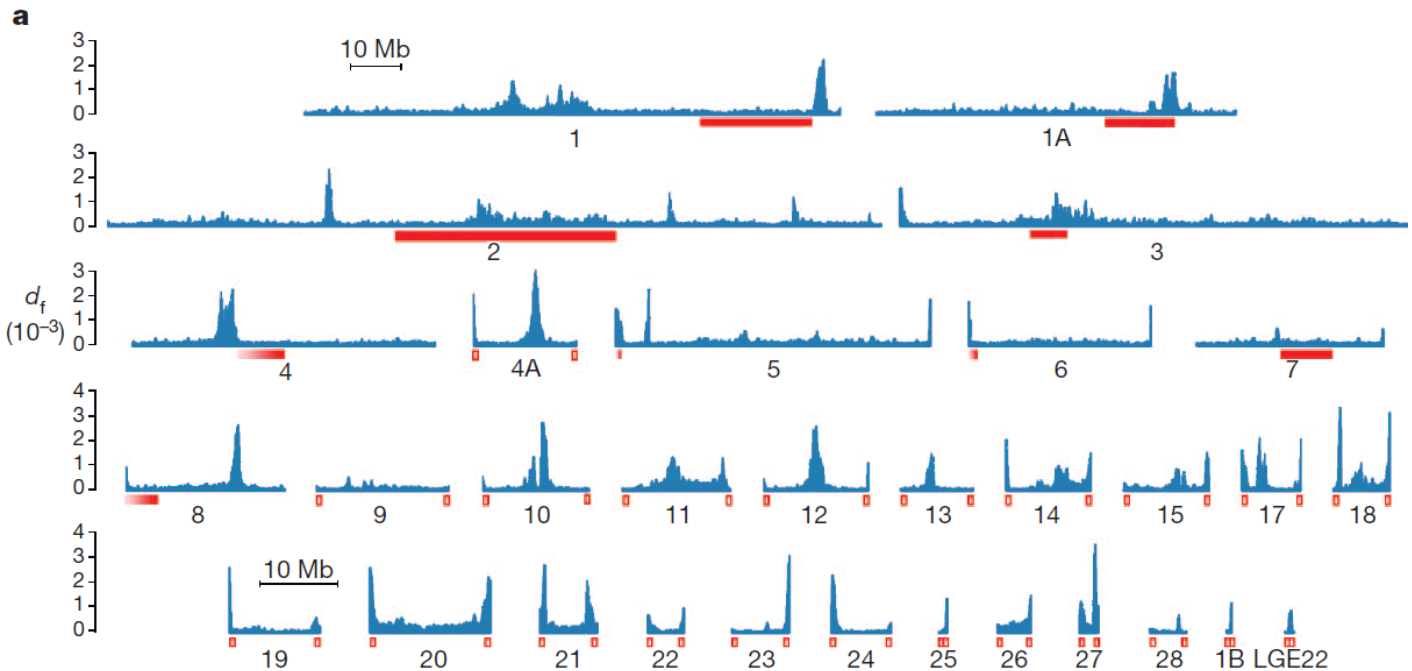
- When an advantageous mutation arises, it is likely to have more descendants.
- If it spreads quickly a large region of the genome around the mutation will not have enough time to recombine before coalescence (thinking backwards in time).
- This region will tend to have a limited number of genealogies and reduced genetic variation.
- In multiple populations, connected by migration, if different alleles are favoured in different populations (local selection) a 'chequer-board' pattern of gene frequencies may arise

Example of local selection

The pied flycatcher (*Ficedula hypoleuca*) has an extensive range over northern Europe, whereas the collared flycatcher (*F. albicollis*) is more southerly, although they overlap in central Europe.

The group of Han Ellegren has performed whole genome resequencing from a number of individuals of both species in order to both look at evidence of natural selection and also to uncover the demographic history. (Ellegren *et al*, *Nature* 2012)

This plot shows a measure of genetic divergence between the two species across the genome.



The pied flycatcher example

- We observe sharp peaks in the level of divergence — some regions of the genome are genetically very different.
- An explanation for this is that gene-flow (hybridisation) has ensured that most of the genomes of the two species are quite similar.
- However, there are some groups of genes where there is divergent selection, favouring different genetic variants in each species.

Detecting selection without defining populations

- Although it is possible to detect selection by examining levels of divergence between populations, sometimes it is difficult to define populations precisely.
- Genetic similarity between groups of populations (because of geography or shared history) can confound many methods for detecting selection because these assume the populations are independent.
- Recently there has been interest in detecting selection using clustering methods, without defining populations.

Principal component methods and population genetics

- DNA sequence data can typically be coded as '0', '1', or '2' depending on the number of copies of the non-reference allele an individual has.
- The reference allele is usually the commonest ('major allele'). *E.g.* if the reference allele is A and the non-reference ('minor allele') is G, then AA is 0, AG or GA is 1, GG is 2.
- So the data consists of a large matrix:

```
0 2 0 1 2 1 1 1 . . .
2 1 0 1 1 2 2 0 . . .
1 2 1 0 1 0 1 2 . . .
. . . . . . . . .
. . . . . . . .   .
. . . . . . . .     .
```

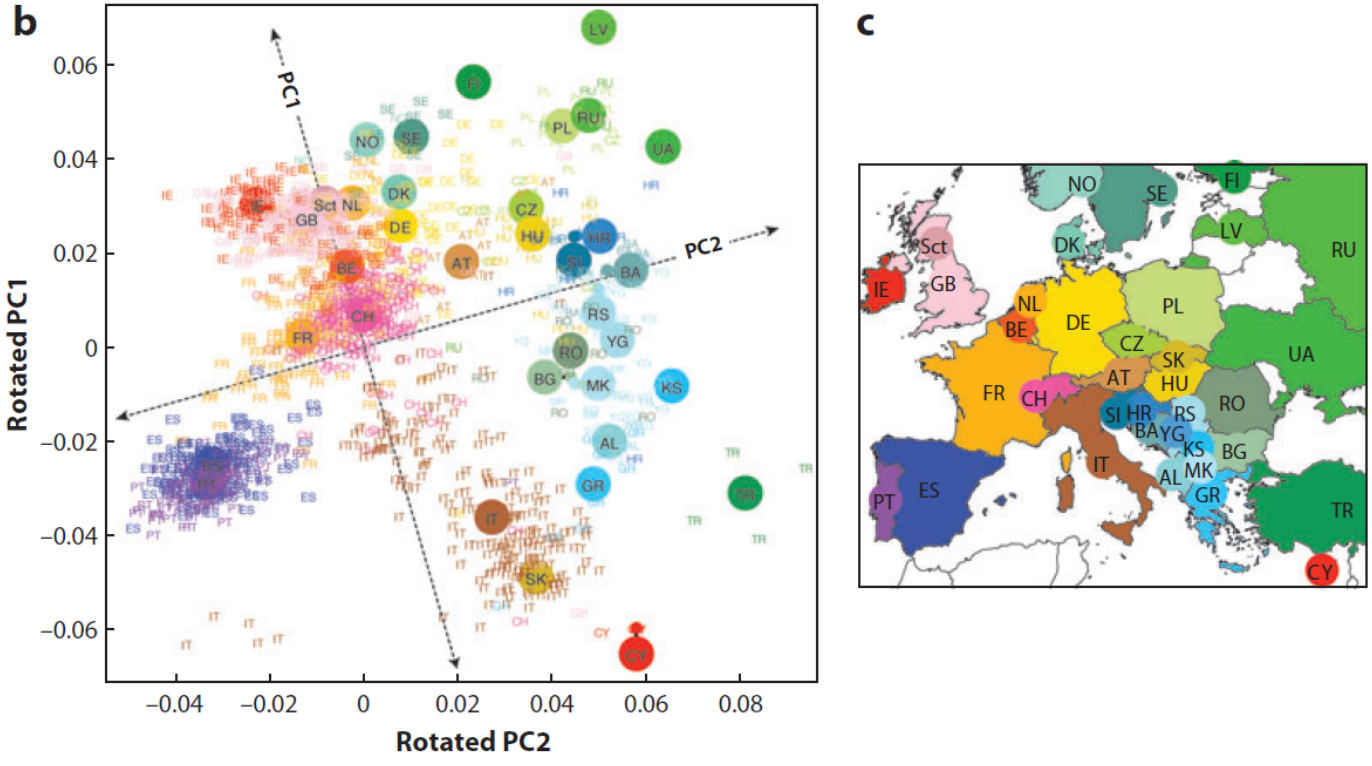
- The rows correspond to individuals, and the columns to SNPs
- A principal components analysis (PCA) aims to summarise the information in the matrix as compactly as possible.

PCA in more detail

- A principal components analysis gives vectors of loadings (length the same as the number of SNPs). One vector for each PC.
- Each vector of loadings can be multiplied by the vector of 0s, 1s, 2s for each individual to **project** the high-dimensional genotype to just one number for that individual for each PC.
- The principal components are uncorrelated
- Usually the 0s, 1s, 2s are themselves transformed beforehand to give each SNP the same variance, but the principle is the same.

This example comes from a study by Novembre *et al*, Nature 2008.

They carried out a PCA of SNPs of European humans, and showed that the positions of people on the first two principal components mirrored geography.



- Each loading value for a SNP represents the weighting of that SNP in the principal component score (values close to zero mean the SNP has little effect on the principal component score).
- There has been recent work on looking at the statistical distribution of loading values in order to pick out outliers which might be candidates for selection (E.g. Luu *et al.*, *Mol. Ecol. Res.*,2017, **17**, 67-77).
- One R package that uses this approach is *pcadapt*.

Plan of Project

- The instructions for carrying out the practical are on this page:
  `https://ritarasteiro.github.io/FieldCourse/pages/PCadapt`
- Follow the instructions in the "Practical pcadapt" link to get a tutorial on using pcadapt to analyse genome data.
- This will also include how to characterise genes on GenBank.
- The aim is to:
  - describe the population structure
  - characterise candidate loci that may be under selection
  - examine distribution of alleles among individuals
- The overall aim is to explain the relationships between populations and the role of natural selection.
- To achieve this will look at a data set taken from the Dryad repository. This is some wolf data, published by Schweizer et al (2016).

Schweizer, R.M., Vonholdt, B.M., Harrigan, R., Knowles, J.C., Musiani, M., Coltman, D., Novembre, J. and Wayne, R.K., 2016. Genetic subdivision and candidate genes under selection in North American grey wolves. Molecular Ecology, 25(1), pp.380–402.