

# Topic 8+9: Population genomics and plotting

# Learning Goals

- Understand the basic principals, how to run and visualize  $F_{ST}$ , STRUCTURE and PCA analyses.
- Be able to list multiple methods for detecting selection in genomic data.
- Be able to list multiple methods for detecting hybrid ancestry.

# Considerations for SNPs

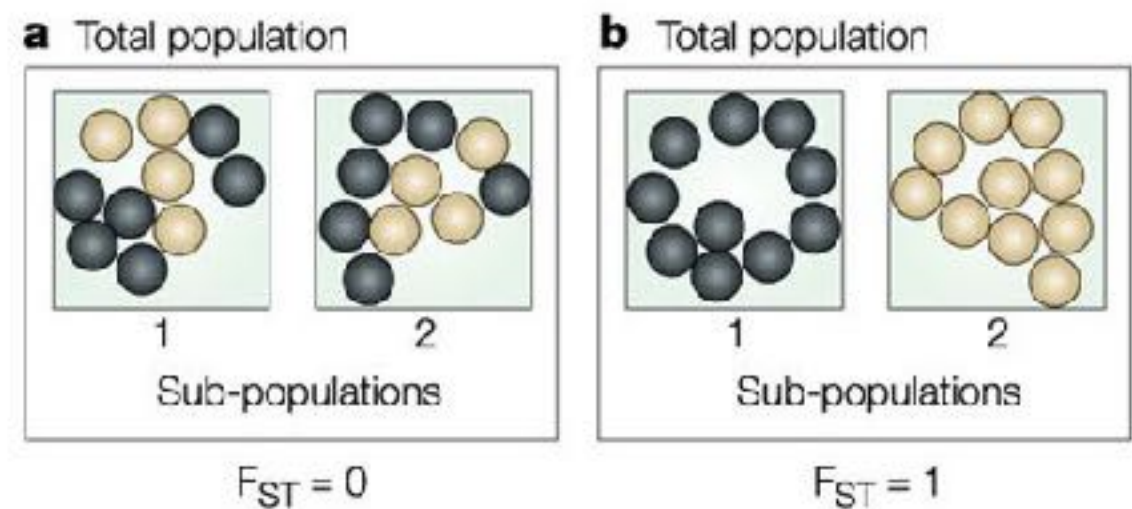
- Ascertainment bias
  - Typically only keep variable sites, can bias diversity estimates
- Linkage
  - With thousands of sites, some will be in close linkage.
- Quality filtering
  - You must filter your SNPs to remove false SNPs, sometimes difficult

# Population structure

- $F_{ST}$
- PCA
- STRUCTURE

# $F_{ST}$

- $F_{ST} = H_T - H_O / H_T$
- $H_T$  = Expected heterozygosity using global allele frequency based on Hardy-Weinberg
- $H_O$  = Average observed heterozygosity



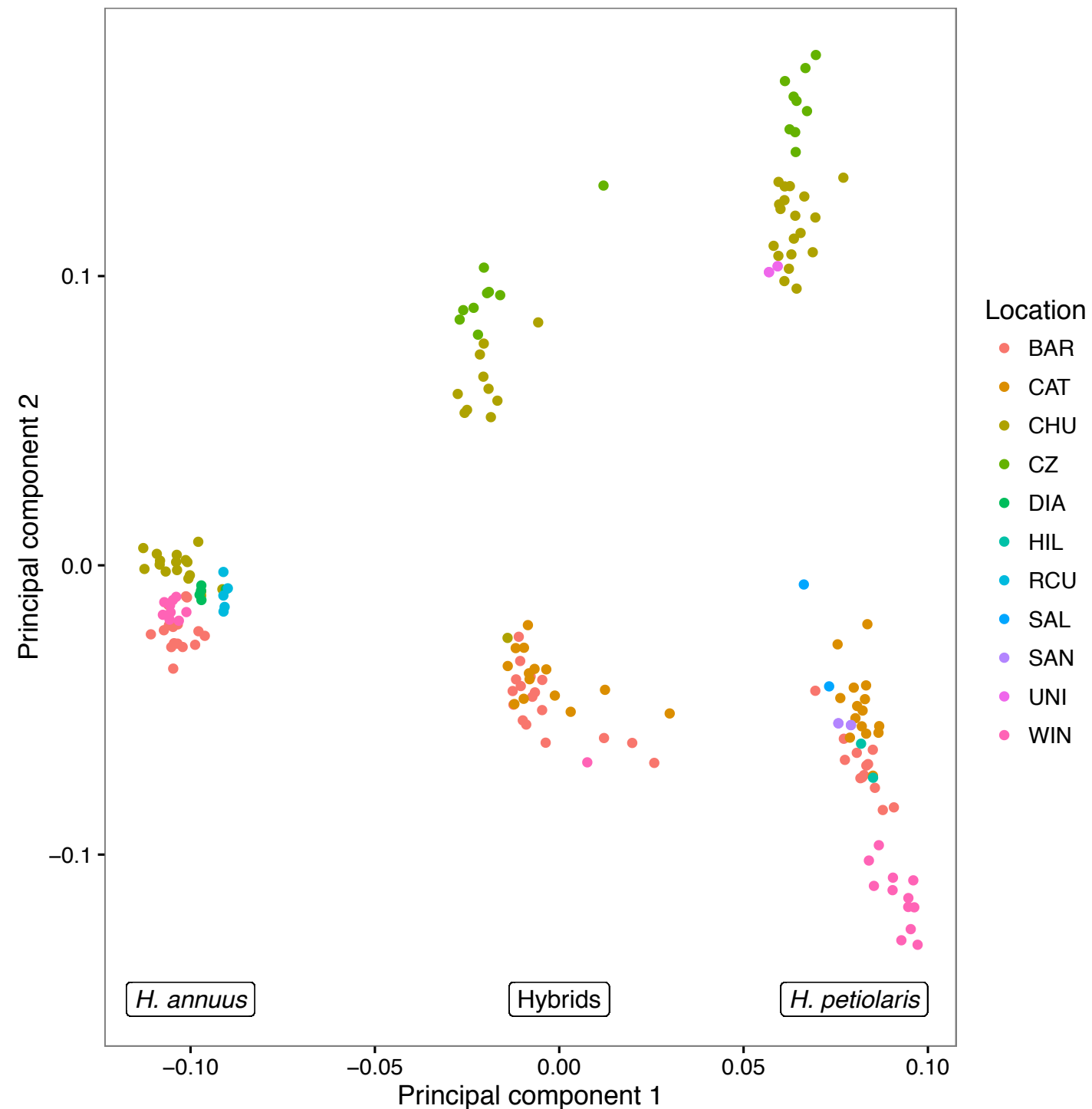
# F<sub>ST</sub> Programs

- hierfstat (R)
- **SNPrelate (R)**
- FSTAT
- Arlequin
- vcftools

# Principal Component Analysis

- Converts a set of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

# Principal Component Analysis





# Principal Component Analysis

- Converts a set of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.
- Great first step to visualize data
- You should prune dataset to unlinked SNPs

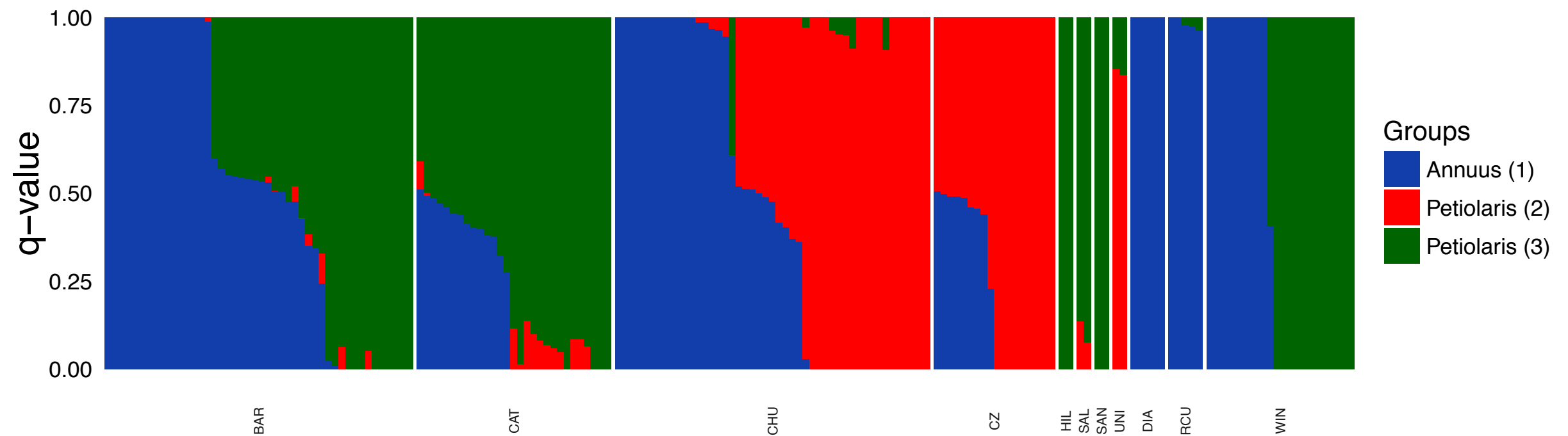
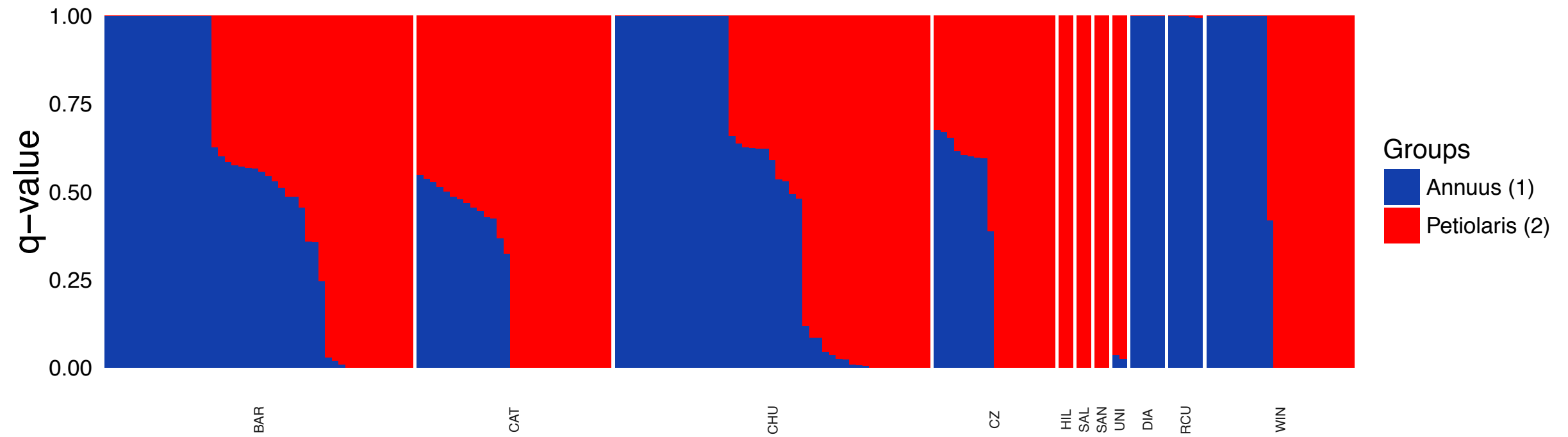
# PCA Programs

- **SNPrelate (R)**
- adegenet (R)
- SPSS

# STRUCTURE

- Models  $K$  populations with a set of allele frequencies at each locus.
- Individuals are assigned to one or more populations based on their genotype
- Can pick the best  $K$  based on your data

# STRUCTURE



# STRUCTURE

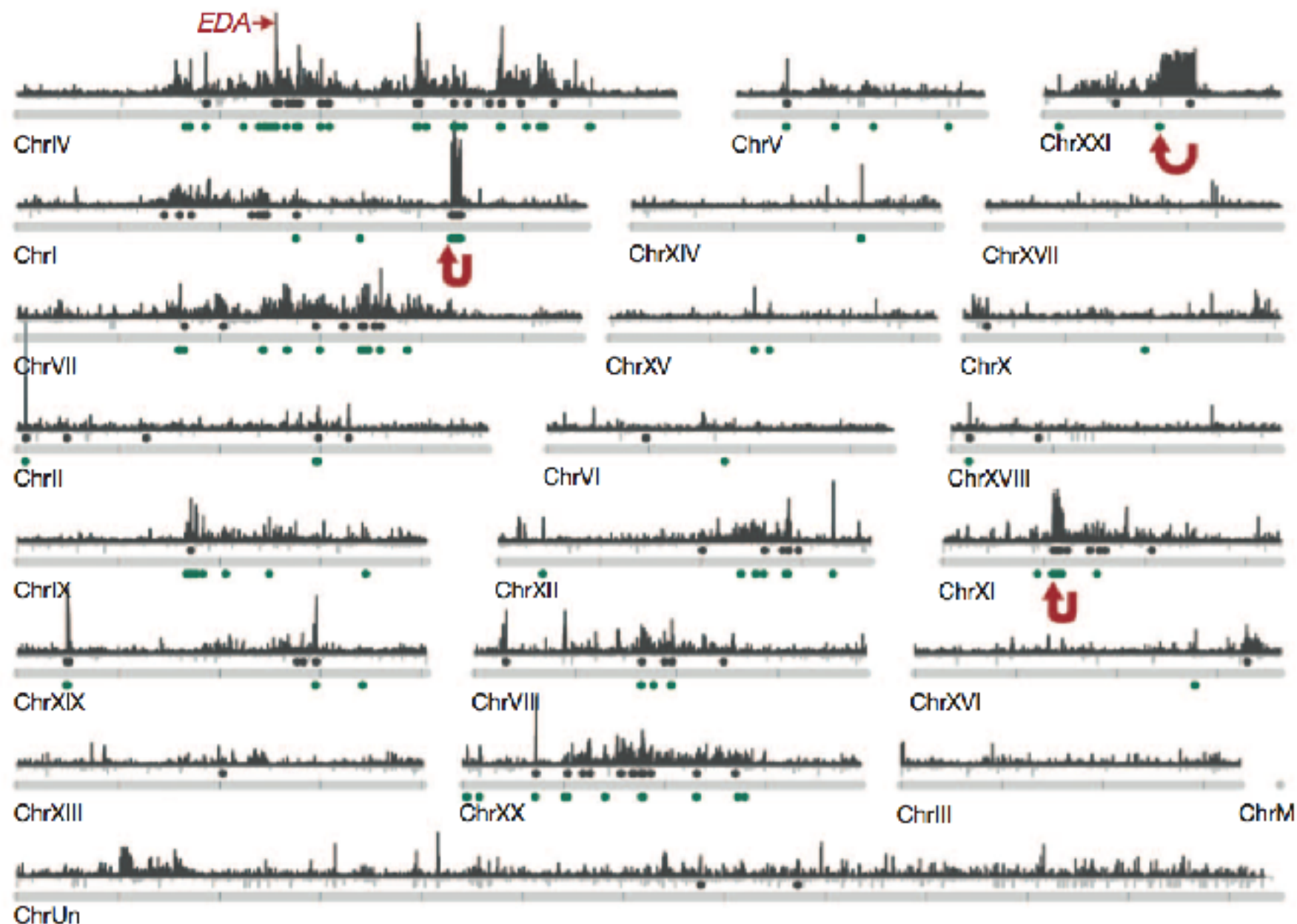
- You should prune dataset to unlinked SNPs
- Run multiple times to confirm consistency

# STRUCTURE programs

- STRUCTURE
- Admixture
- **FASTstructure**
- NGSadmixture

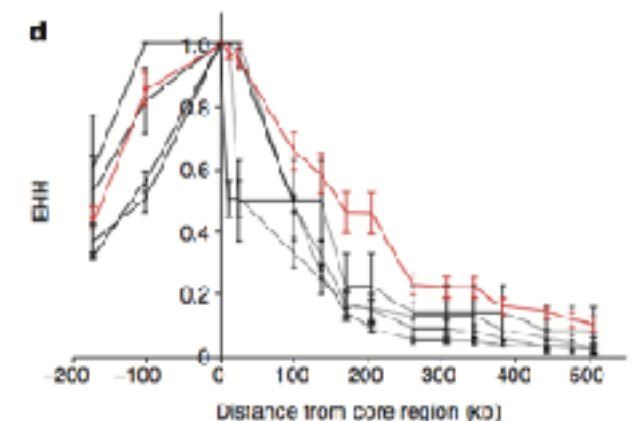
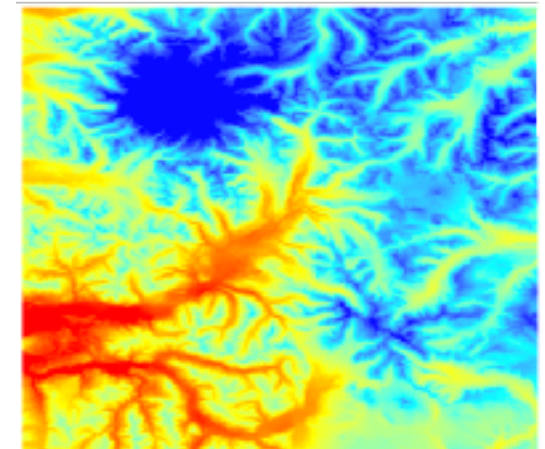
# Signatures of adaptation

- Which loci are contributing to local adaptation?



# Outlier tests and signatures of selection

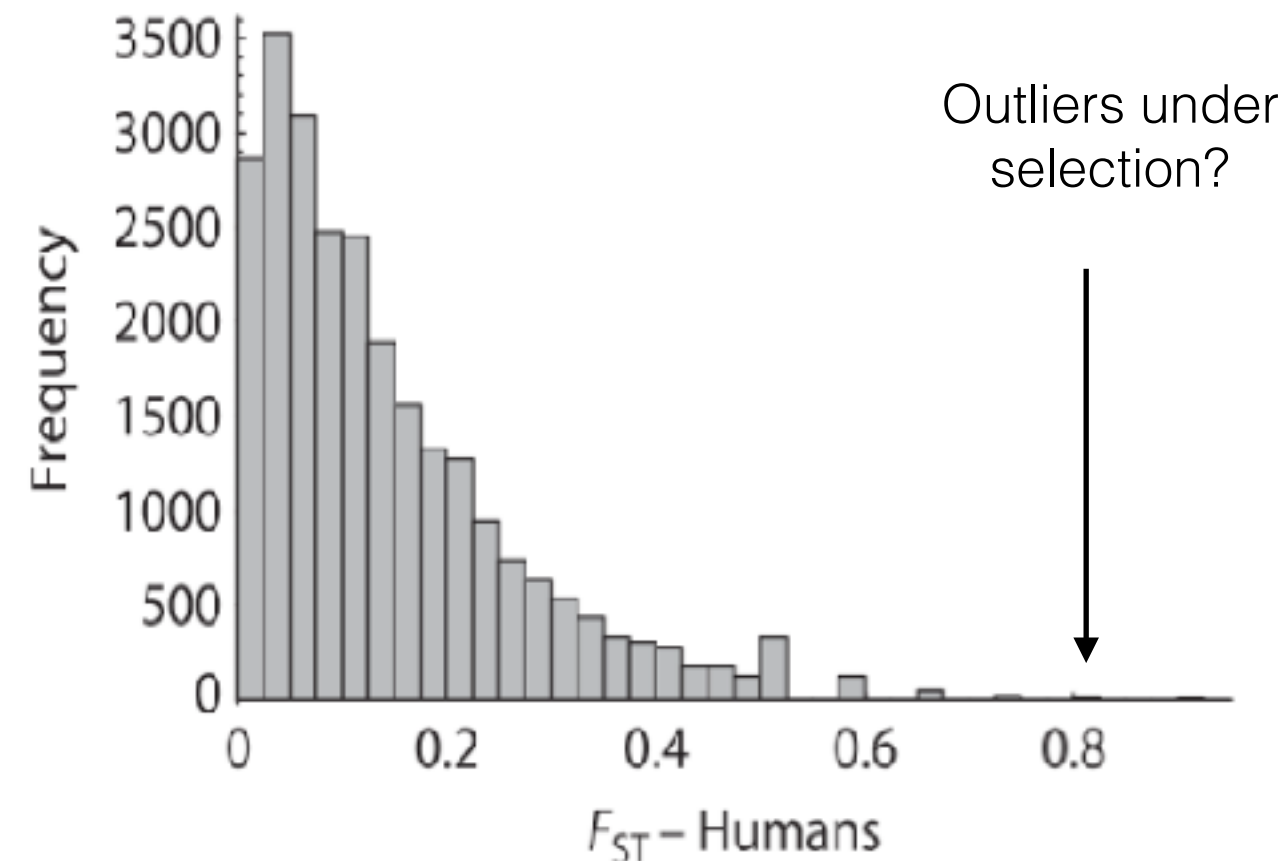
- Landscape level: differences in allele frequency among populations or environments
- Phenotypic level: associations between alleles and locally adapted phenotypes
- Sequence level: Changes in allelic diversity along a chromosome





# $F_{ST}$ outlier tests

## $F_{ST}$ - outlier tests

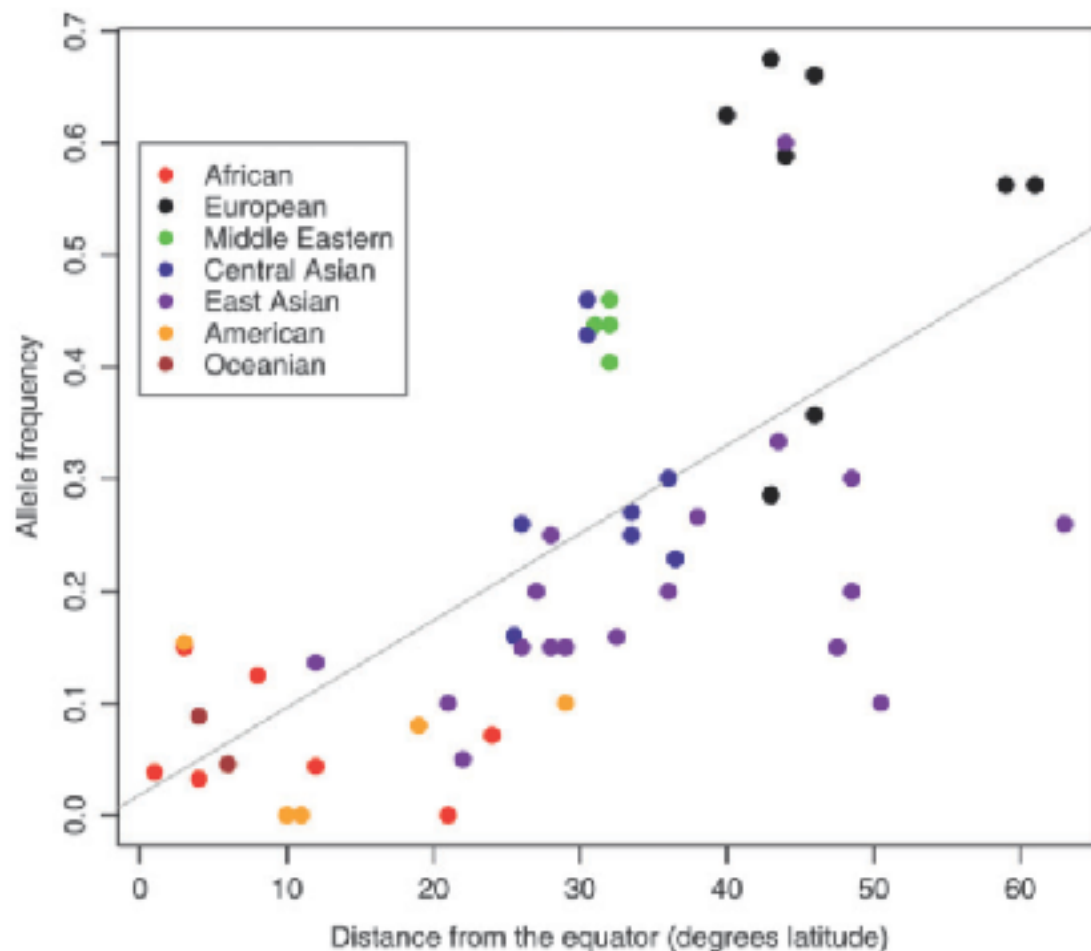


- Compare the outliers to some background distribution expected under a null model to evaluate significance
- **BayeScan**: uses a Dirichlet distribution, akin to assuming no migration nor mutation since common ancestry
- **FDIST**: assumes an island model, uses IM to estimate demography
- BayENV
- OUTFLANK

# Other causes of $F_{ST}$ outliers

- Background selection against deleterious mutations.
- Cryptic hybrid zones involving multiple loci involved in reproductive isolation.
- Stochastic effects at wave-edge of an expanding population.
- Species-wide selective sweep.

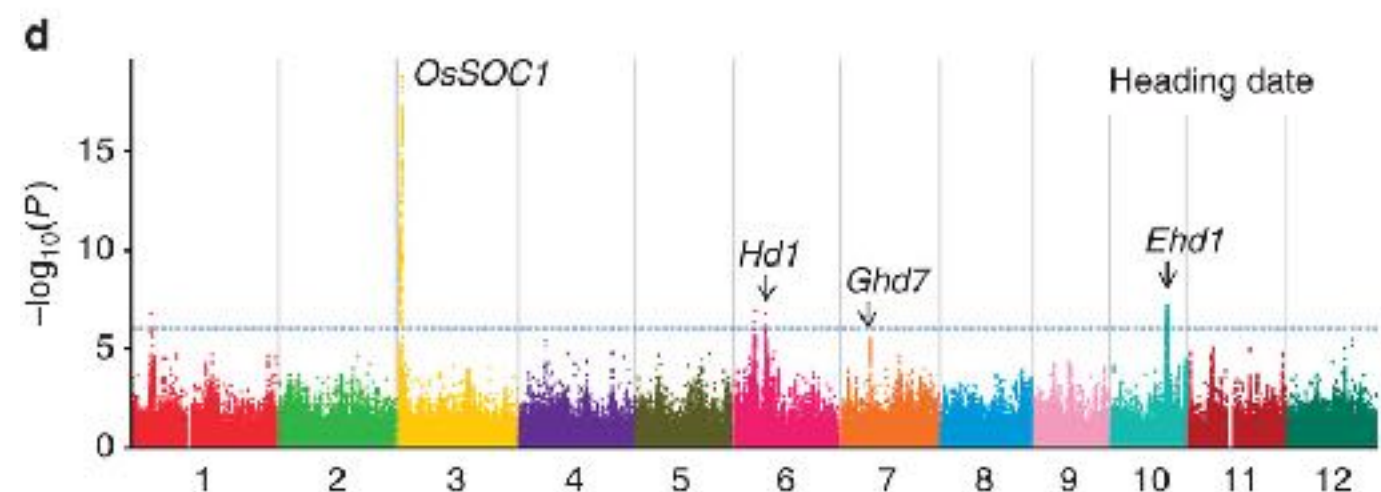
# Environment-allele associations



- Tests correlation between SNP and environment after controlling for population structure
- **Bayenv**: estimates covariance matrix representing population structure from separate set of “neutral” loci
- **LFMM**: Latent-factor mixed model; estimates the population structure from SNPs in test panel

# SNP-phenotype associations (GWAS): one allele at a time

- Regression of phenotype on SNP
- Use PCA or STRUCTURE as a covariate in a linear model or a kinship matrix of relatedness in a mixed effect model
- Yields an estimate of the association between SNP and phenotype beyond what would be expected due to population structure



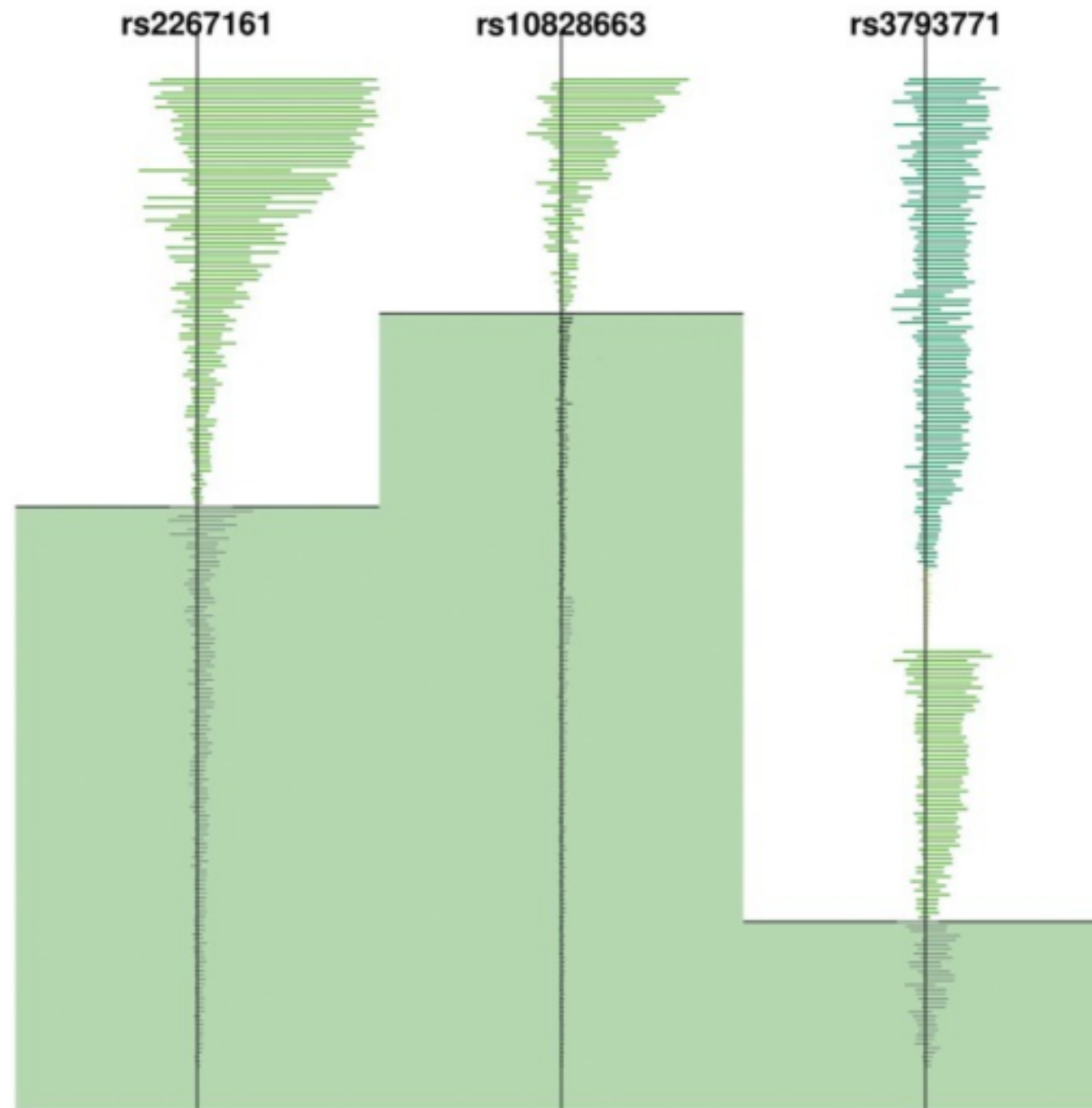
# GWAS programs

- Tassel
- ANGSD
- GWAStools (R)
- GenABEL (R)
- GCTA

# Haplotype length

- Integrated Haplotype Score (iHS)
  - Measures levels of LD surrounding a derived allele compared to an ancestral allele at the same position.
- Extended Haplotype Homozygosity (EHH)
  - Measures the length of haplotypes in a region.

# Haplotype length



# Haplotype length programs

- selscan
- hapbin
- sweep

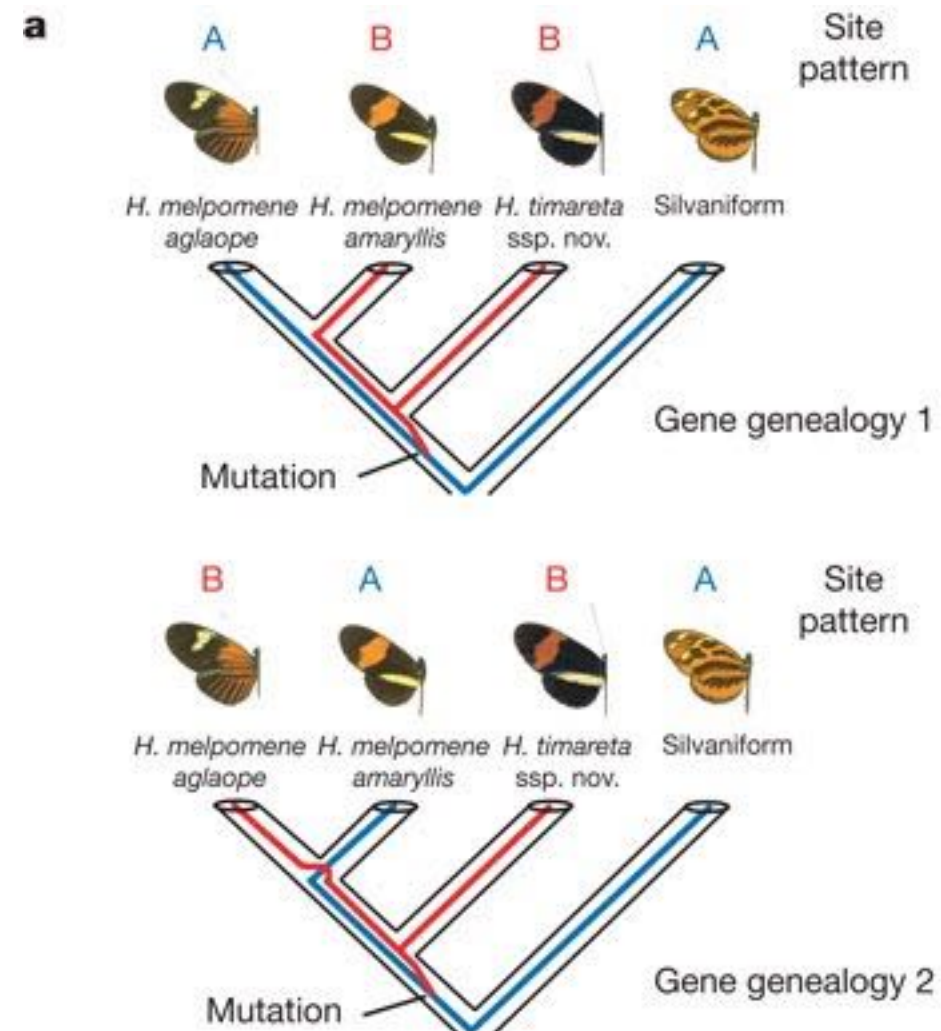


# Hybrid ancestry

- Are genotyped samples hybrids?
- What generation hybrid?
- Where does each loci come from?

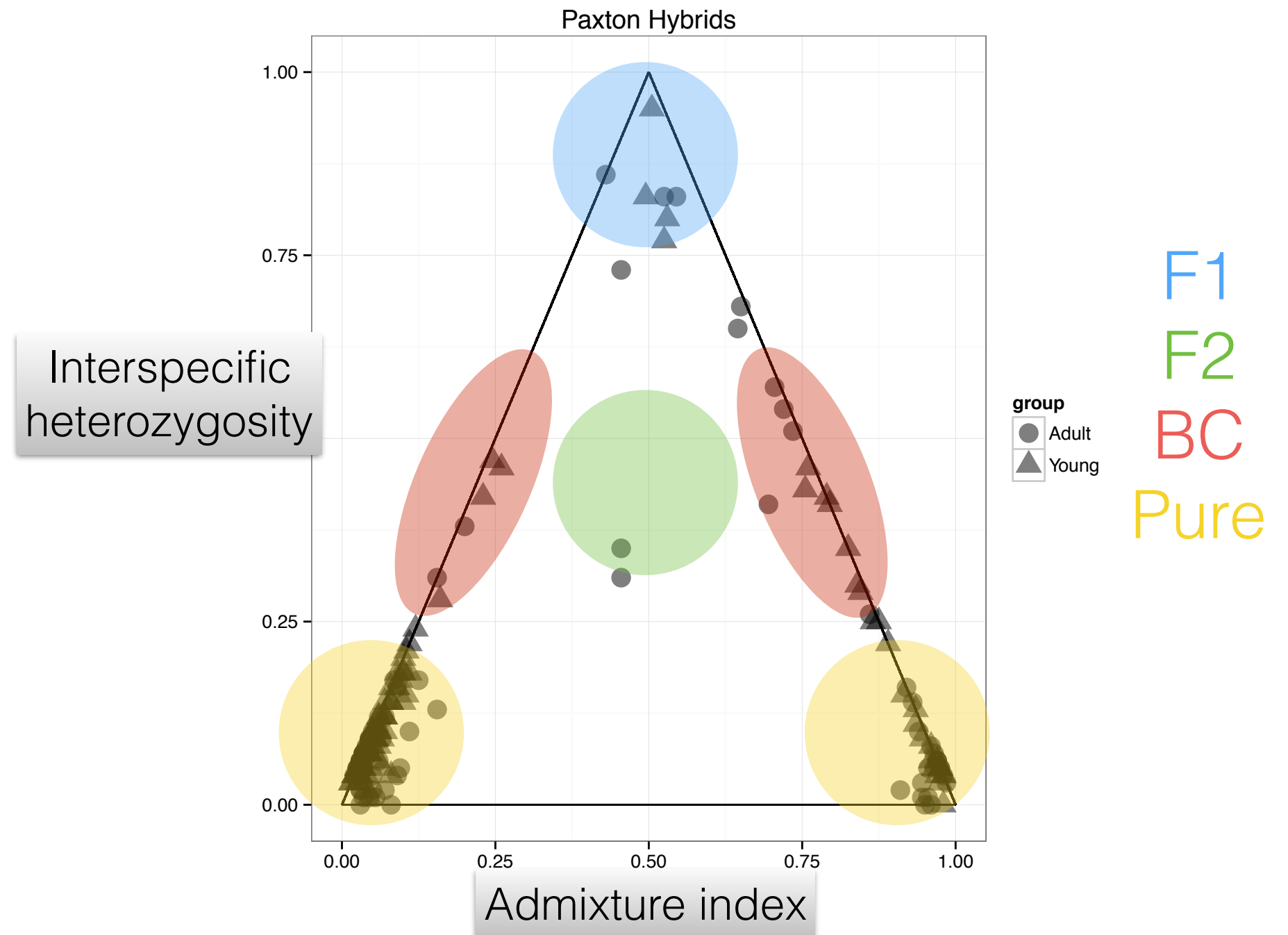
# Are genotyped samples hybrids?

- STRUCTURE
- PCA
- ABBA-BABA



# What generation hybrid?

- **Hlest**
- Hindex



# Where does each loci come from?

- SABER
- STRUCTURE linkage model
- MSG
- Ancestry\_HMM

