

Topic 8+9: Population genomics and plotting

Biol 525D - Bioinformatics for Evolutionary Biology
2019

Learning Goals

- Understand the principals behind basic population genetic visualization methods
 - F_{ST} , STRUCTURE and PCA analyses.
- Be able to plot results of these programs

Considerations for SNPs

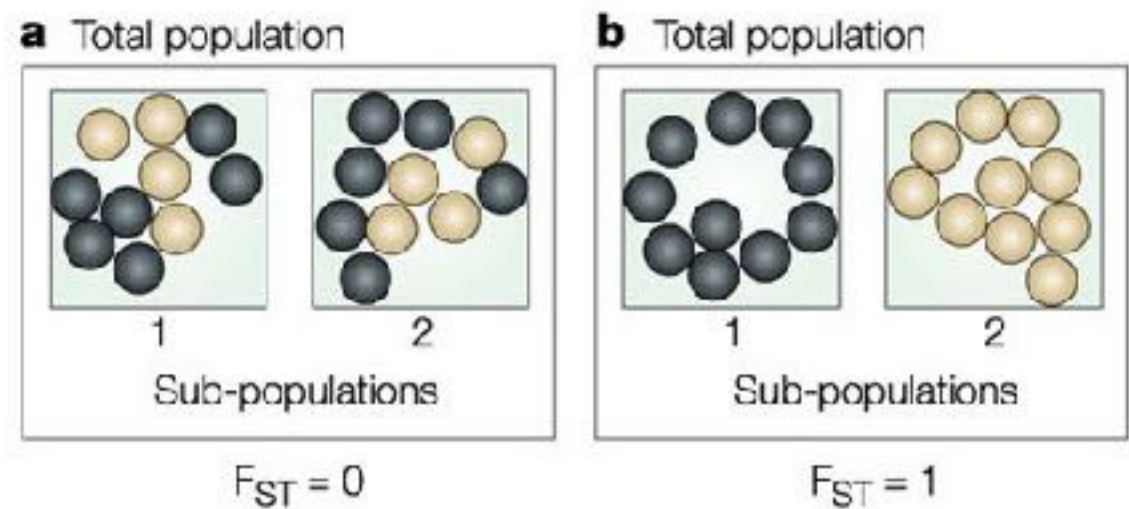
- Ascertainment bias
 - Typically only keep variable sites, can bias diversity estimates
- Linkage
 - With thousands of sites, some will be in close linkage.
- Quality filtering
 - You must filter your SNPs to remove false SNPs, sometimes difficult

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



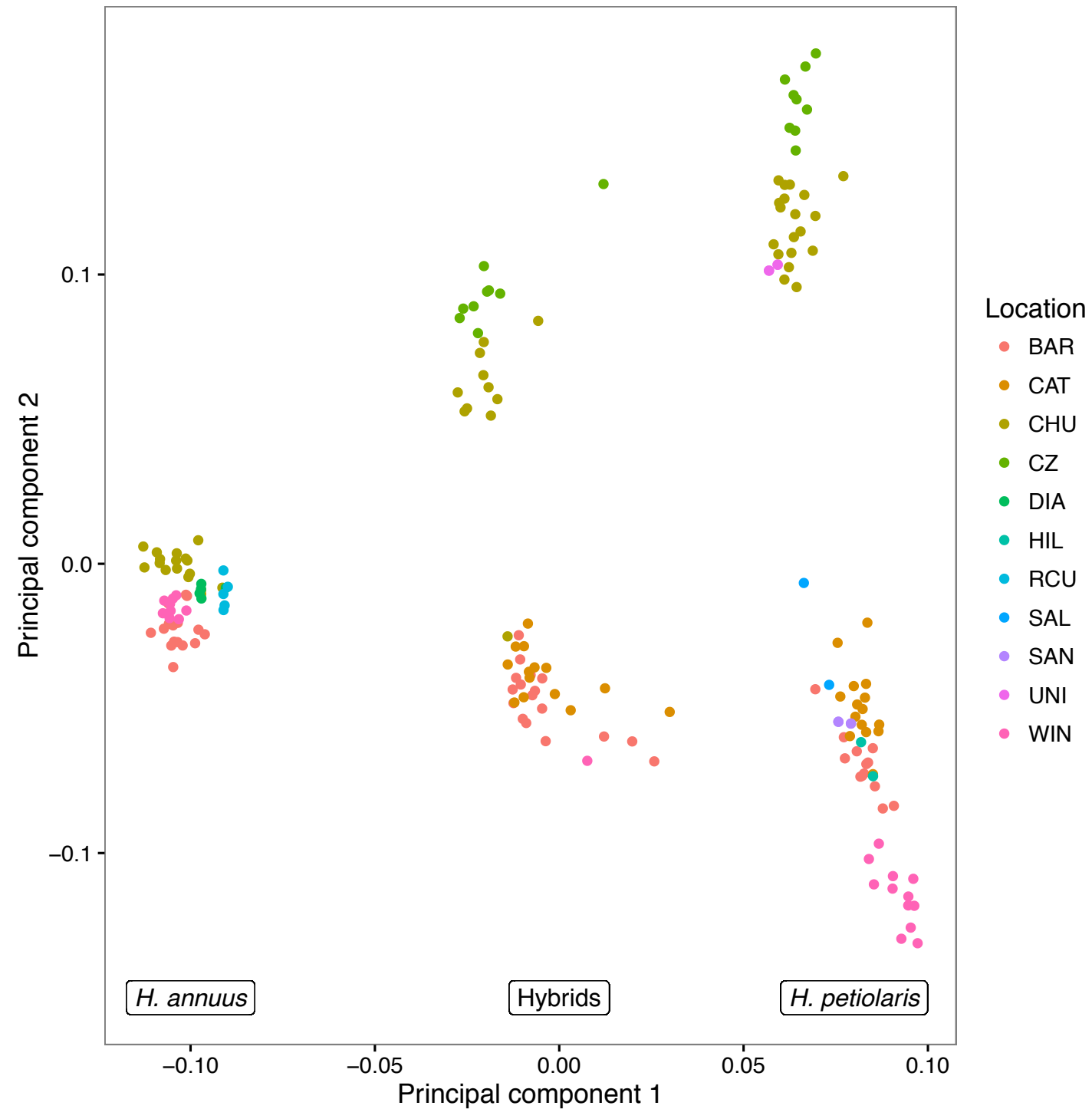
F_{ST} Programs

- hierfstat (R)
- **SNPrelate (R)**
- FSTAT
- Arlequin
- vcftools
- scikit-allel (python)

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

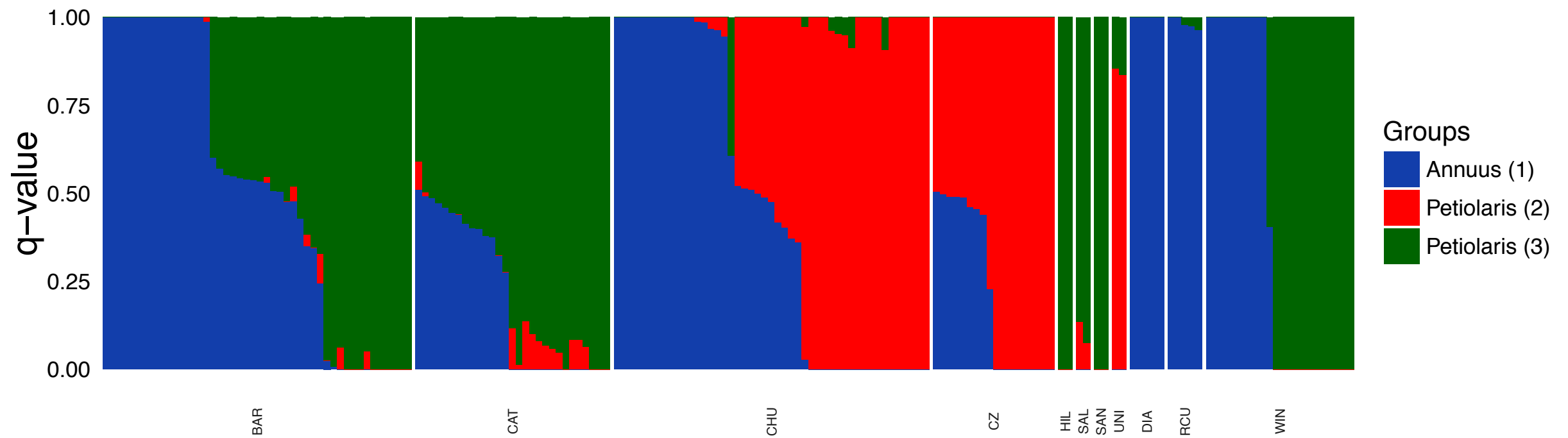
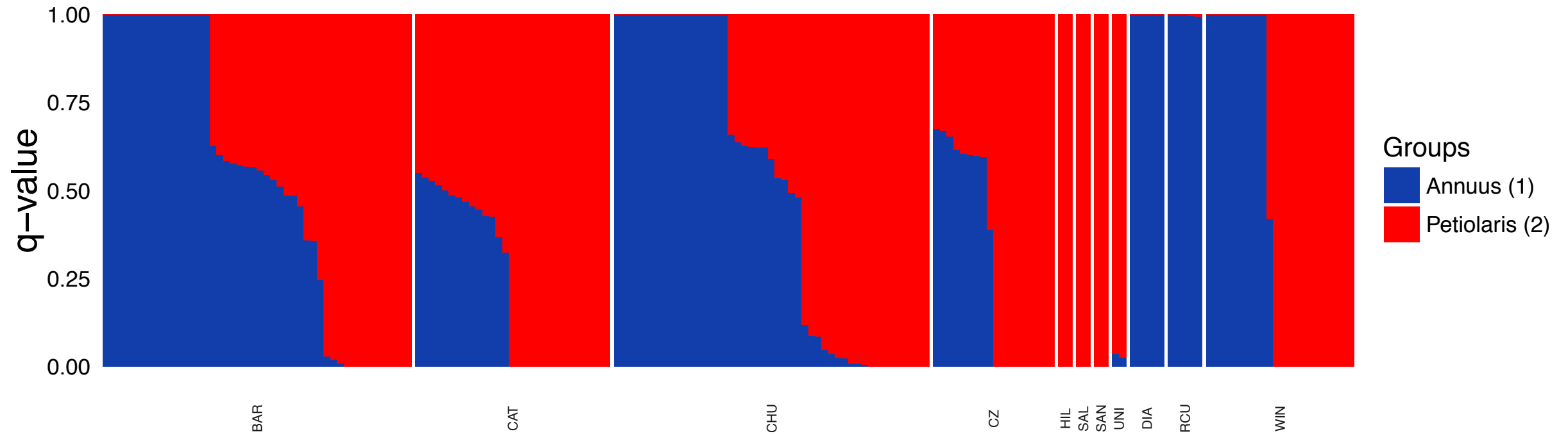
PCA Programs

- **SNPrelate (R)**
- adegnet (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

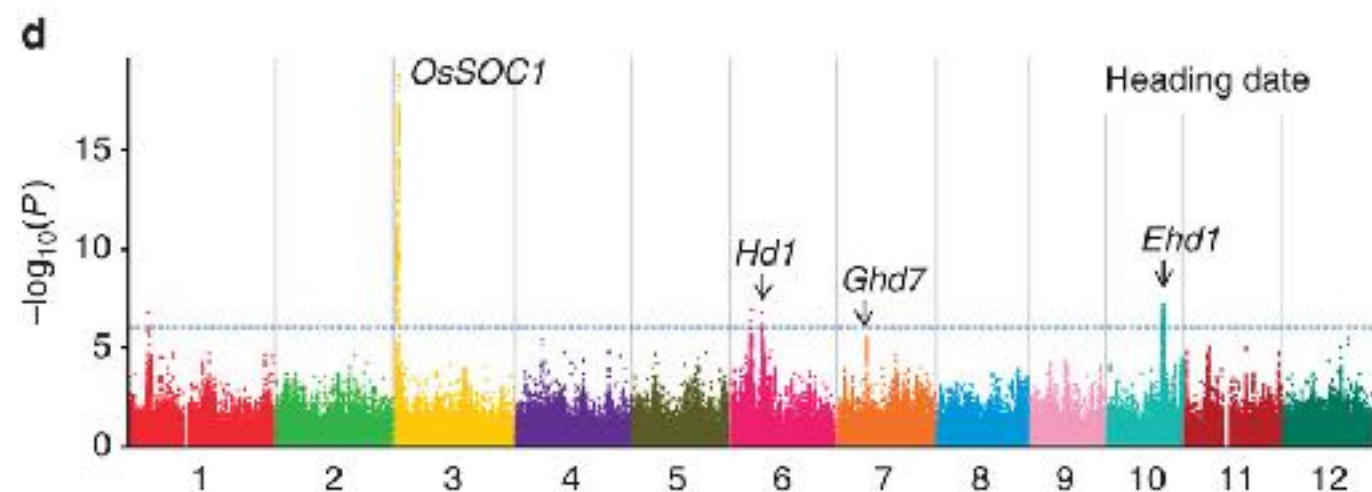
- You should prune dataset to unlinked SNPs
- Run multiple times to confirm consistency

STRUCTURE programs

- STRUCTURE
- Admixture
- **FASTstructure**
- NGSadmix

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

Plotting

- dplyr for data manipulation
- ggplot2 for plotting

