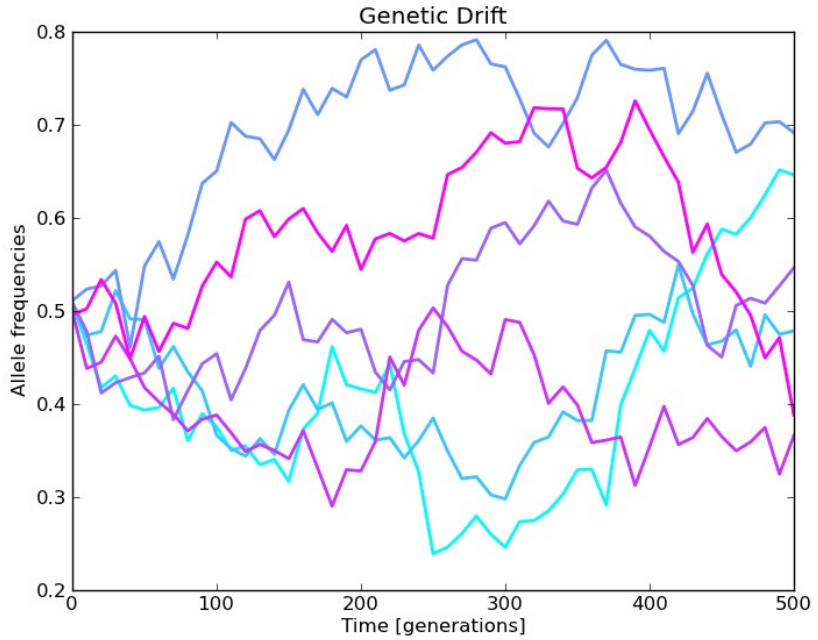
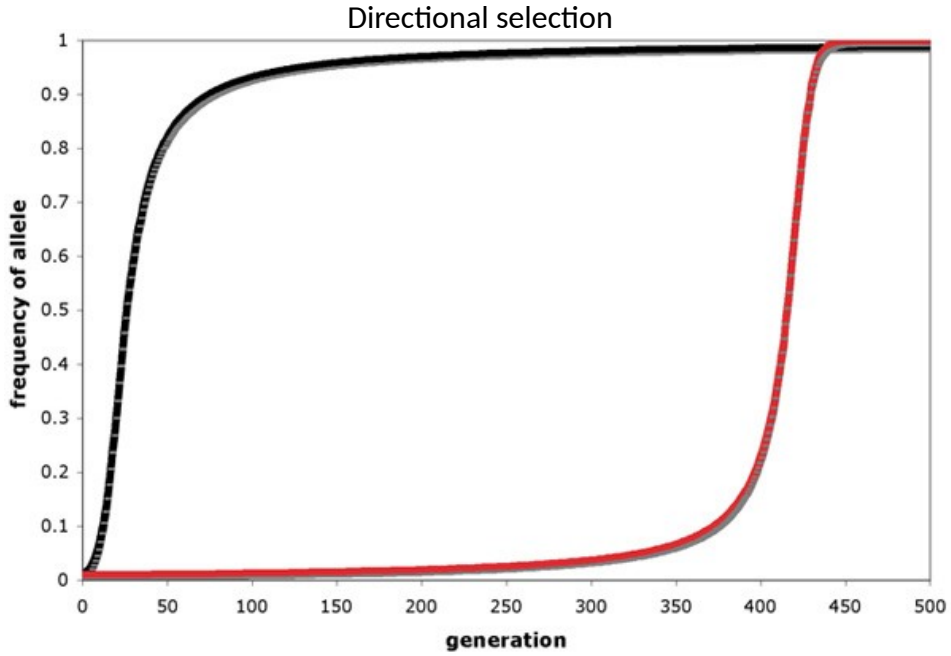
Detecting directional selection (selective sweeps)

Tuomas Toivainen / Evolutionary genomics course /7.4.2016

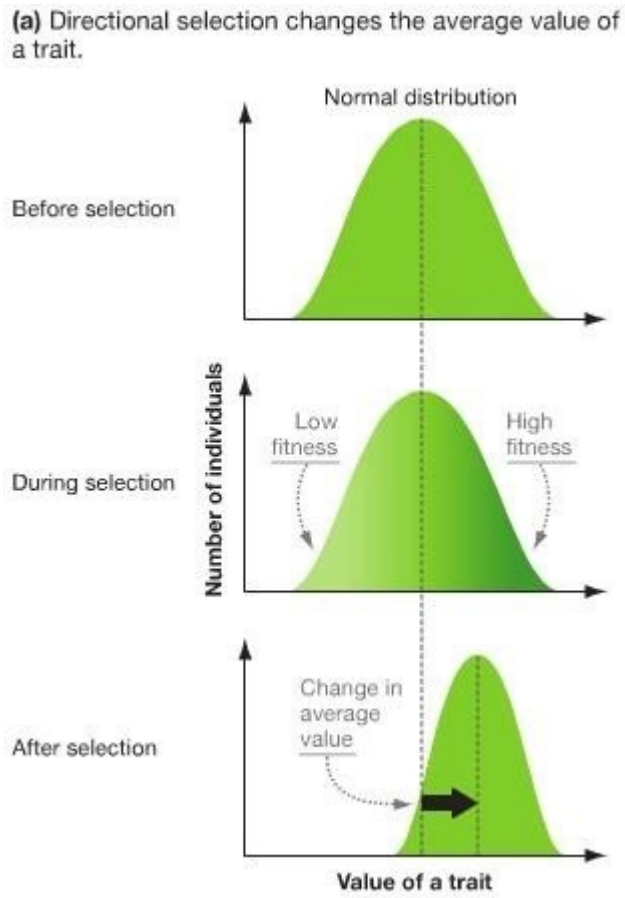
Evolution

- The change in allele frequencies (within species)
- Interplay of random genetic drift and selection
- Selection effective when 2Nes > 1
- Where s is selection coefficient and Ne effective population size
- In small populations drift can dominate selection
- Mutations produce new variation





Directional selection targets phenotypes



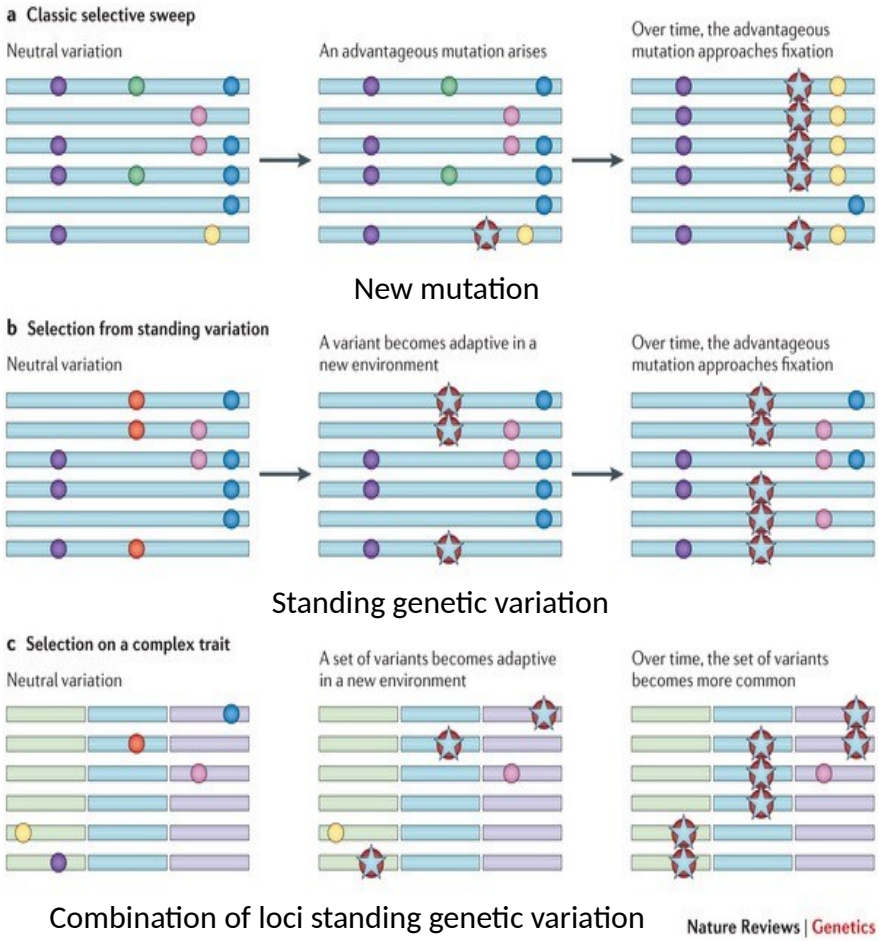
- New selection pressure due to eg. environmental change
- Depends on genetic architecture of a trait, how population respond to selection?
- Few large effect loci or tens of loci

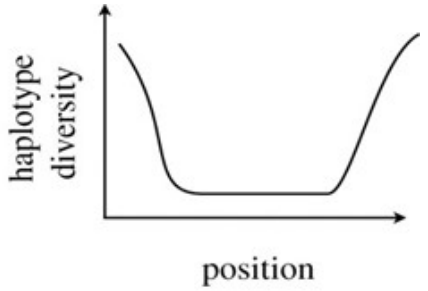
What kind of genetic variation is used for adaptation?

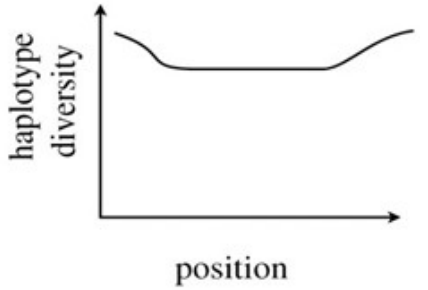
Hard sweep

Soft sweep

Polygenic



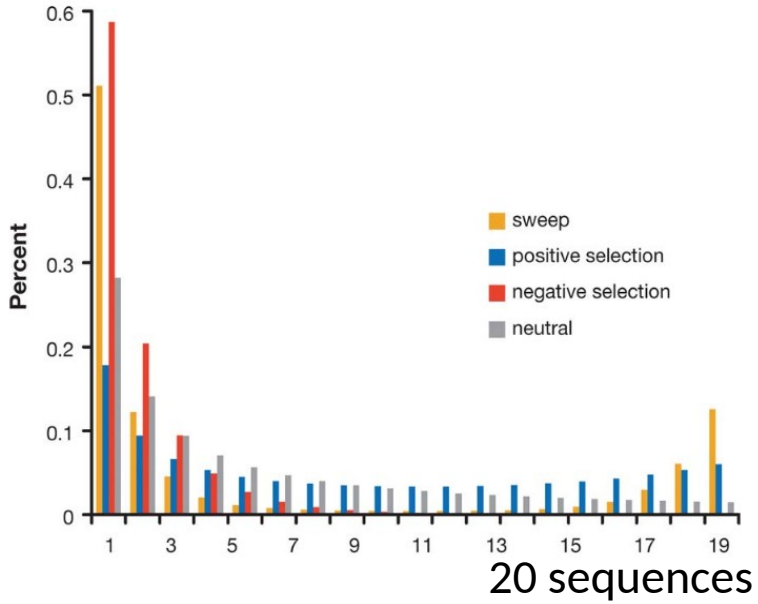




Scheinfeldt and Tishkoff 2013

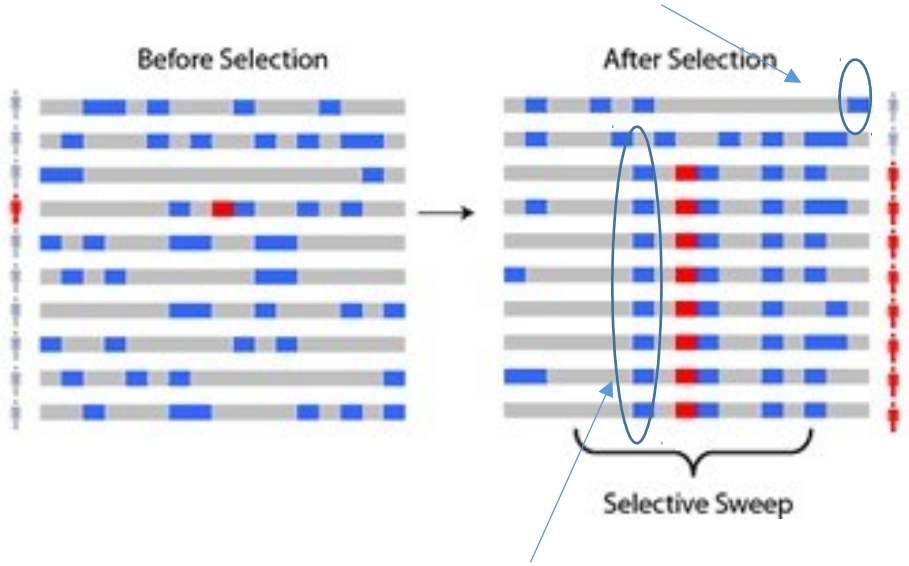
- Large effect mutations result in “hard” selective sweeps
- In polygenic adaptation allele frequencies change only slightly (even in thousands of loci)

Detecting a hard sweep using site frequency spectrum



Implemented in SweeD software (Pavlidis et al. 2013)

Singleton (1/10)



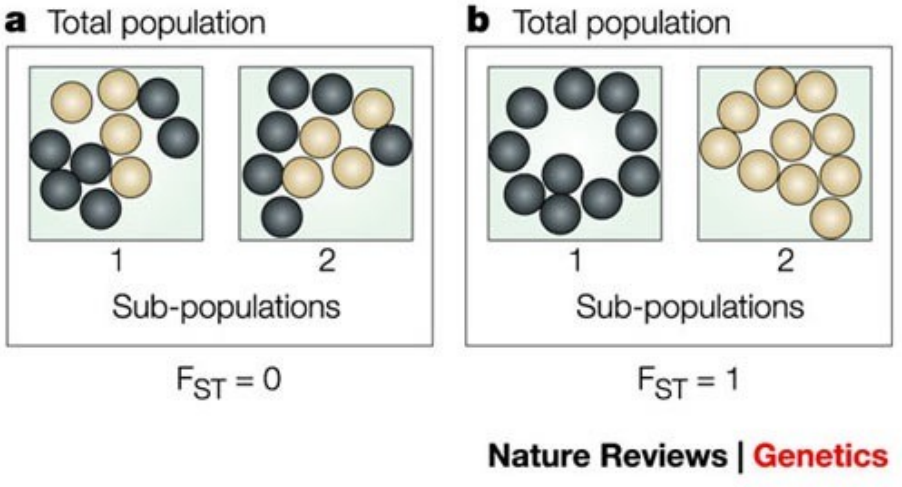
High frequency derived allele (8/10)

Other methods

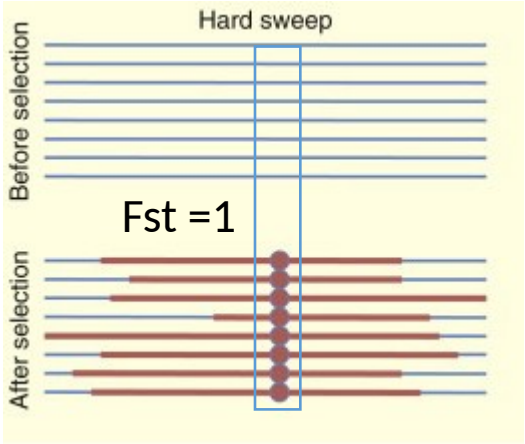
- Linkage disequilibrium based tests
- Genetic differentiation between populations Fst

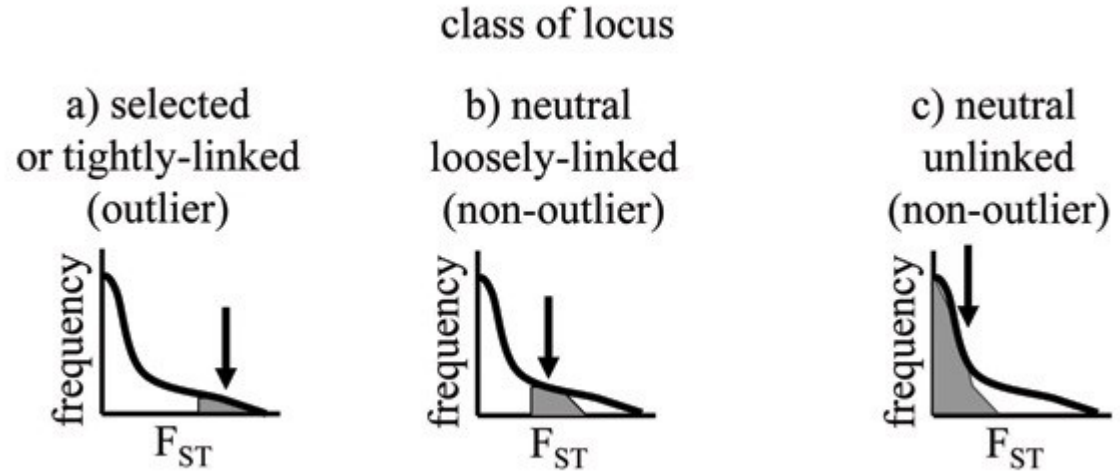
Fst

- genetic differentiation (allele frequency differences) between populations



- In hard sweep selection targets new mutation, result in high Fst between populations





- Fst outliers can point to selected regions

Woodland strawberry (Fragaria vesca ssp. vesca)

- Small perennial plant
- Reproduces by mixed mating (mostly selfing, occasionally outcrossing)
- Propagates also efficiently by runners
- Small genome, 211Mb /7 chromosomes

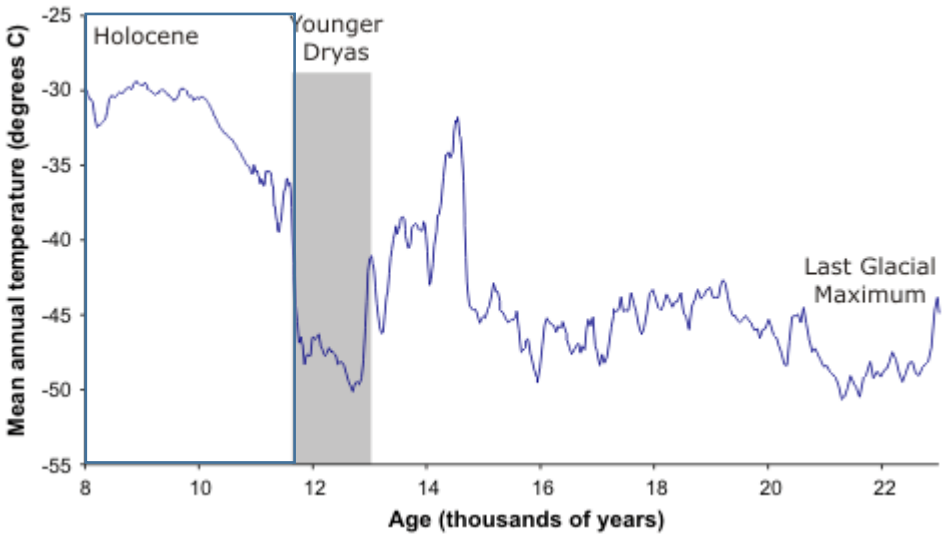
Fragaria vesca ssp. vesca

Wide geographical distribution
-Excellent model for climate adaptation studies

Brown: ssp. vesca
Green: ssp. americana
Yellow: ssp. bracteata
Red: ssp. californica

- Adapted to different environments from southern Spain to most northern parts of Norway
- Which genes underlie adaptation to current locations?

Environmental conditions have changed radically in Europe after the last glacial maximum



- Temperature has elevated steeply after the last glacial maximum
- We are still expecting to see clear signals of sweeps concerning climate warming

Preparation of strawberry NGS data for population genetic analyses

1. **Alignment of reads against reference sequence**

- 150bp paired end reads were aligned against the referense sequence (211Mb) with BWA

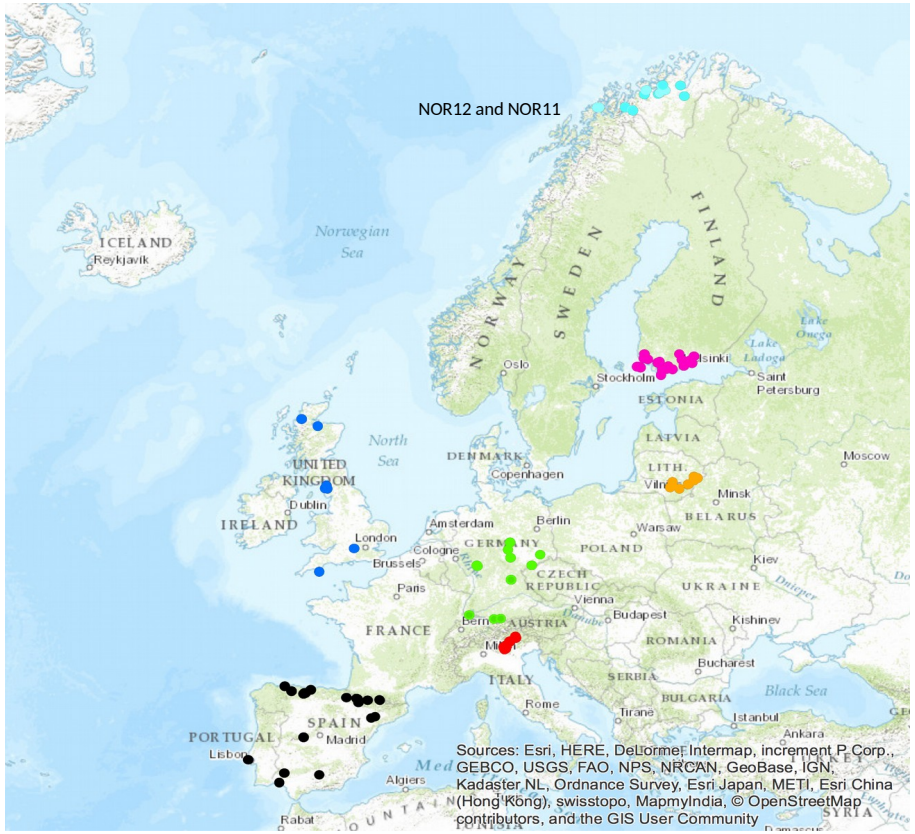
2. **SNP calling and filtering with GATK**

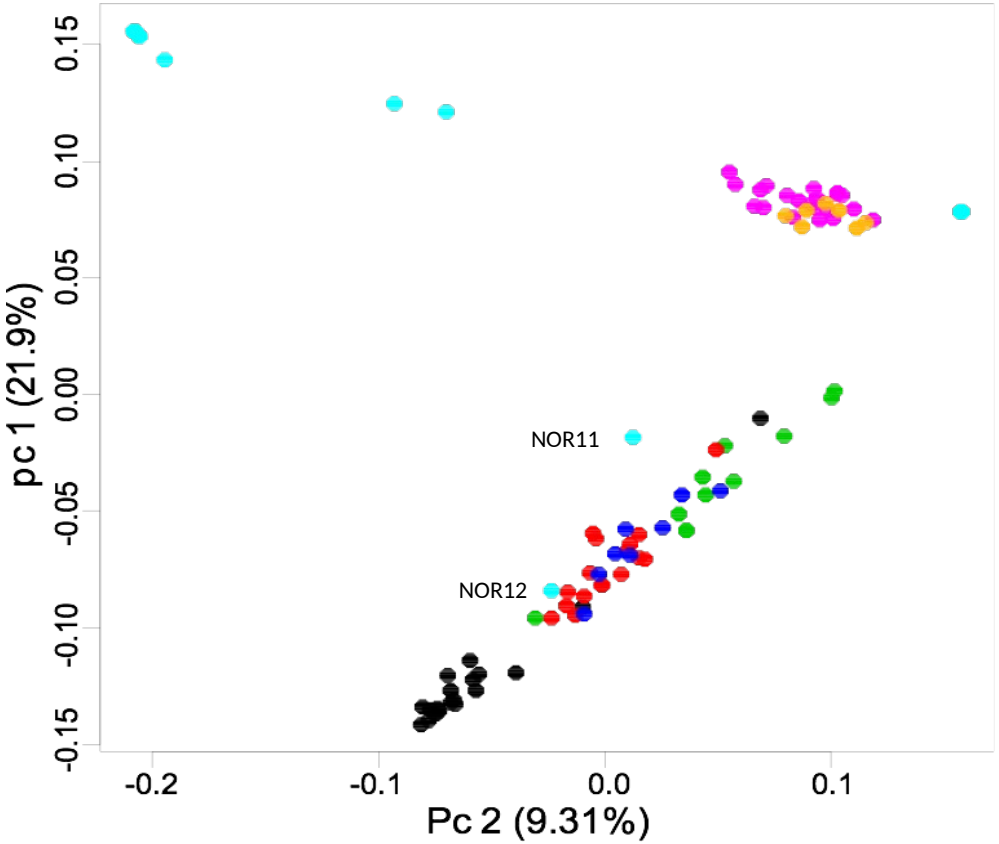
- indels realigned with GATK
- SNP calling for individual samples followed by joint calling (120 samples)

3. **Additional filtering with bcftools and vcftools**

- SNPs less than 20bp around indels were removed
- Minimum coverage/site 10 and maximum 50
- Excessively heterozygote sites within populations ($p<0.001$) were removed
- Mean coverage 18,
- 2 million high quality SNPs

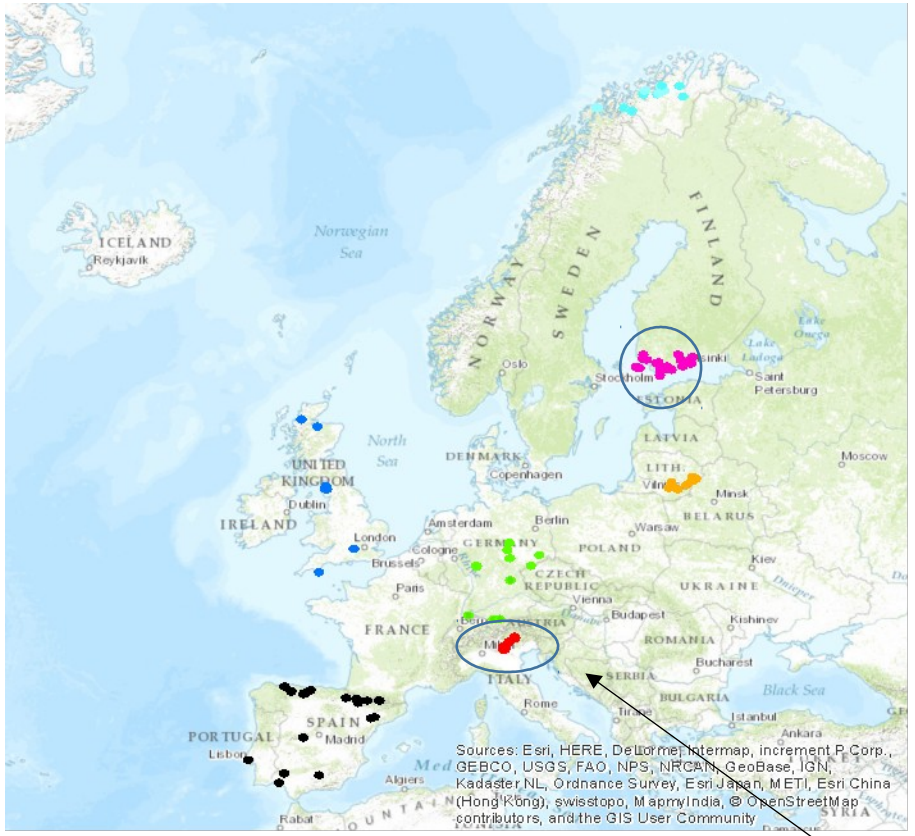
Geographical and genetic structure of 120 strawberry samples





Data for the practical:

- 17 Italian strawberry samples
- 21 Finnish strawberry samples
- Chromosome 6



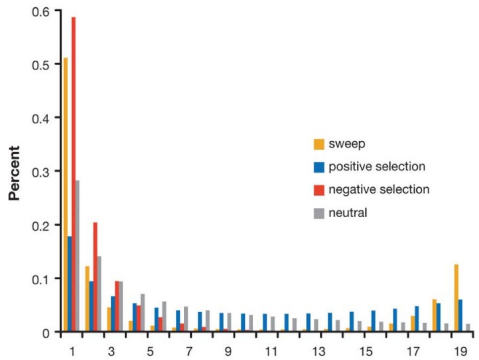
- Find potentially selected genomic regions in Italian samples

SweeD (**Sweep Detector, Pavlidis & Alachiotis**)

- SFS will be calculated empirically using all SNPs (obtained from the entire genome)
- In principle robust to demographic events (eg. bottlenecks)

- Calculates composite likelihood ratio along a chromosome (Sweep model/Neutral model)
- Large values suggest a sweep

Genome wide SFS



- Chromosome will be scanned with user specified grid size (eg. 10 kb windows) to find skews in SFS
- In practice, also simulations are being conducted to control demographic events statistically

Input file format for SweeD

Number of derived alleles

Number of chromosomes /site

Genomic position

position	x	n	folded
193	0	34	0
1781	0	34	0
1804	0	34	0
1899	8	34	0
1945	34	34	0
15021	0	34	0
15075	3	34	0
15117	19	34	0
15714	3	34	0
15783	2	34	0
15803	3	34	0

Only polymorphic sites are used for this analysis

- Only unfolded variable sites (ancestral allele known (or predicted) are used for analysis
- *Fragaria iinumae* has been used as an outgroup species to infer ancestral allele