

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.
- Become familiar with approaches for studying selection, and the architecture of (adaptive?) traits
 - Selection scans, GWAS
- Be able to plot results of these programs

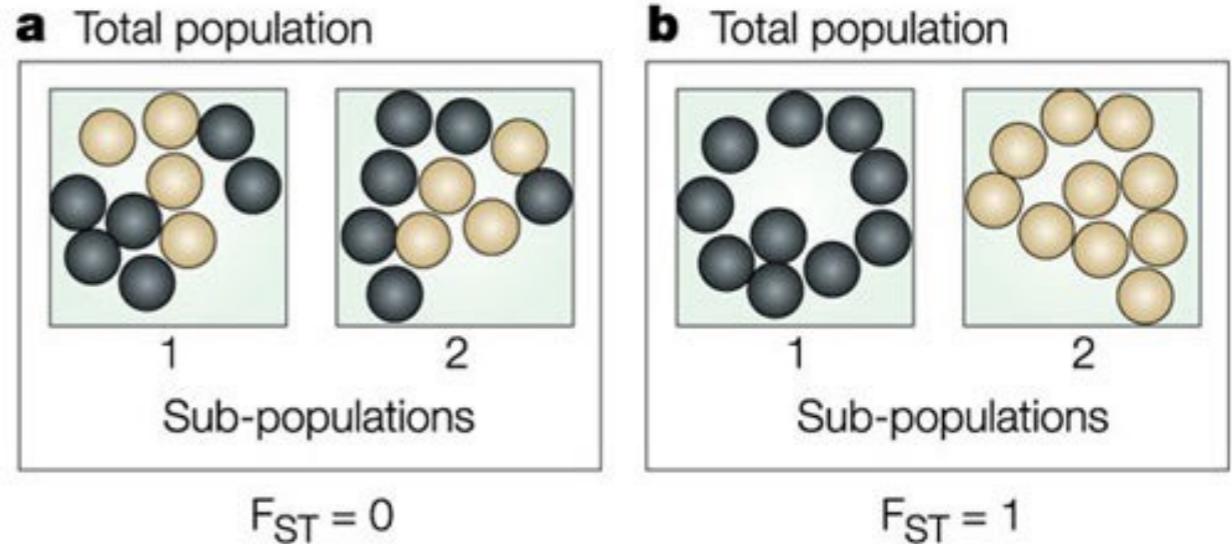
Considerations for SNPs

- Ascertainment bias
 - Typically only keep variable sites, can bias diversity estimates
- Linkage
 - With thousands (or millions) 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}



Nature Reviews | Genetics

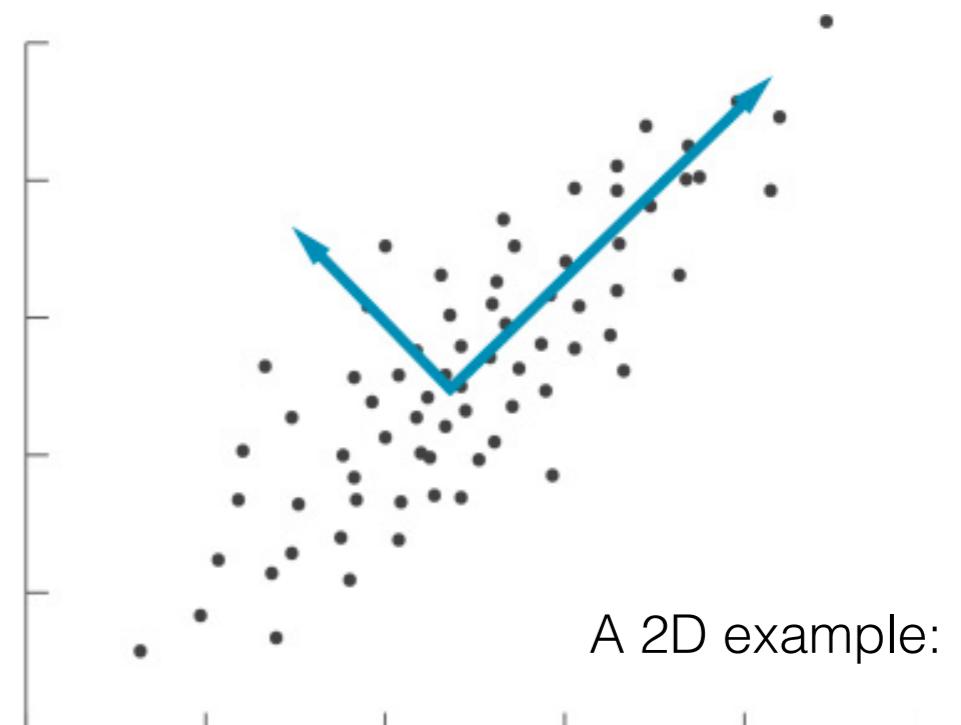
- $F_{ST} = (\pi_{\text{between}} - \pi_{\text{within}}) / \pi_{\text{between}}$
- $F_{ST} = 1 - (H_s/H_T)$
 - H_T = Expected heterozygosity (based on Hardy-Weinberg) of the **total population**
 - H_T = Expected heterozygosity (based on Hardy-Weinberg) of the **subpopulation**

F_{ST} Programs

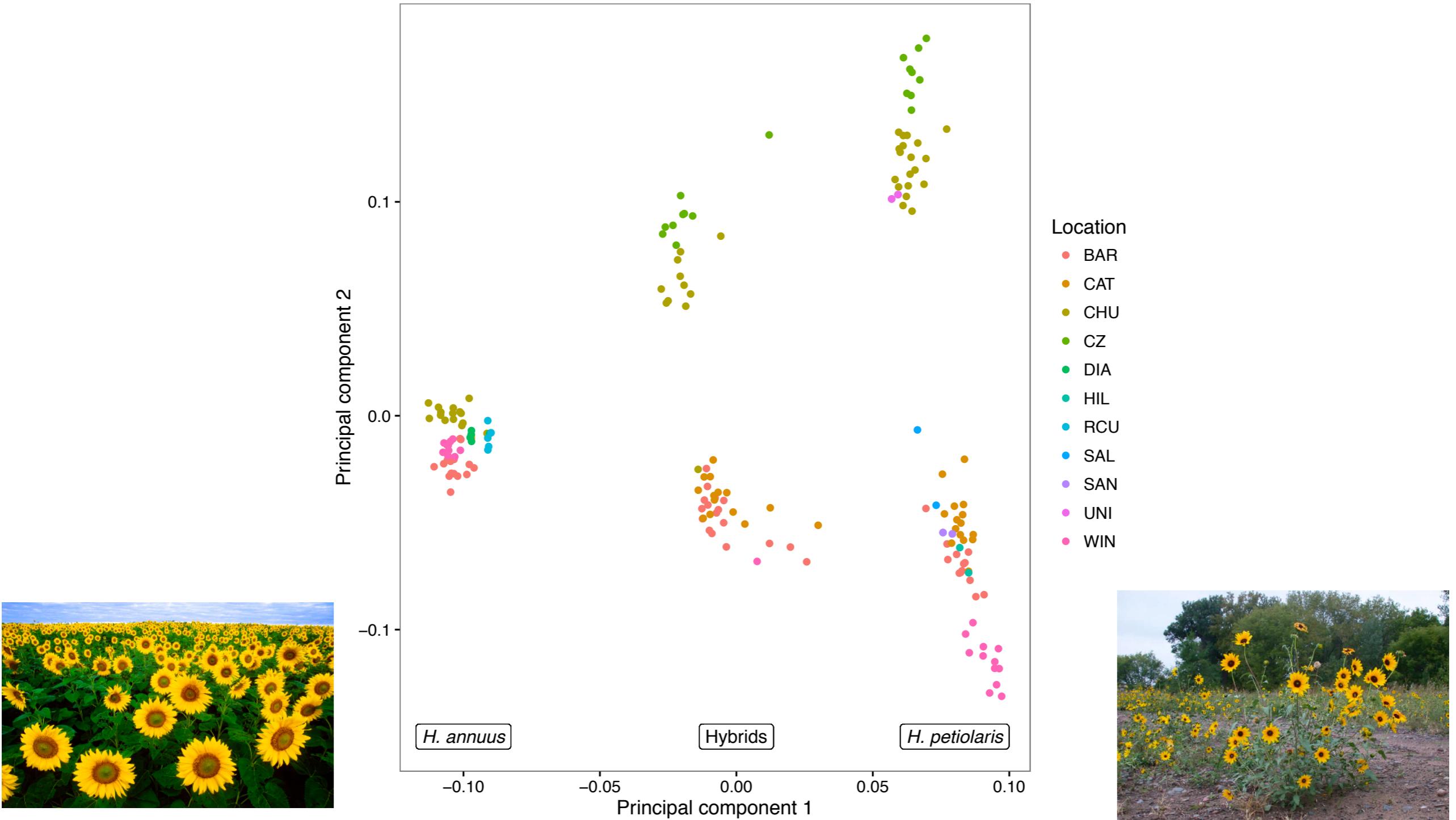
- hierfstat (R)
- SNPrelate (R)
- FSTAT
 - Many different Fst estimators:
*Wright, Weir and Cockerham,
and Hudson et al.*
- Arlequin
- **vcftools**
- scikit-allel (python)

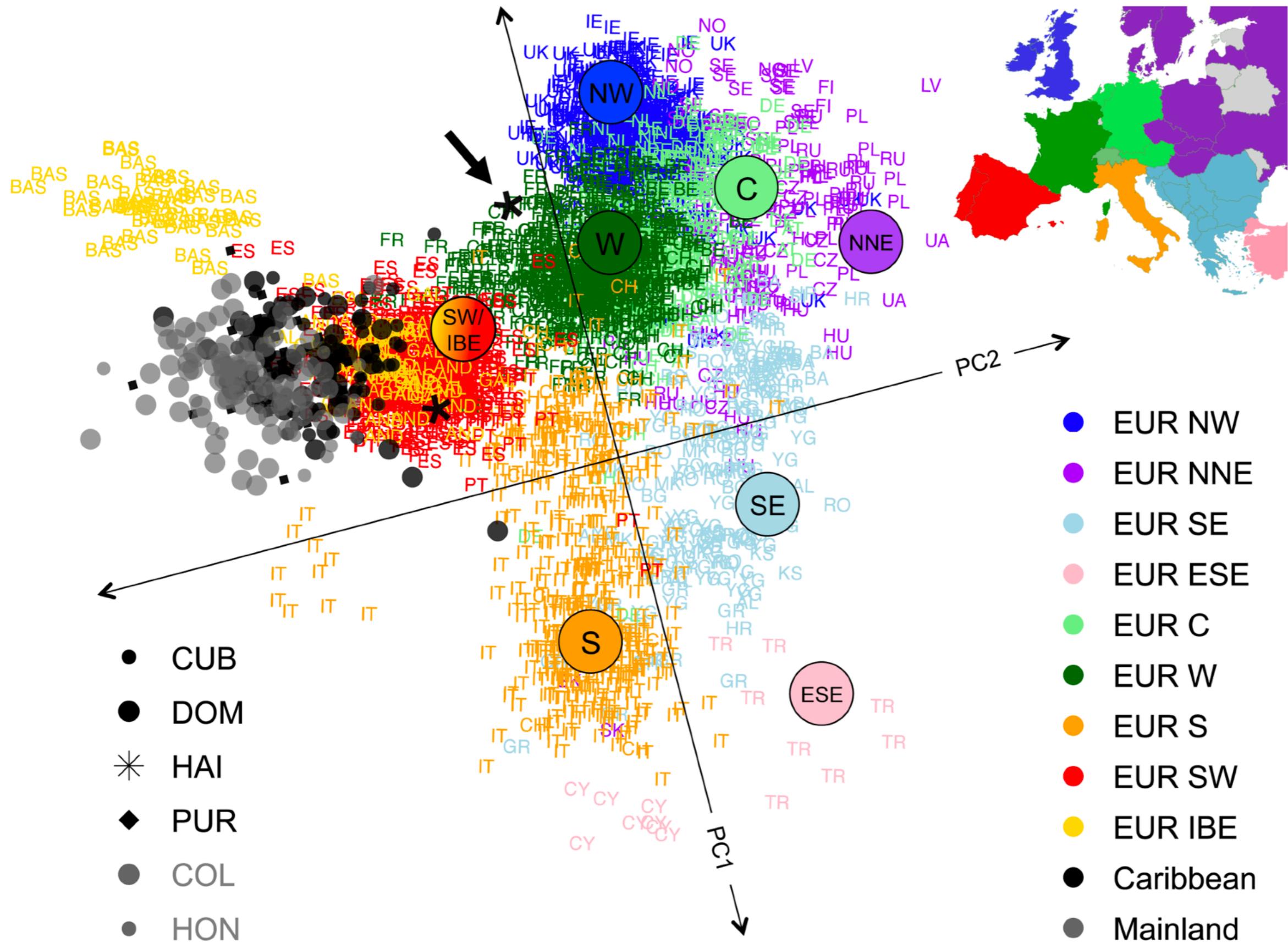
Principal Component Analysis

- Model-free approach to assaying populations structure
- 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

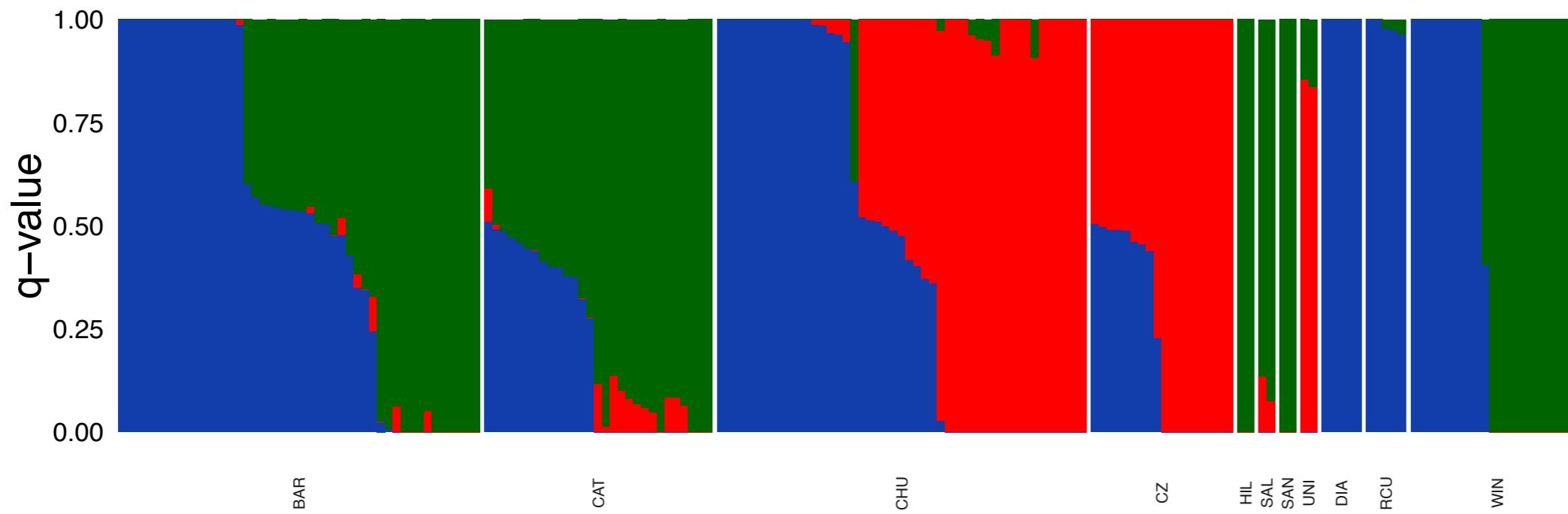
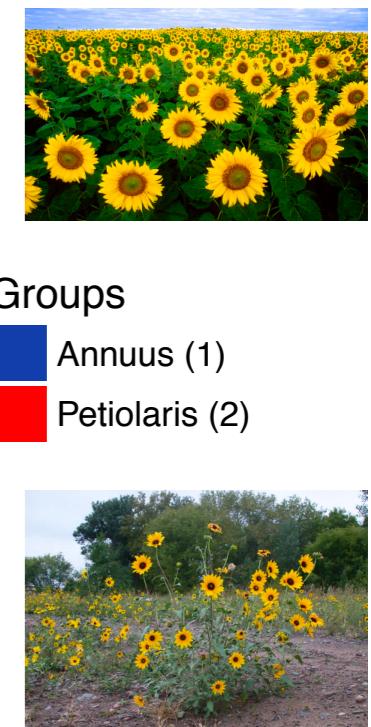
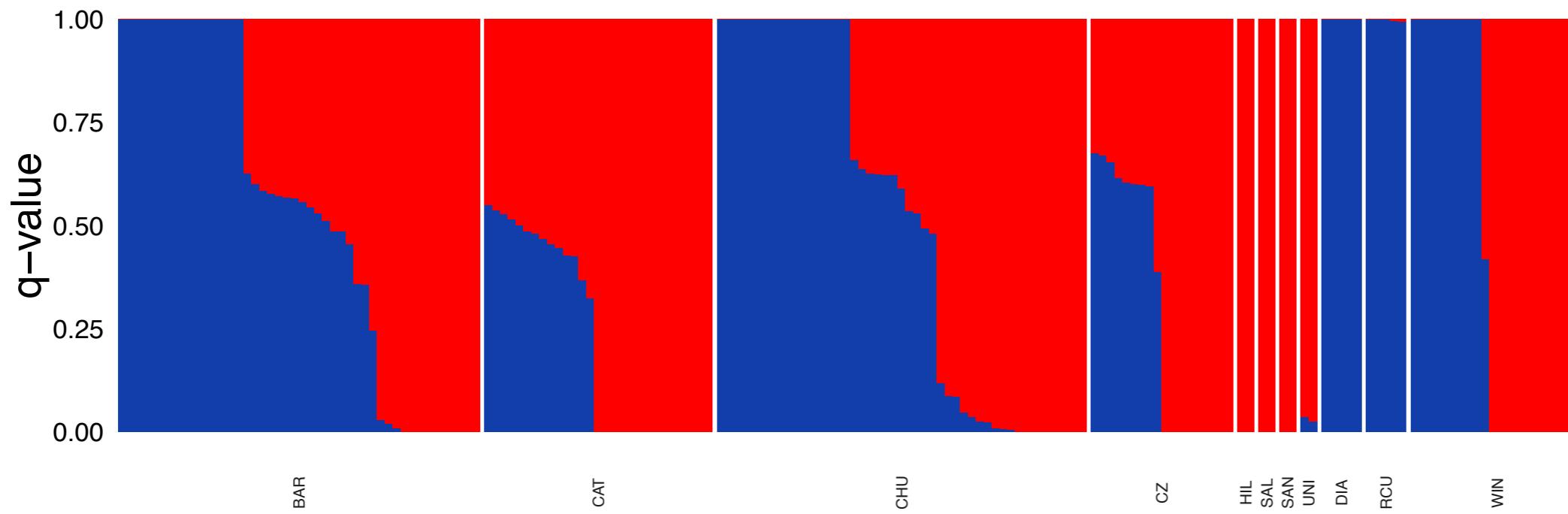
PCA Programs

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

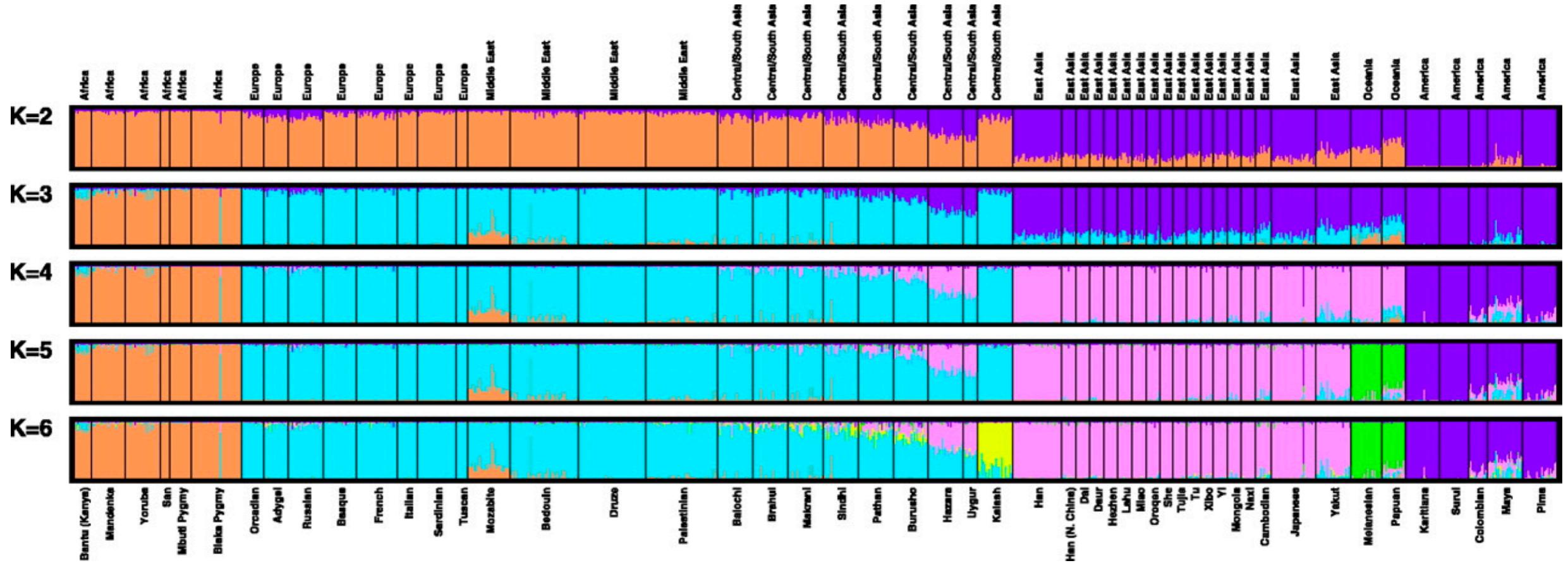
STRUCTURE

- **Models** K populations with a set of allele frequencies at each locus, *assumes hardy-weinberg*
- Individuals are assigned to one or more populations based on their genotype
- Can pick the best K based on *the fit of the model to your data*

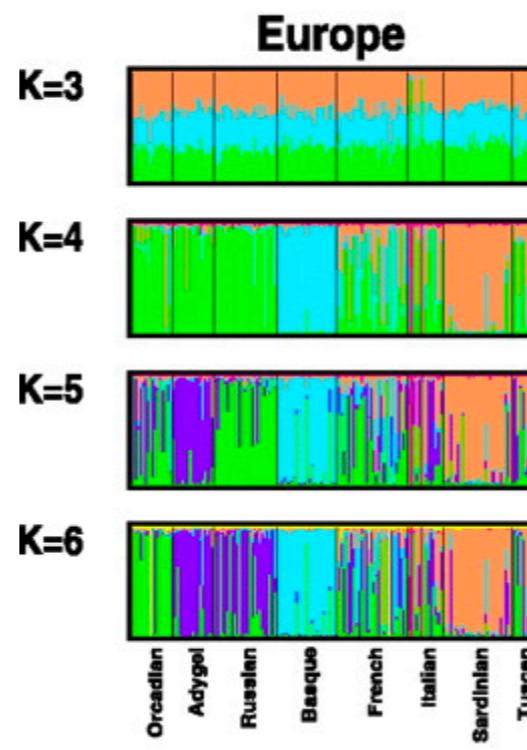
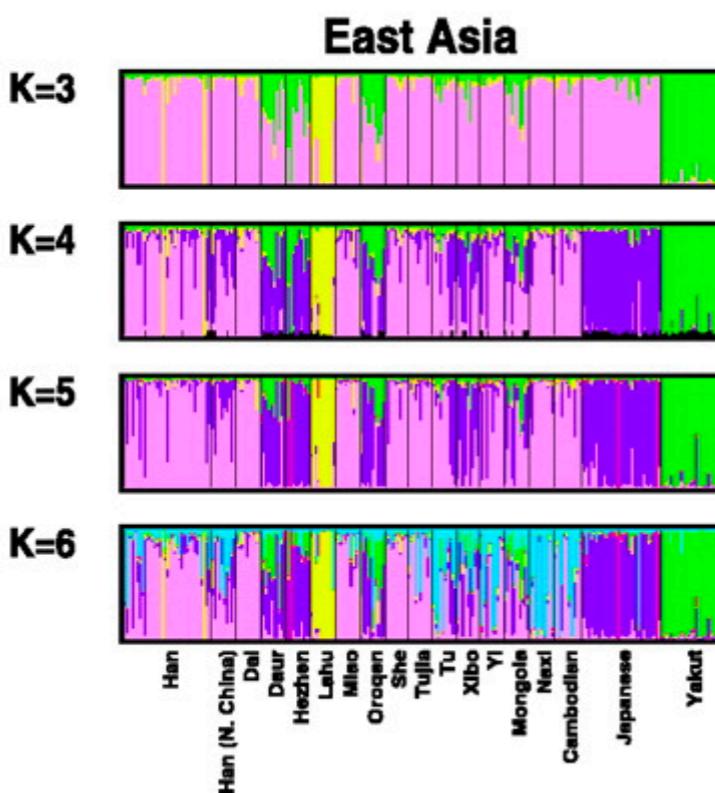
STRUCTURE



STRUCTURE



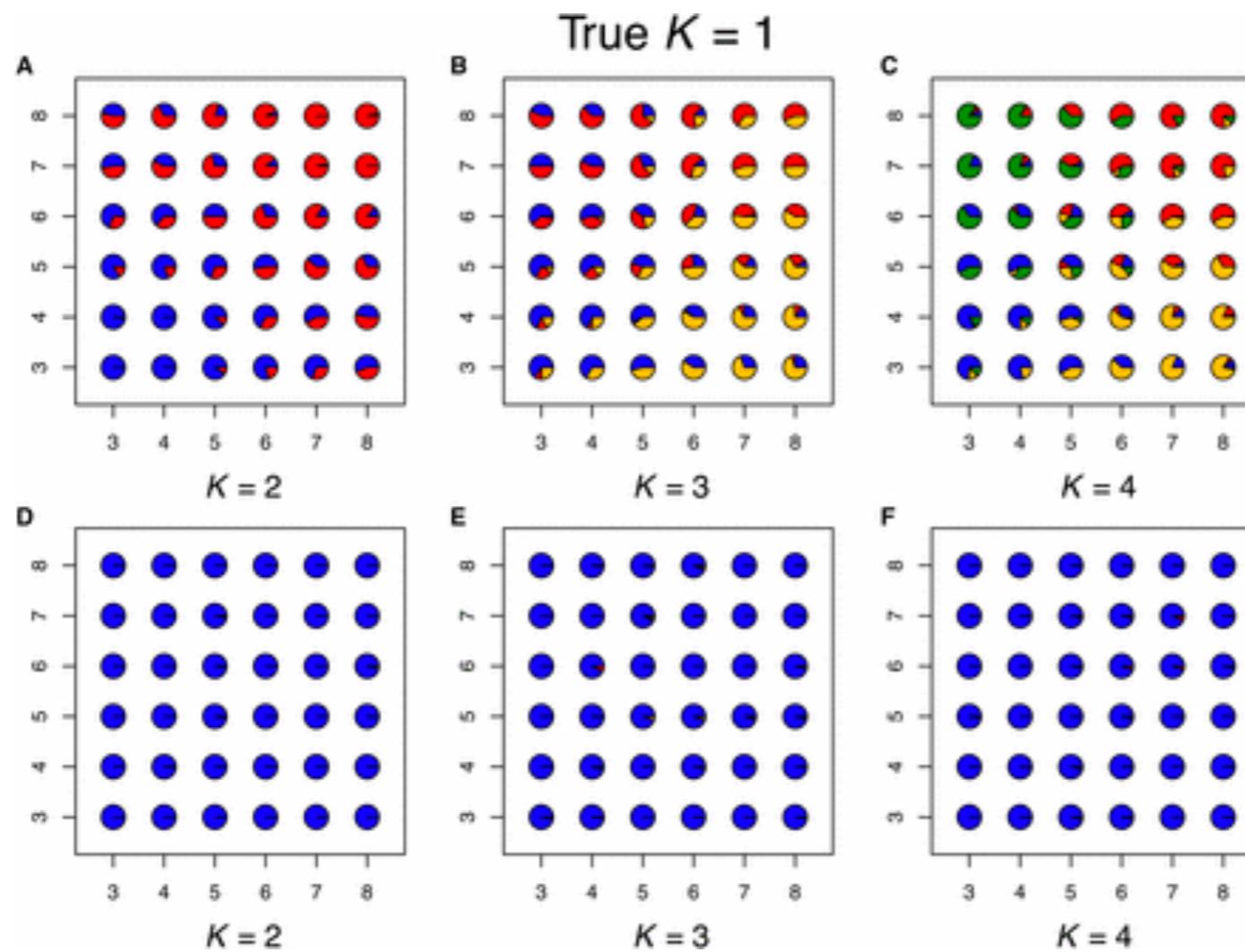
Rosenberg et al 2002



Inferences
depend on the
scale of analysis

STRUCTURE

Also, validity of HW assumptions
(random mating, large N_e , no gene flow)



Assuming random mating:, lots of “groups” that seem more similar to one another

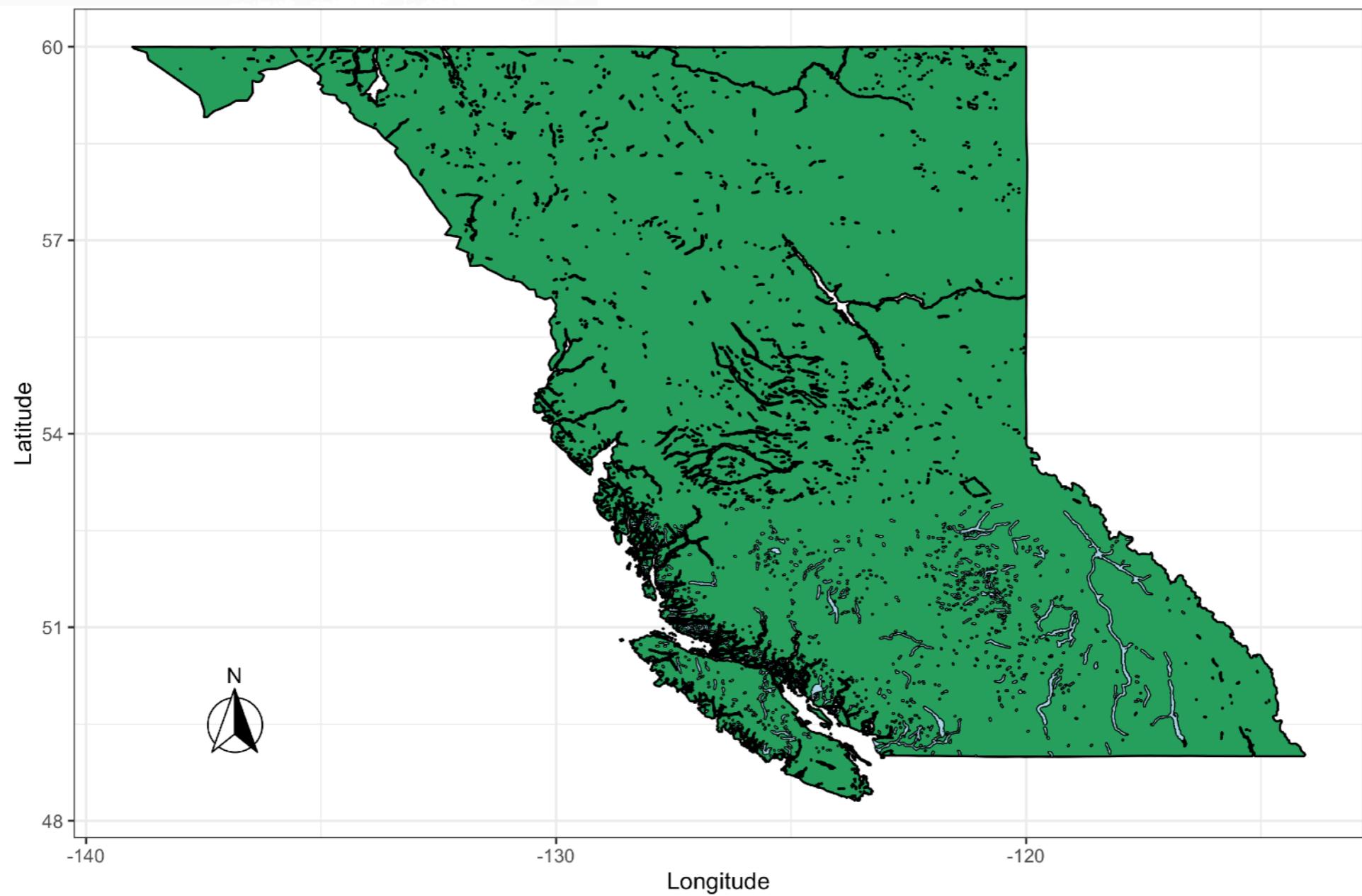
Accounting for non-random mating:
No groupings of excess covariance in allele frequencies

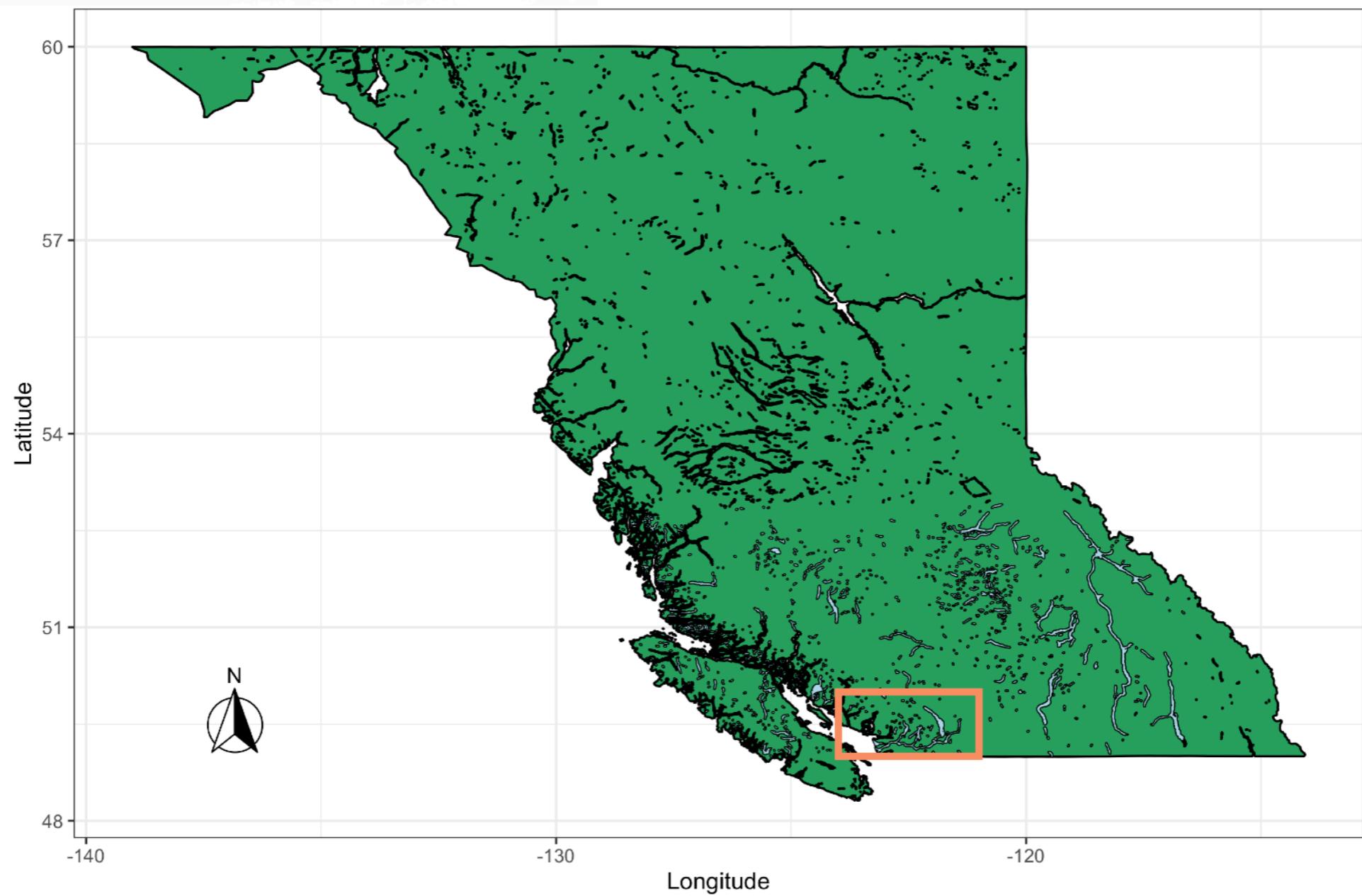
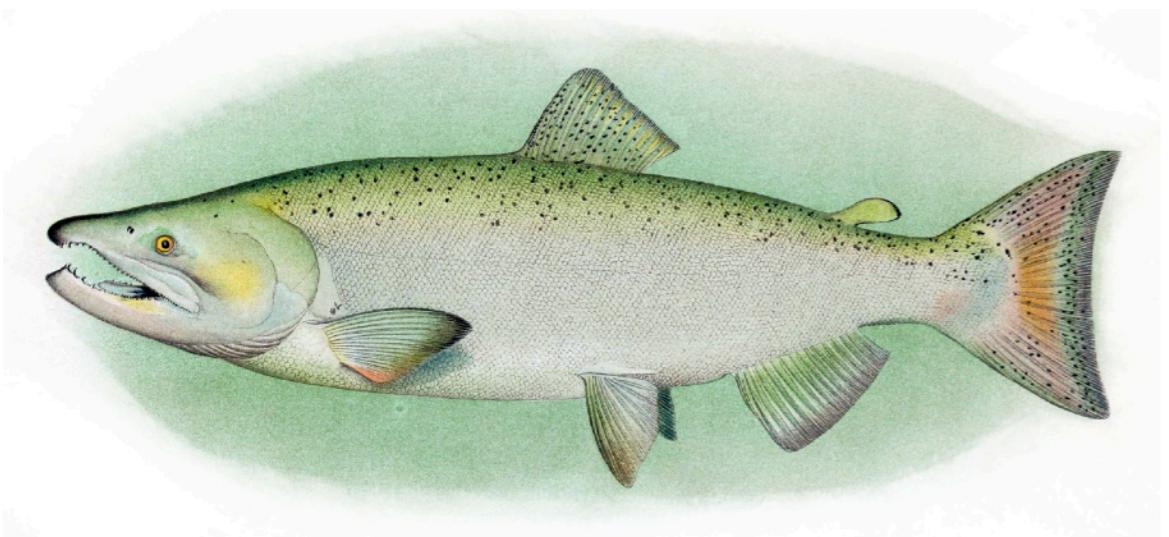
STRUCTURE programs

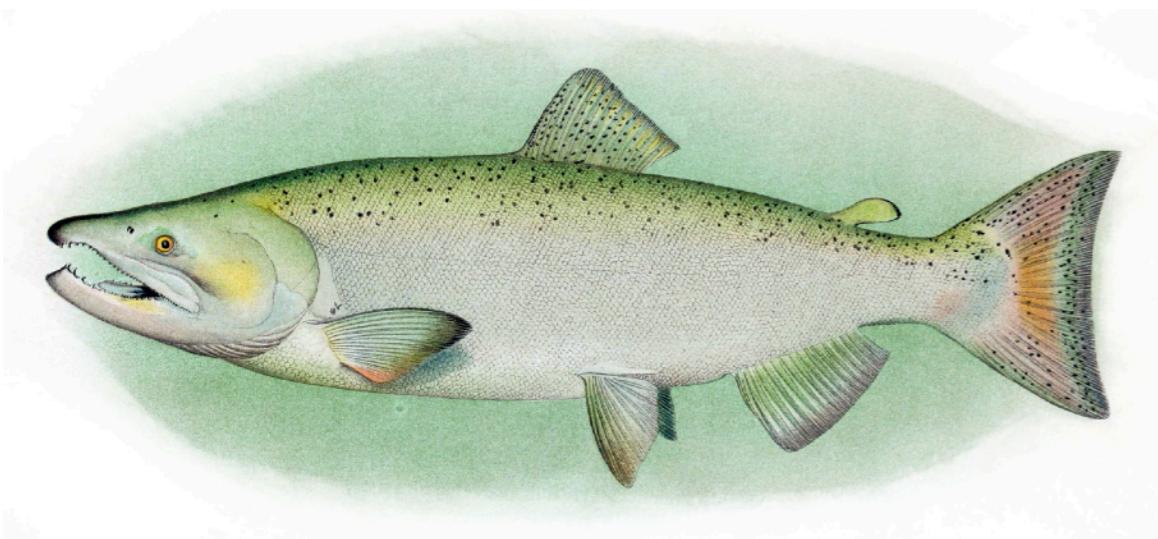
- STRUCTURE (original implementation, lots of model options, very slow & made for small genetic data sets)
- **Admixture** (fast - made for large genomic datasets)
- FASTstructure (fast - made for large genomic datasets)
- NGSadmix (ANGSD implementation, based on genotype likelihoods)

STRUCTURE

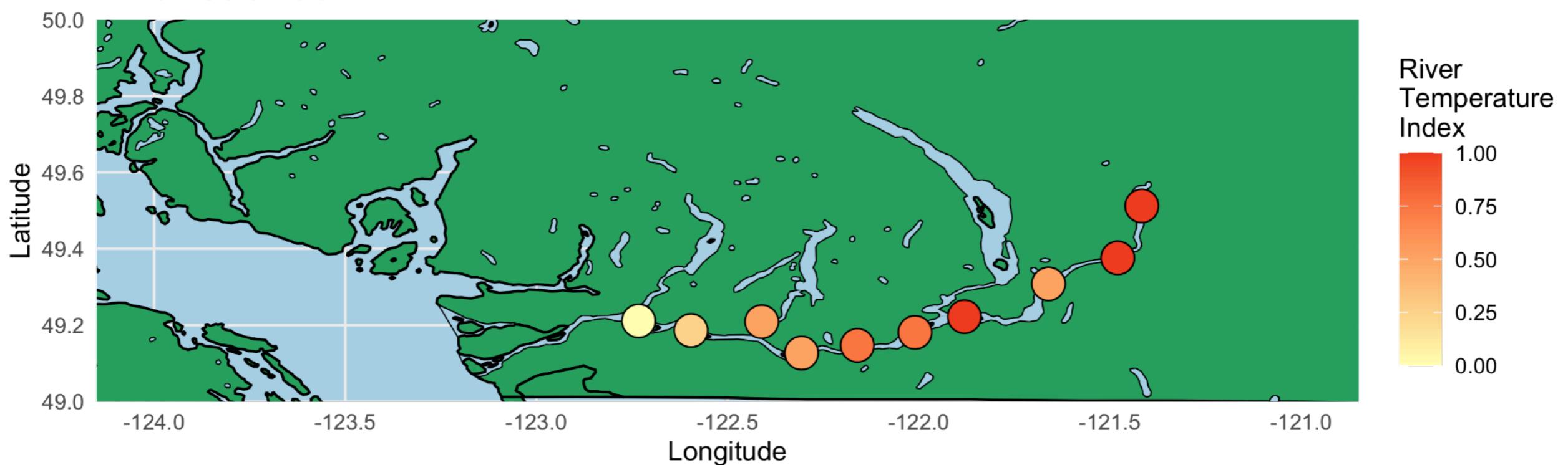
- You should prune dataset to unlinked SNPs
- Run multiple times to confirm consistency - “Cross validation”
- Admixture has a bootstrap option, to get CIs of admixture proportions

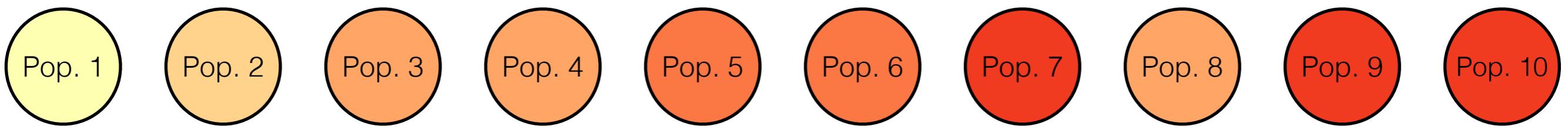
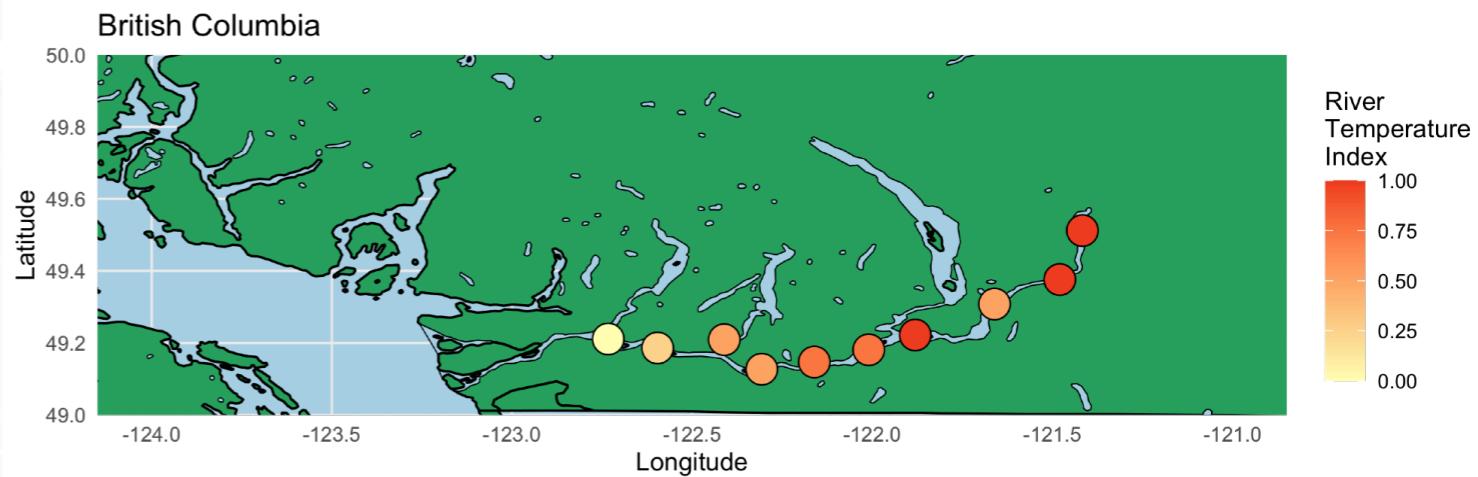


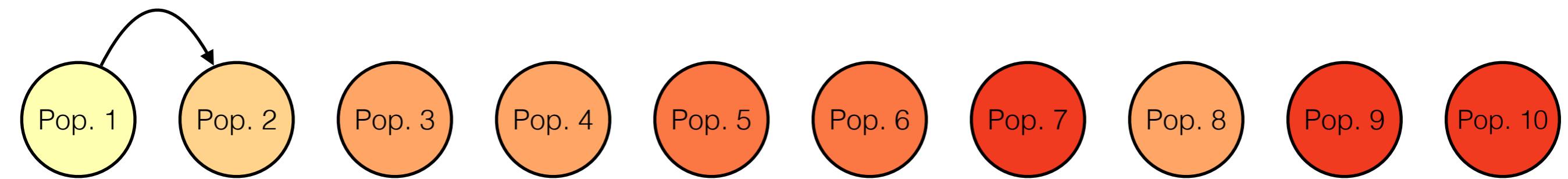
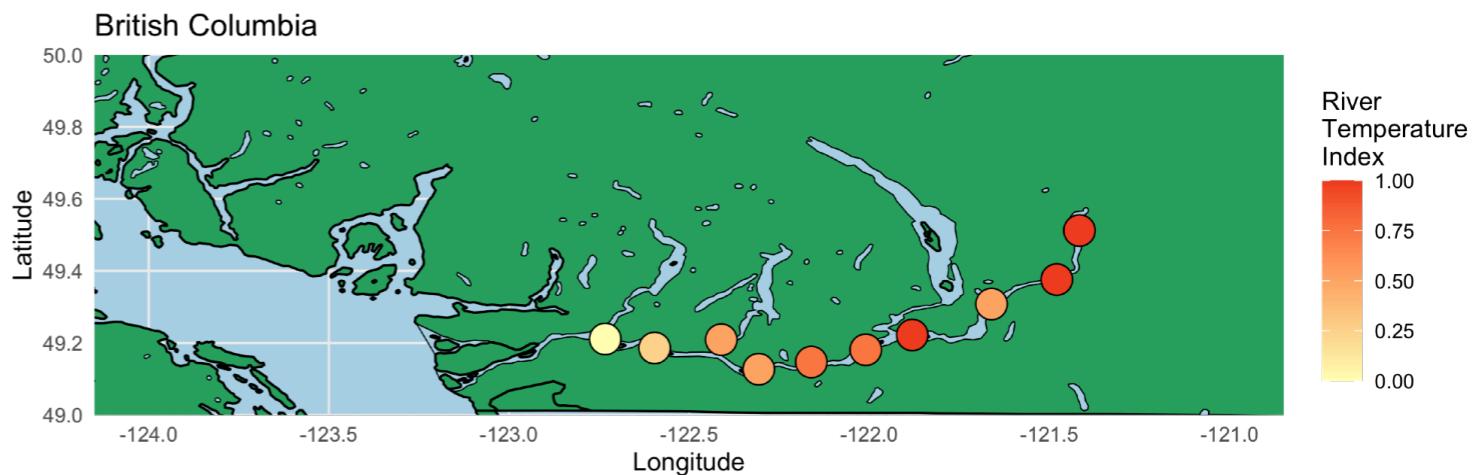
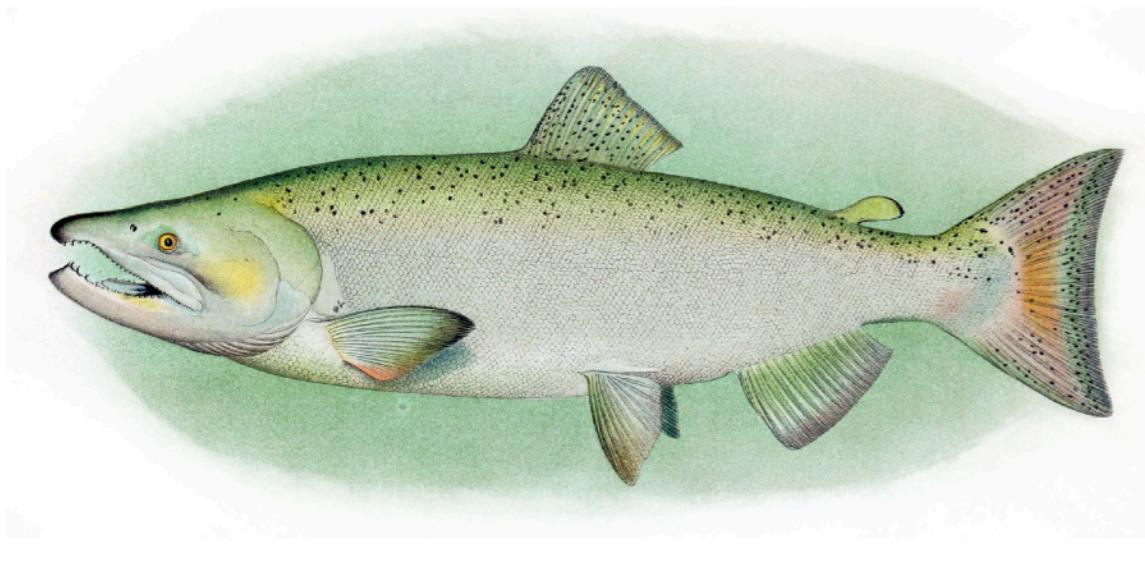


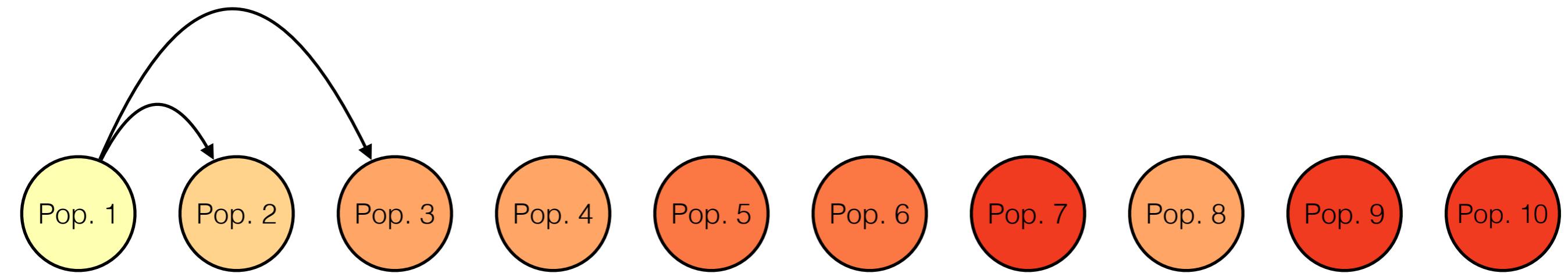
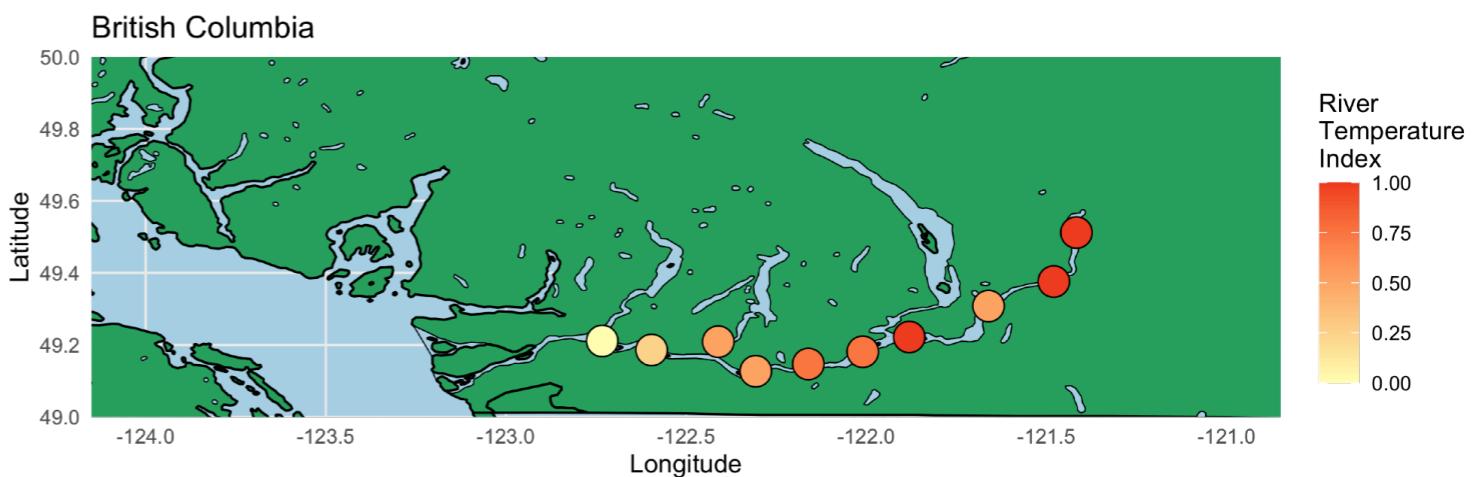
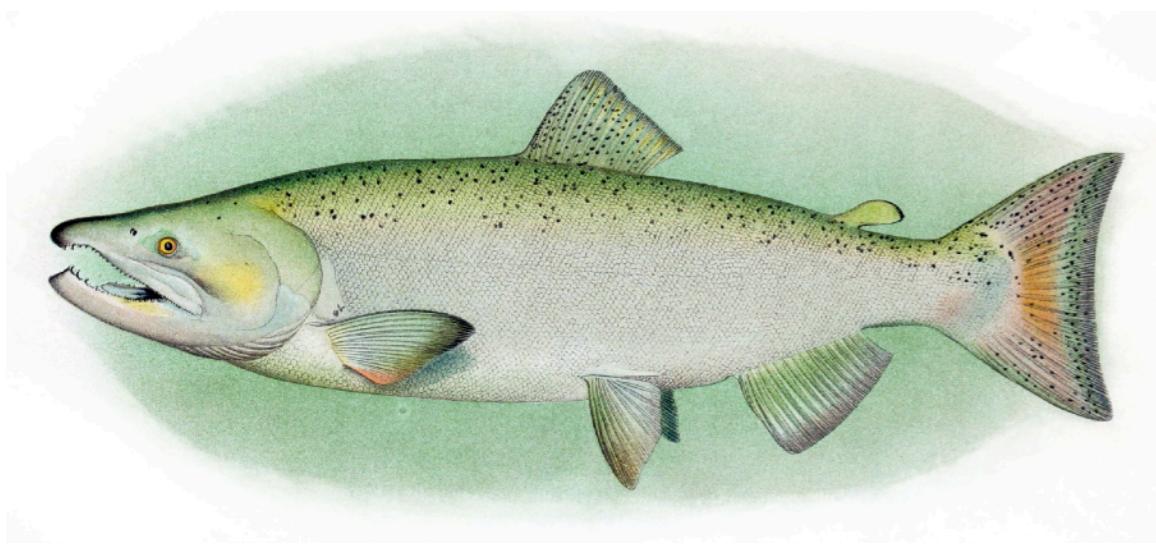


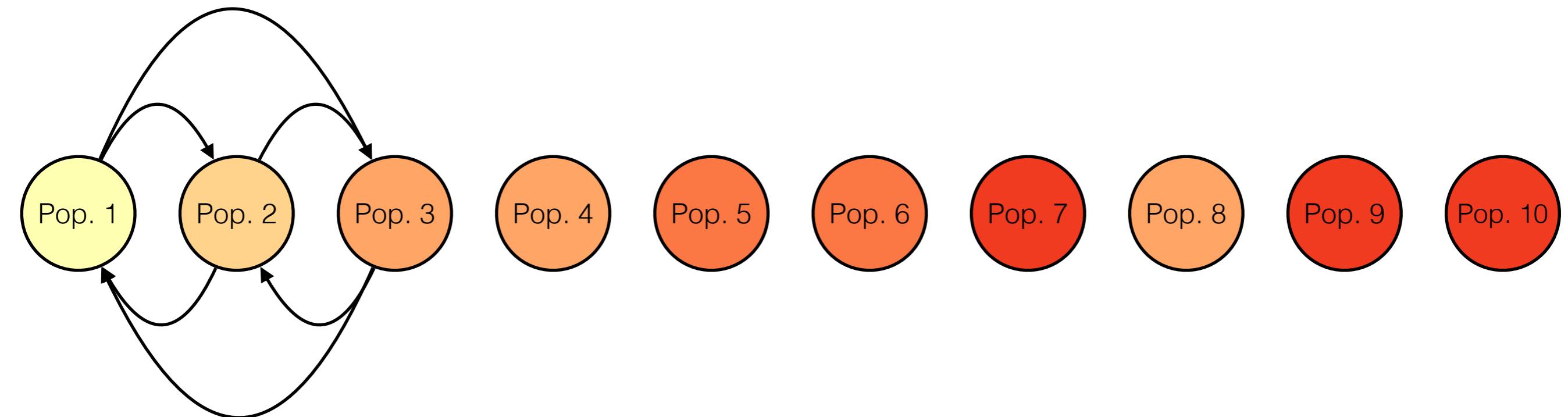
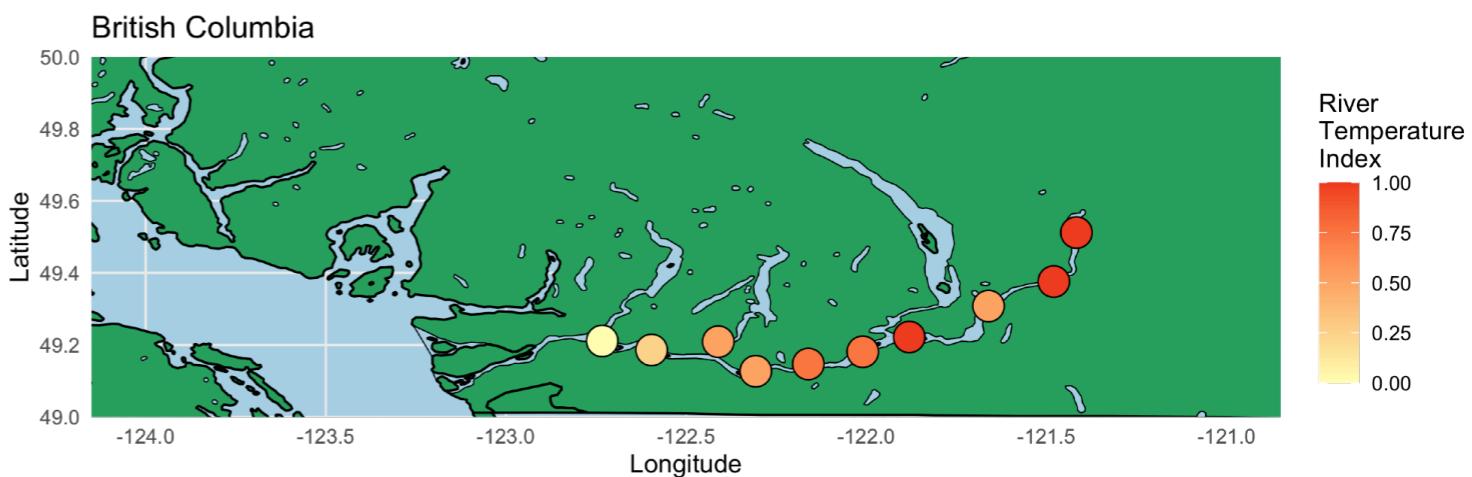
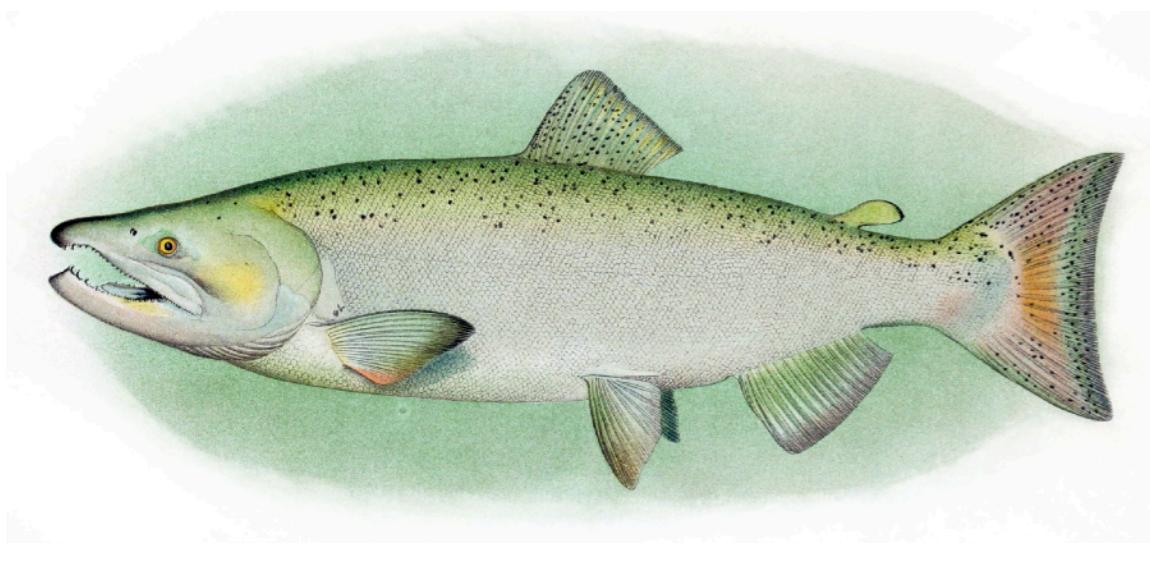
British Columbia

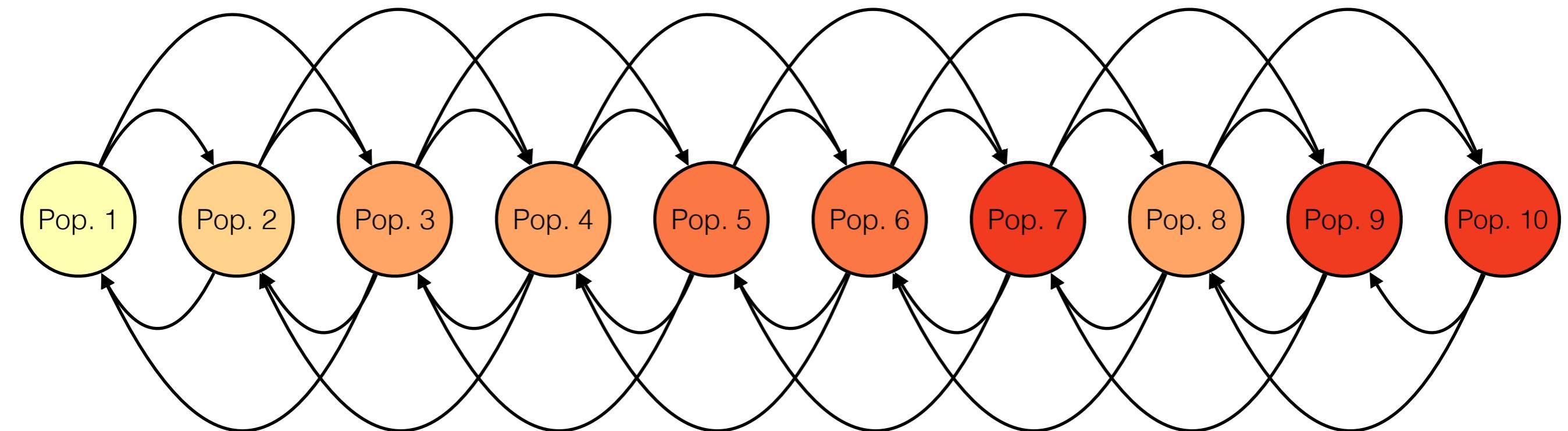
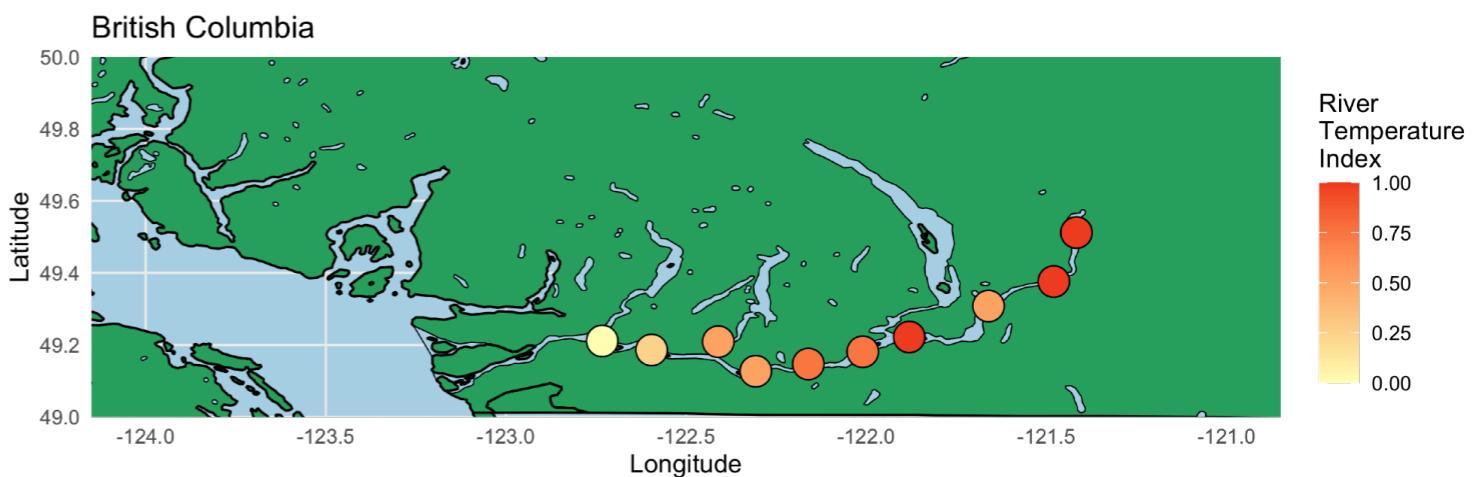
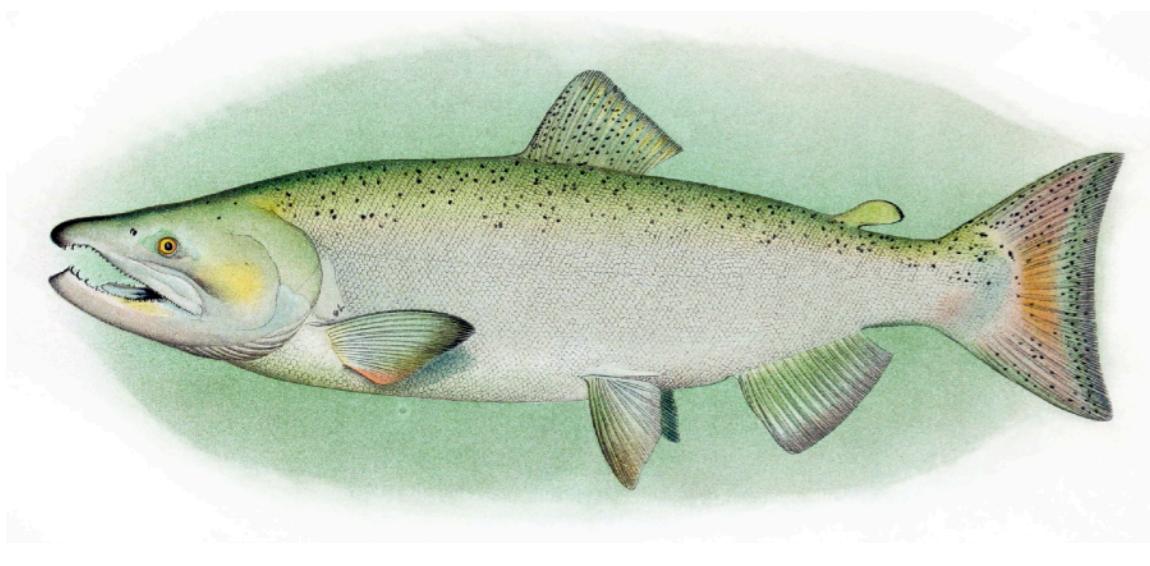


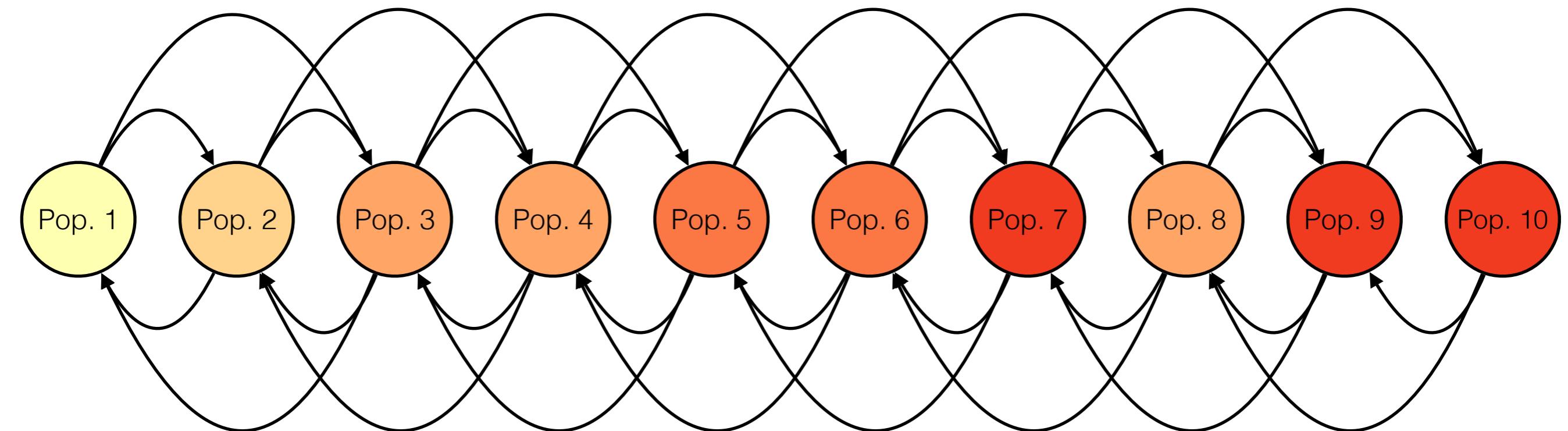
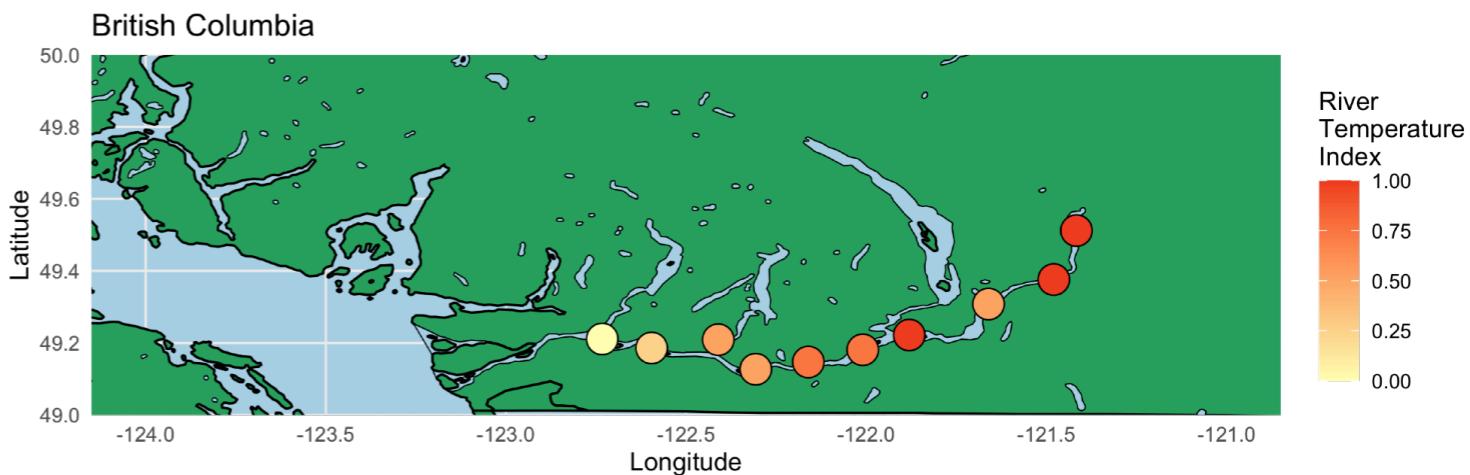
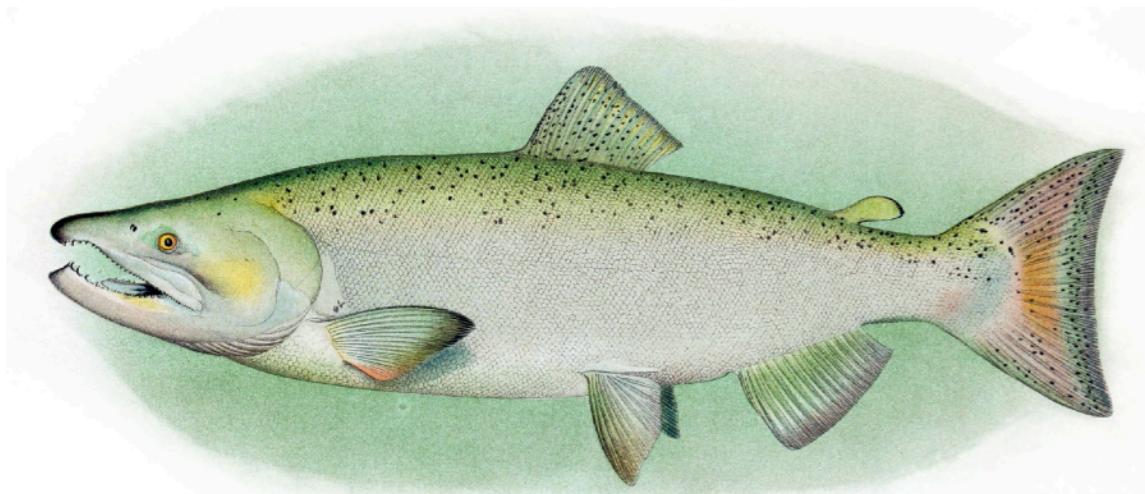










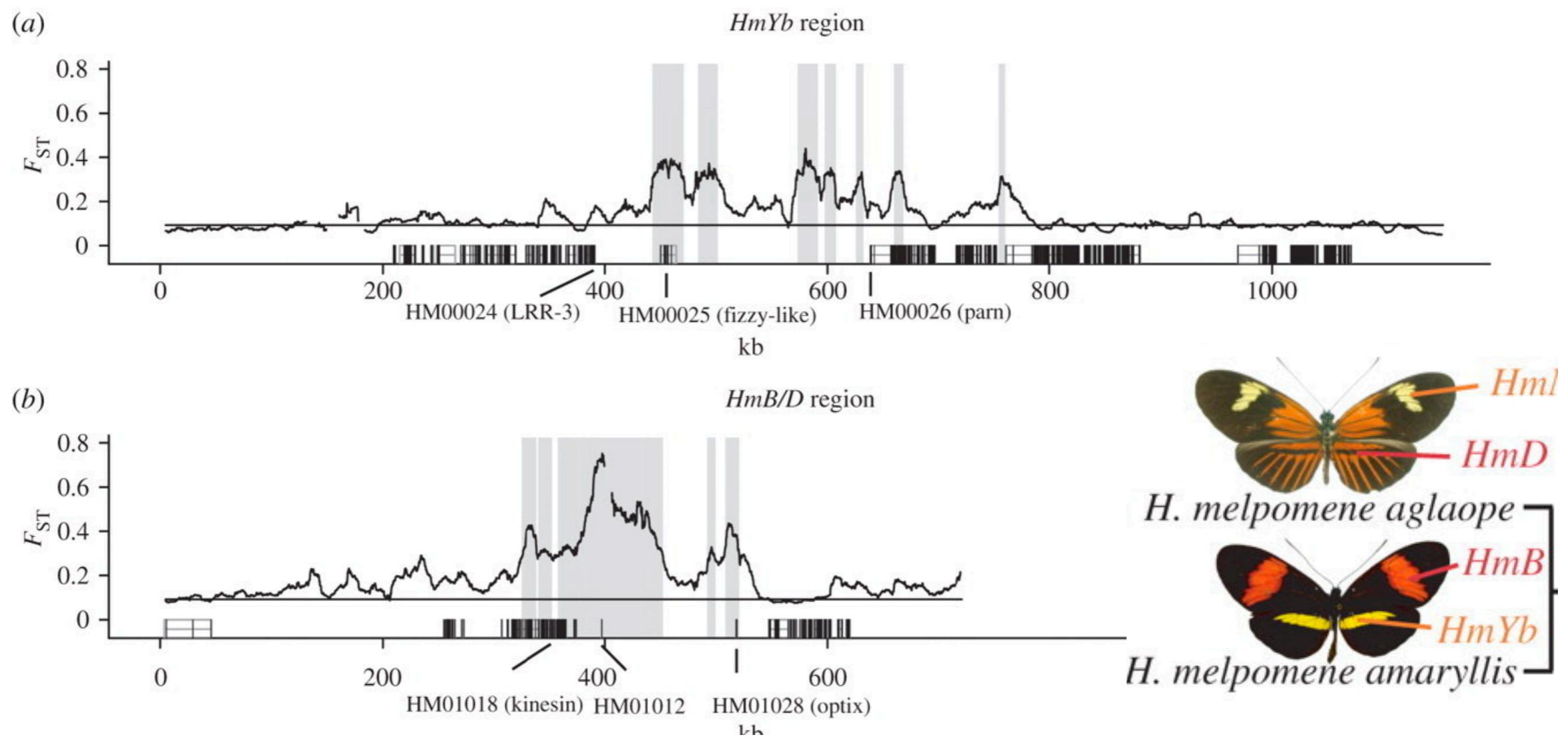


Given our knowledge of the system:
How would you examine patterns of population
structure? What would you look out for?

Population genomics of adaptation (and phenotypes)

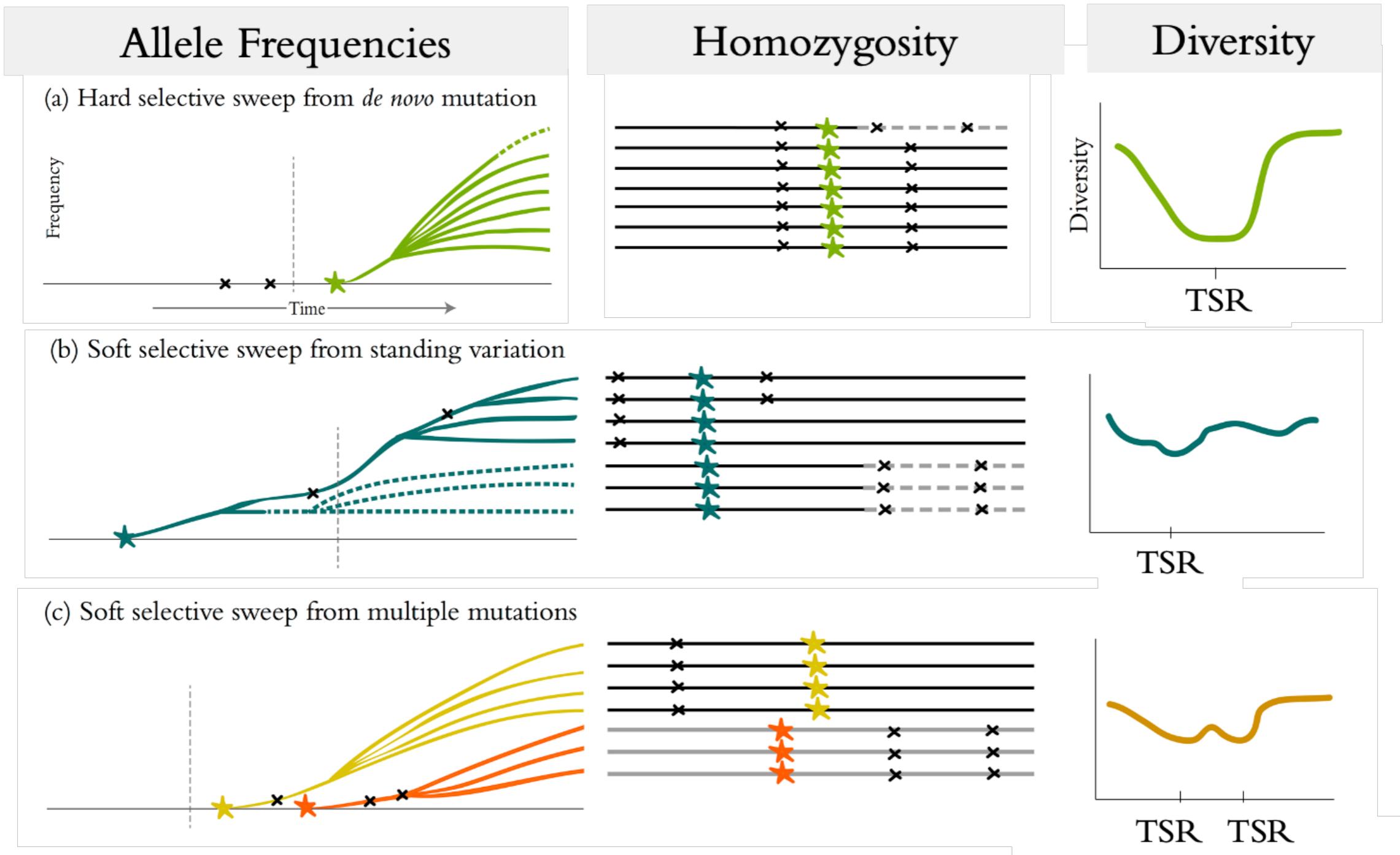
- Scans for selection
 - Fst revisited
 - Selective sweeps
- GWAS

F_{ST} : Scans for selection



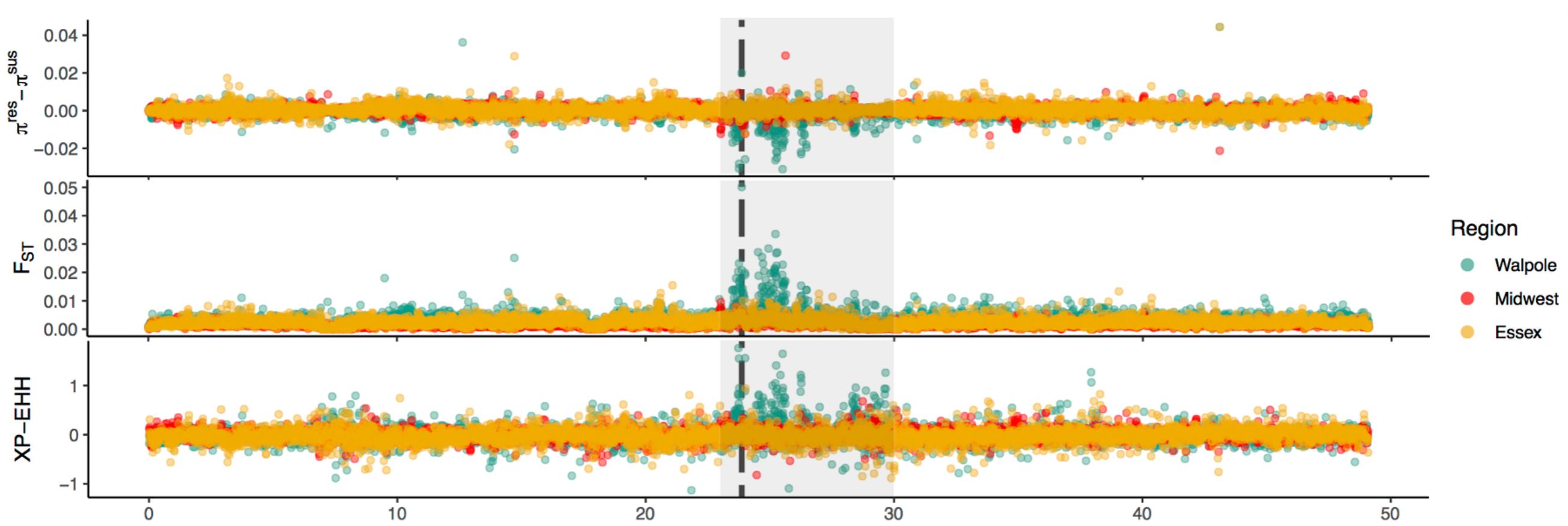
Nandeau et al., 2012

Scans for selection: Sweeps



Scans for selection: Sweeps

Herbicide resistance allele



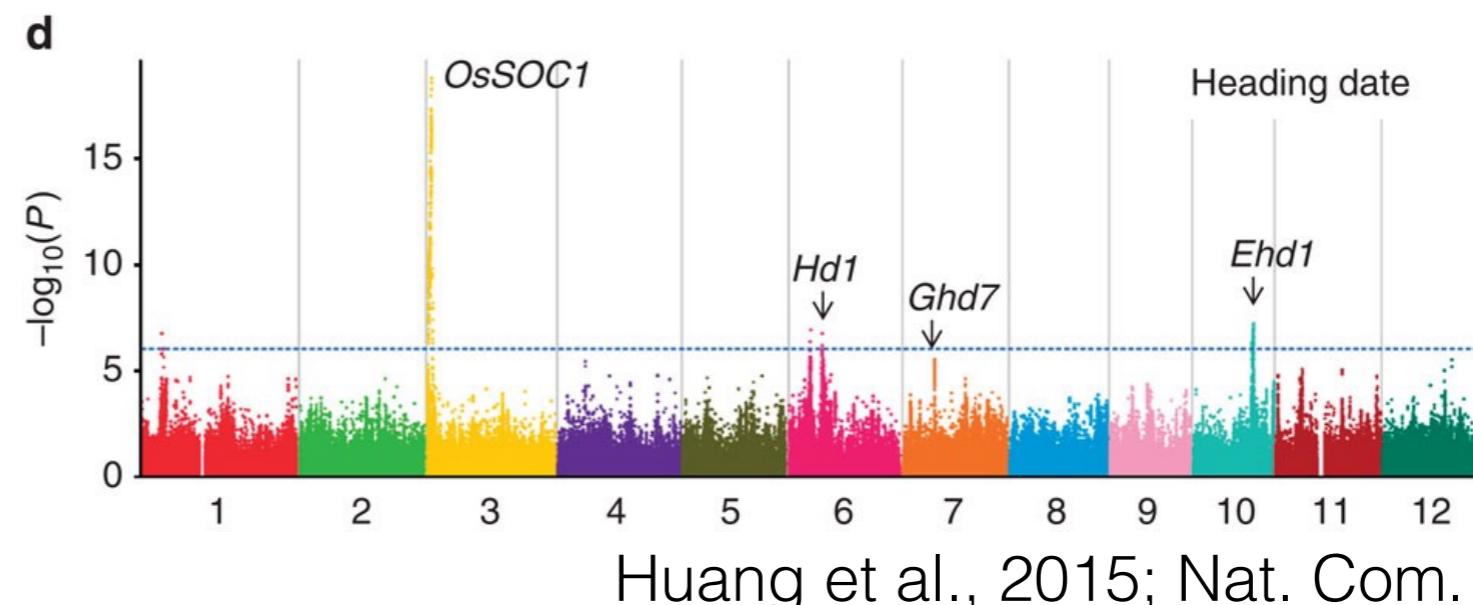
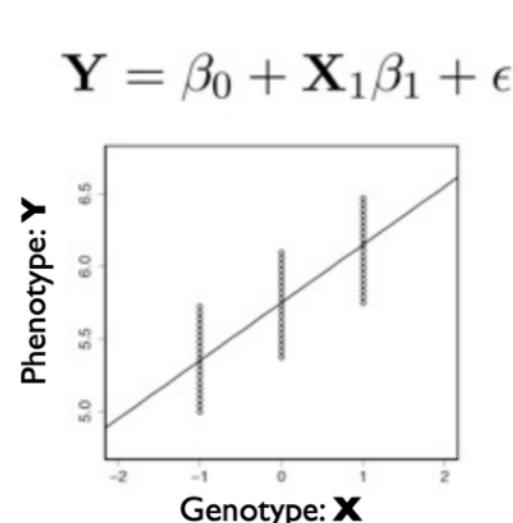
Loss of diversity, excess homozygosity, and allele frequency differentiation around locus in resistant haplotypes

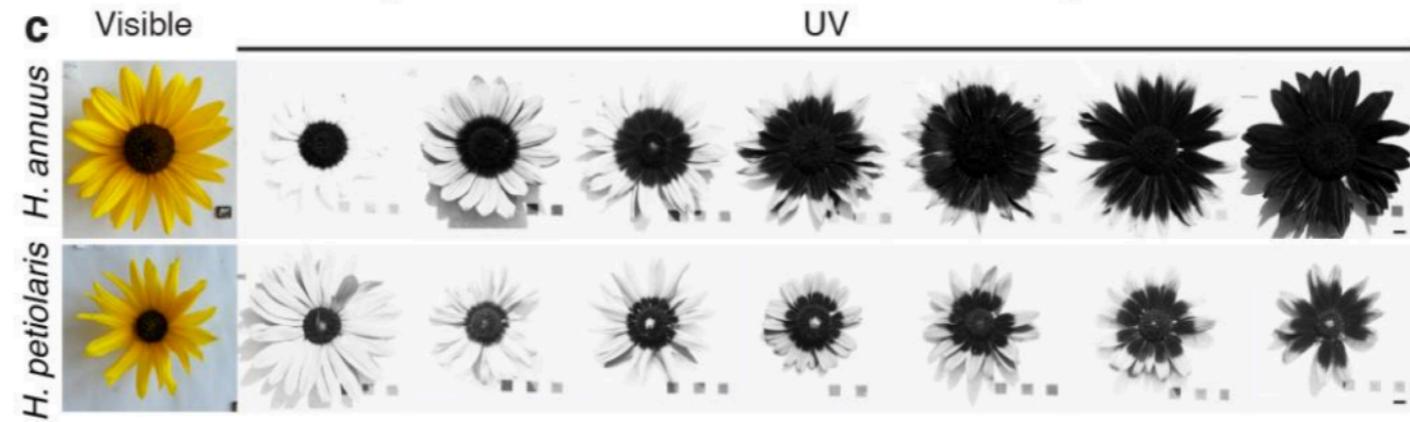
Population genomics of adaptation (and phenotypes)

- Scans for selection
 - Fst revisited
 - Selective sweeps
- GWAS

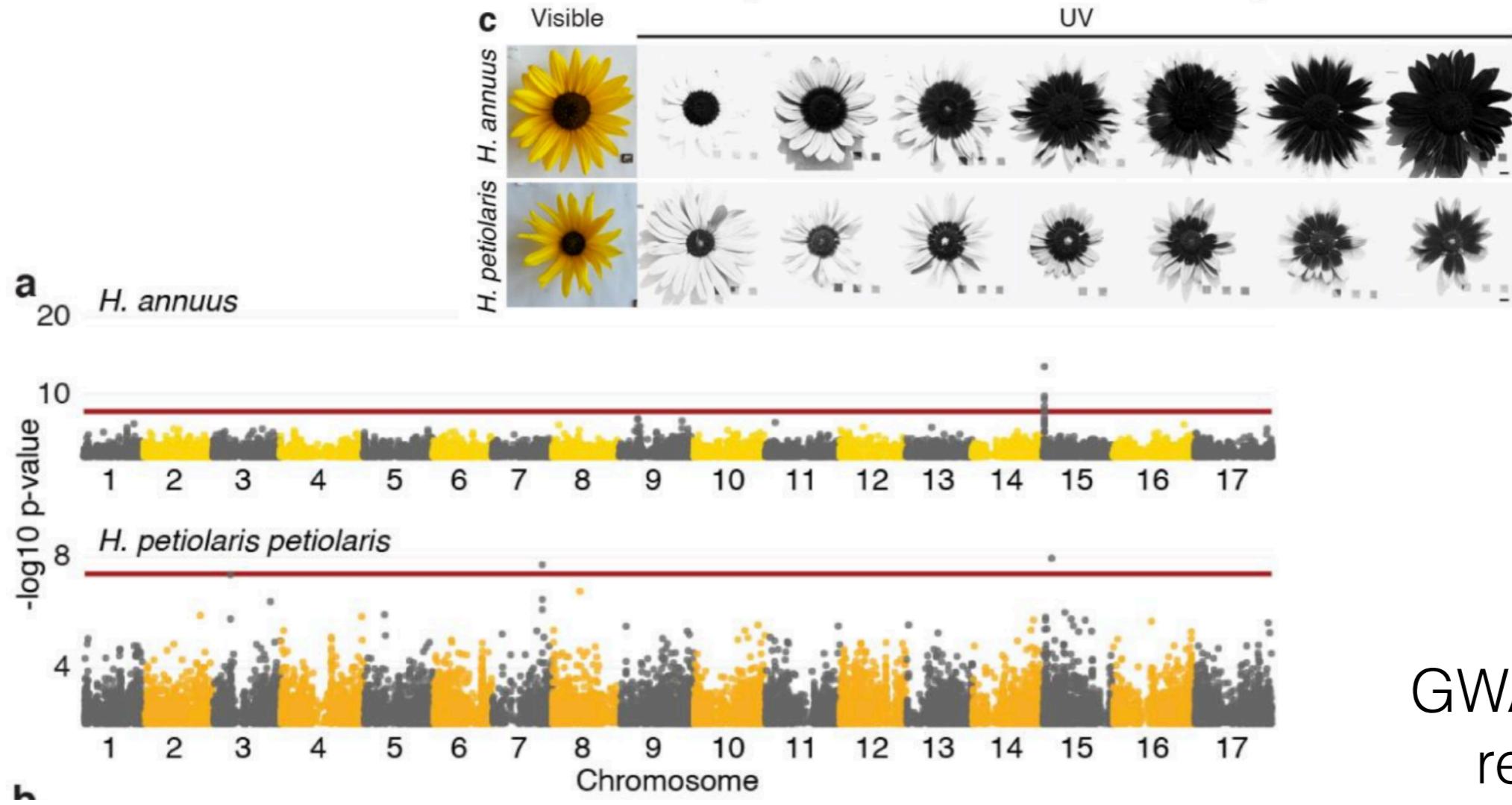
SNP-phenotype associations (GWAS): one allele at a time

- Regression of phenotype on SNP
- Yields an estimate of the association between SNP and phenotype beyond what would be expected due to population structure

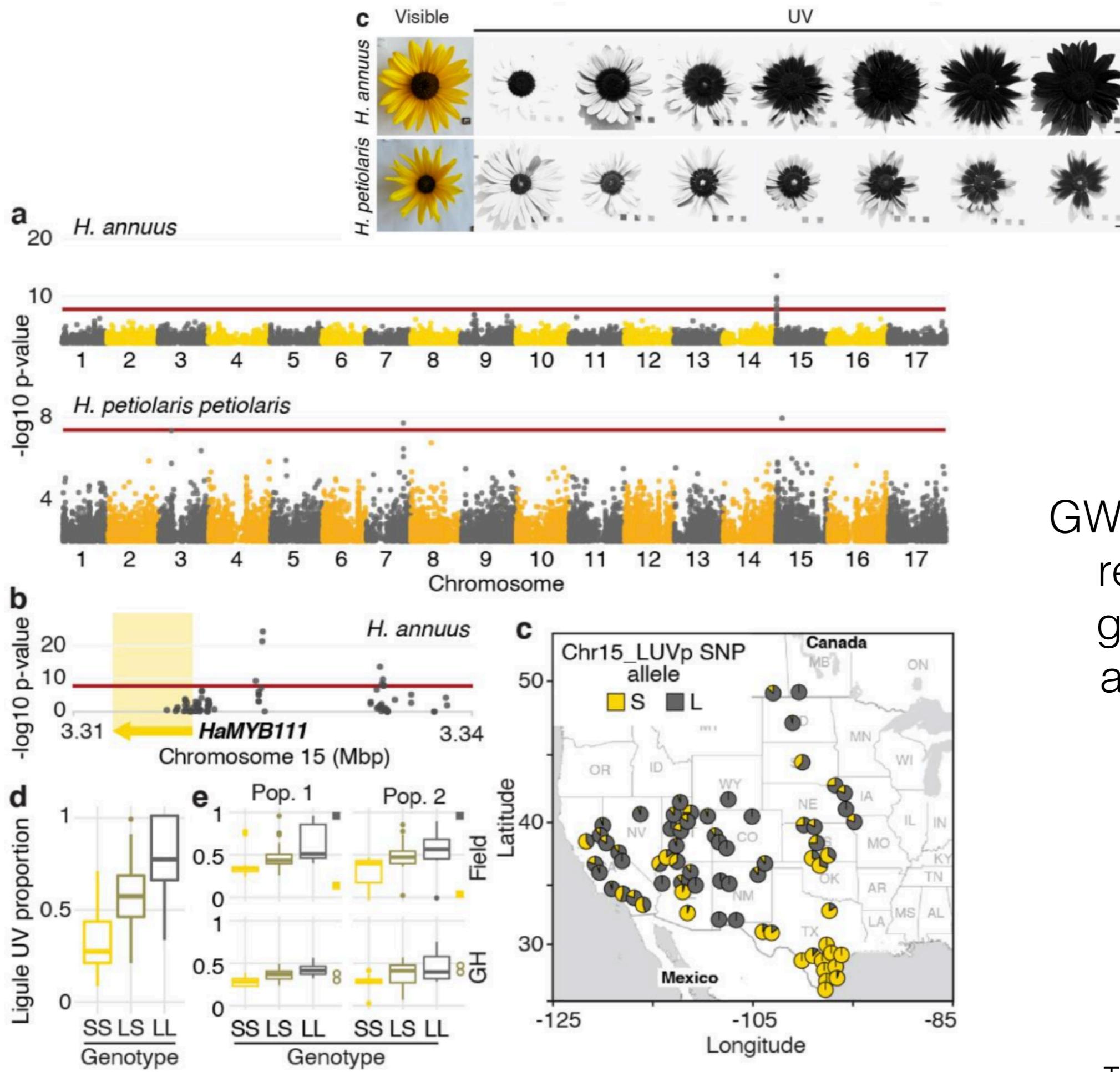




GWAS can help
resolve the
genetics of
adaptation



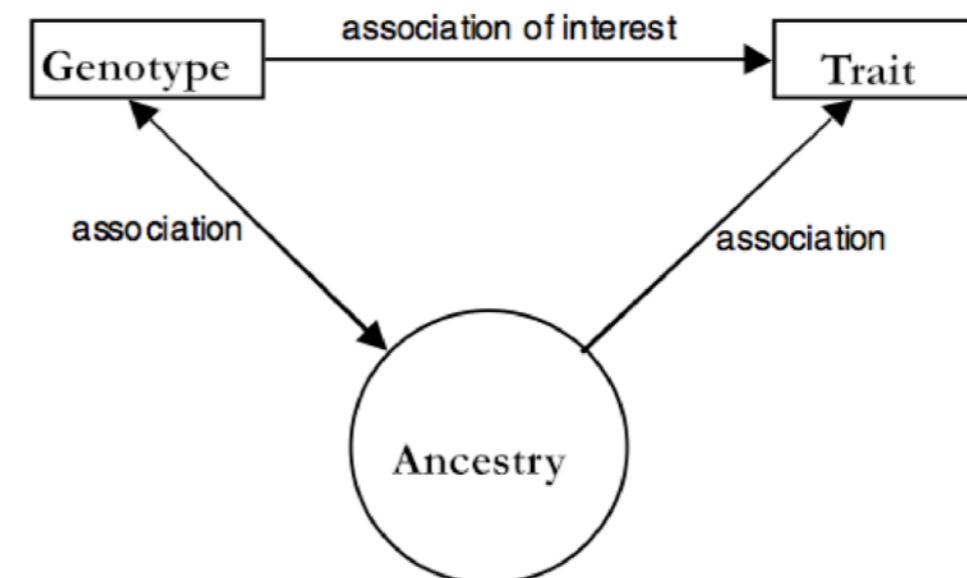
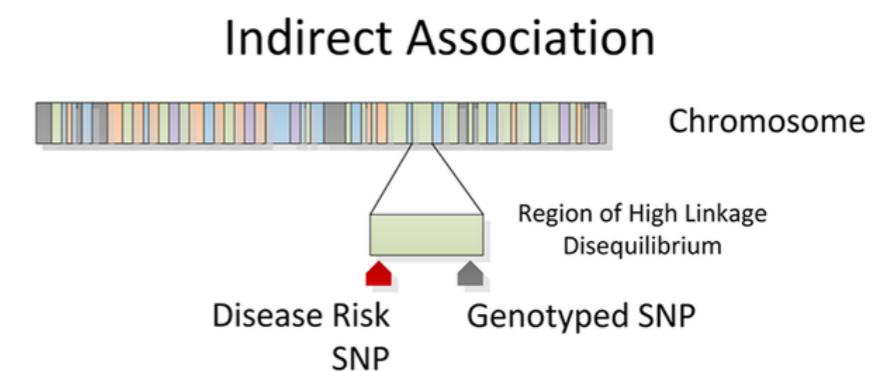
GWAS can help resolve the genetics of adaptation



GWAS can help resolve the genetics of adaptation

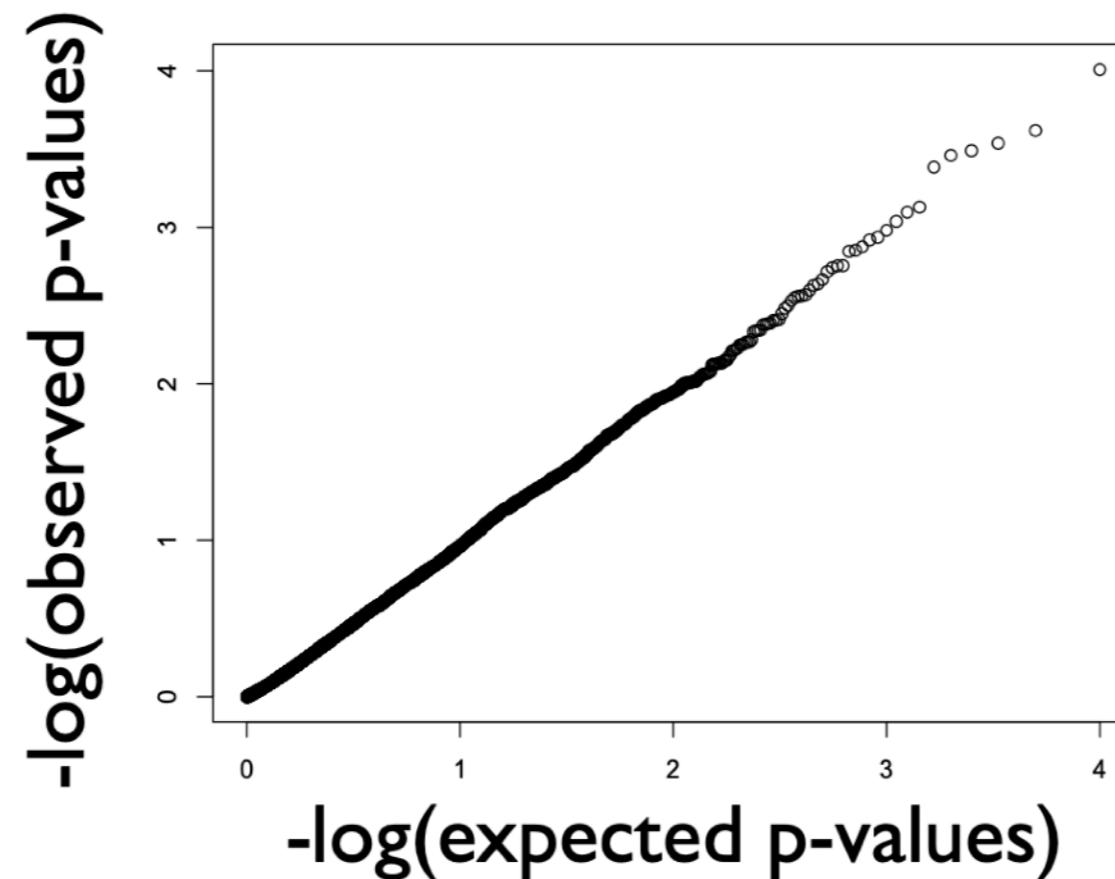
GWAS issues:

- Correlation != Causation
- Population structure
- Multiple testing



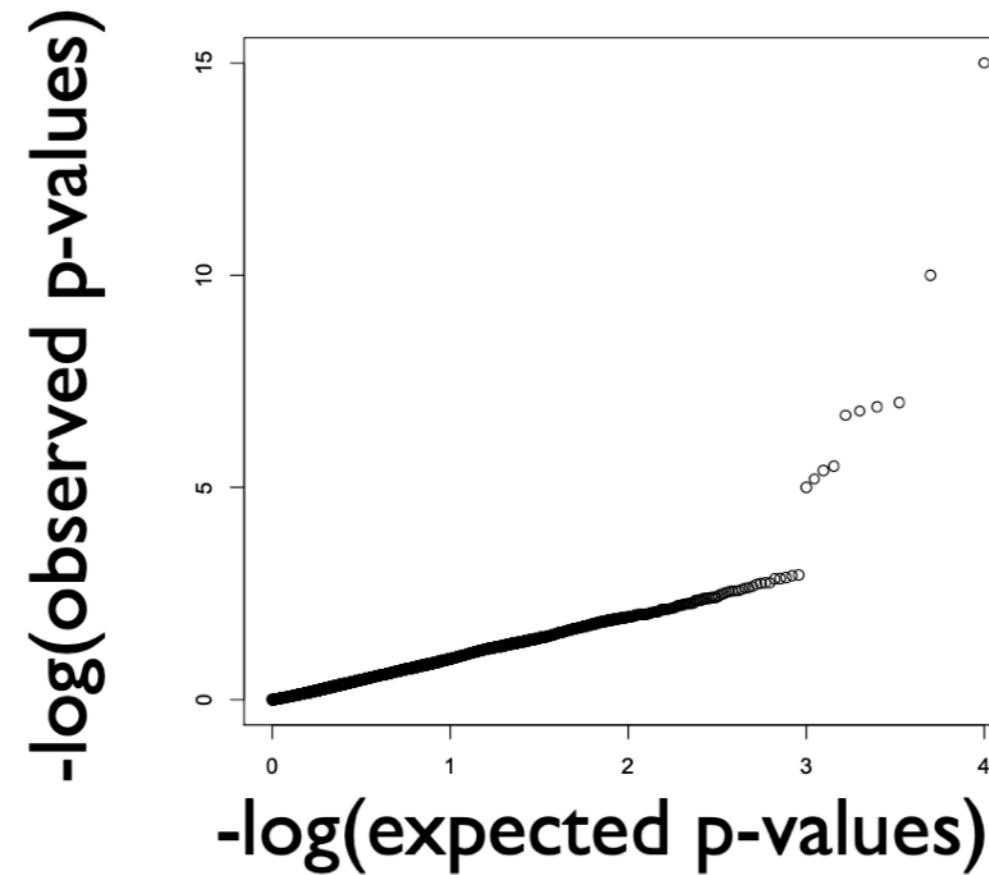
Evaluating your statistical GWA model: QQplots

- In an ideal GWAS case where there ARE NO causal polymorphisms, your QQ plot will be a line:



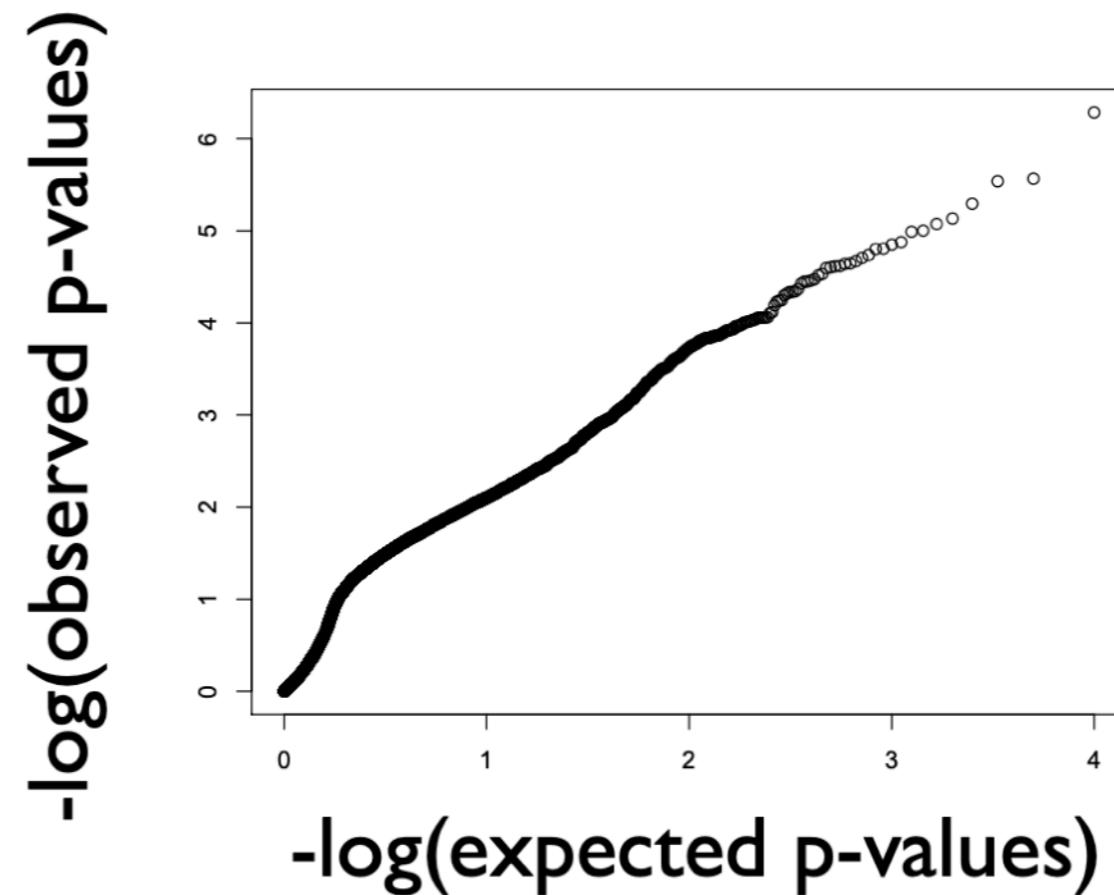
Evaluating your statistical GWA model: QQplots

- In an ideal GWAS case where there ARE causal polymorphisms, your QQ plot will be a line with a tail (!!):



Evaluating your statistical GWA model: QQplots

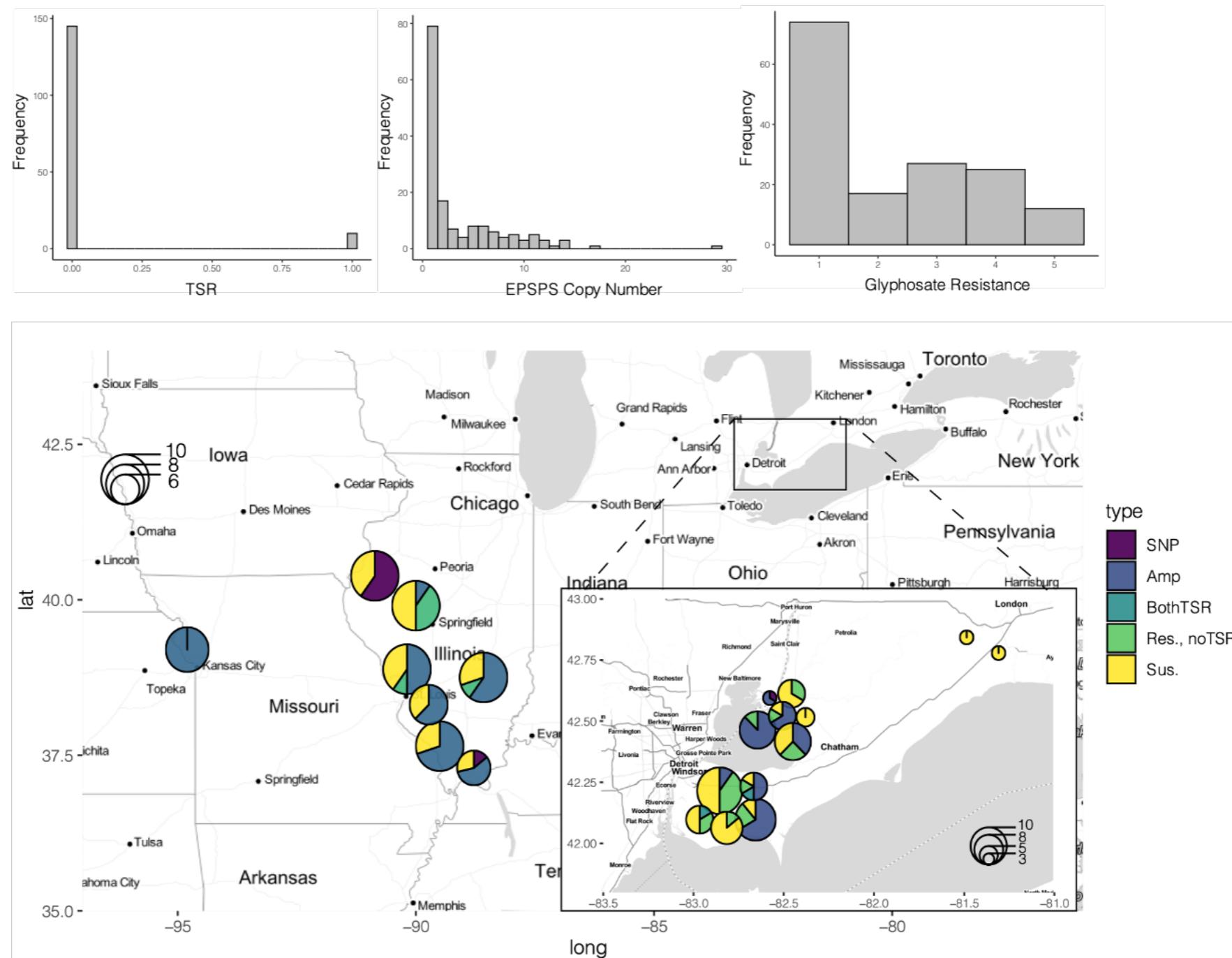
- In practice, you can find your QQ plot looks different than either the “null GWAS” case or the “ideal GWAS” case, for example:



What might be driving this?

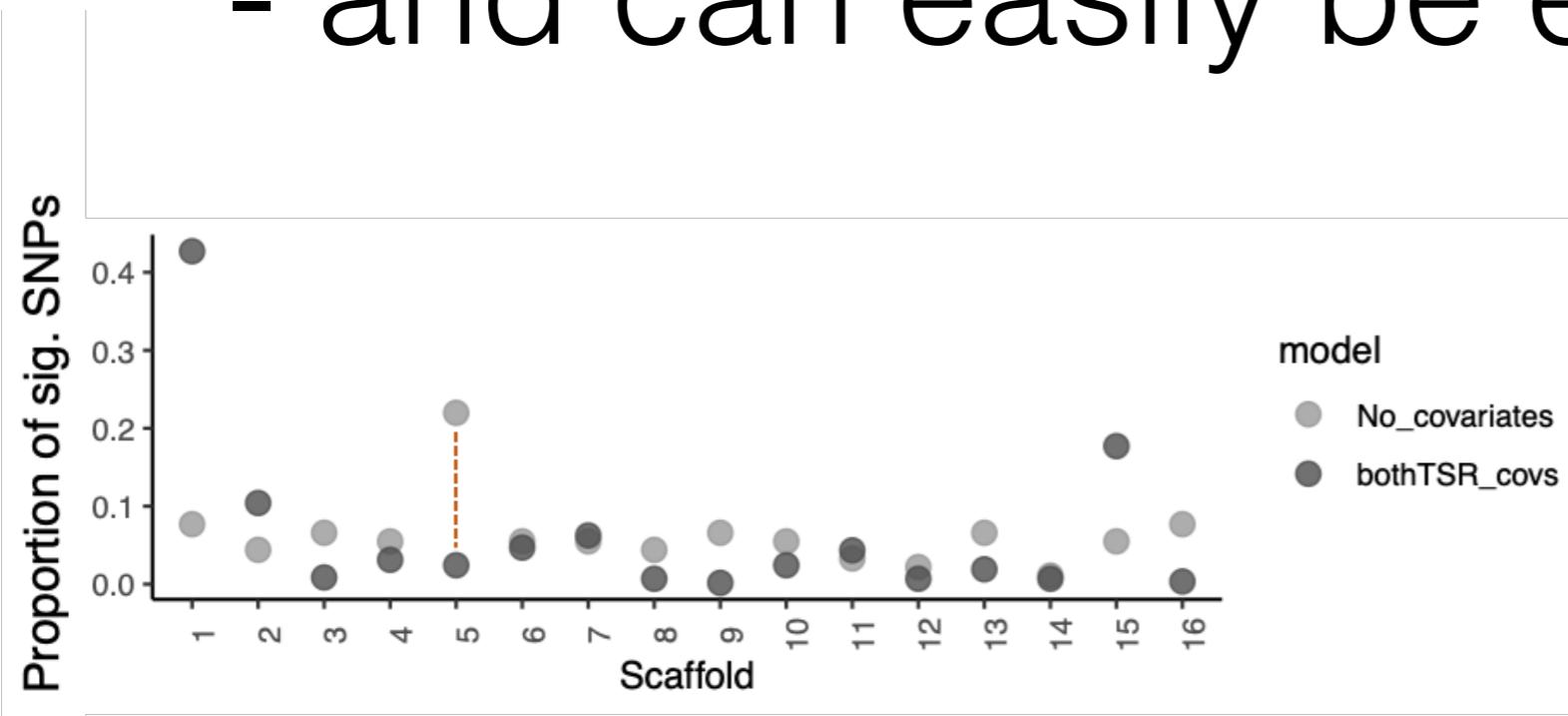
GWAS is a simple linear regression

- and can easily be extended

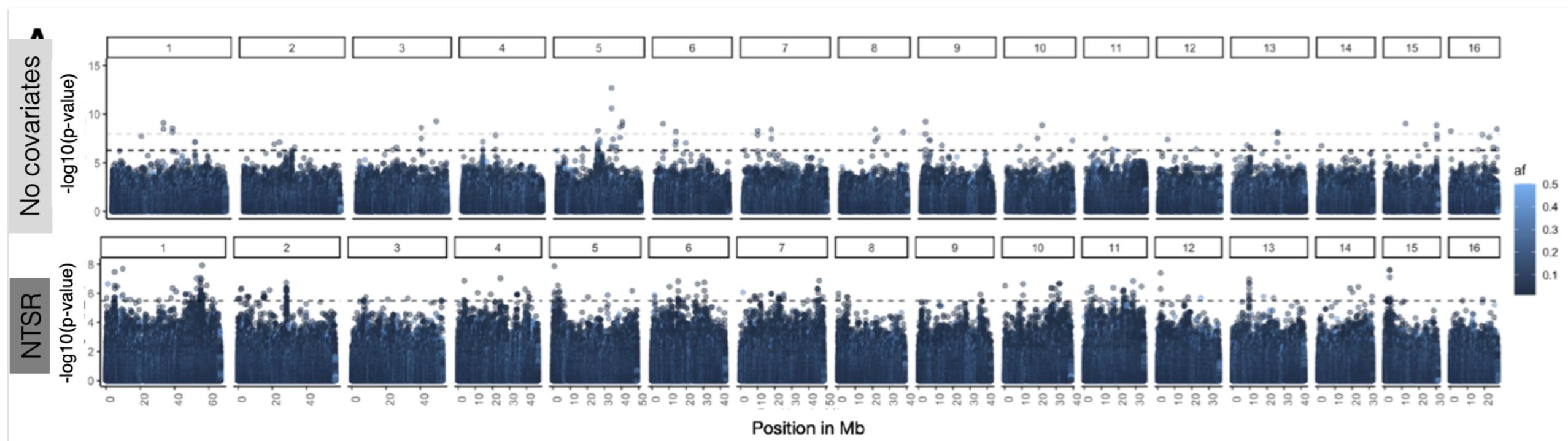


GWAS is a simple linear regression

- and can easily be extended



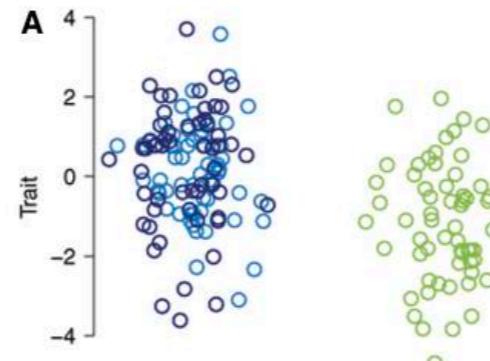
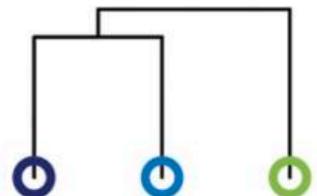
- i.e. Multiple regression
- Can also incorporate interaction effects etc.



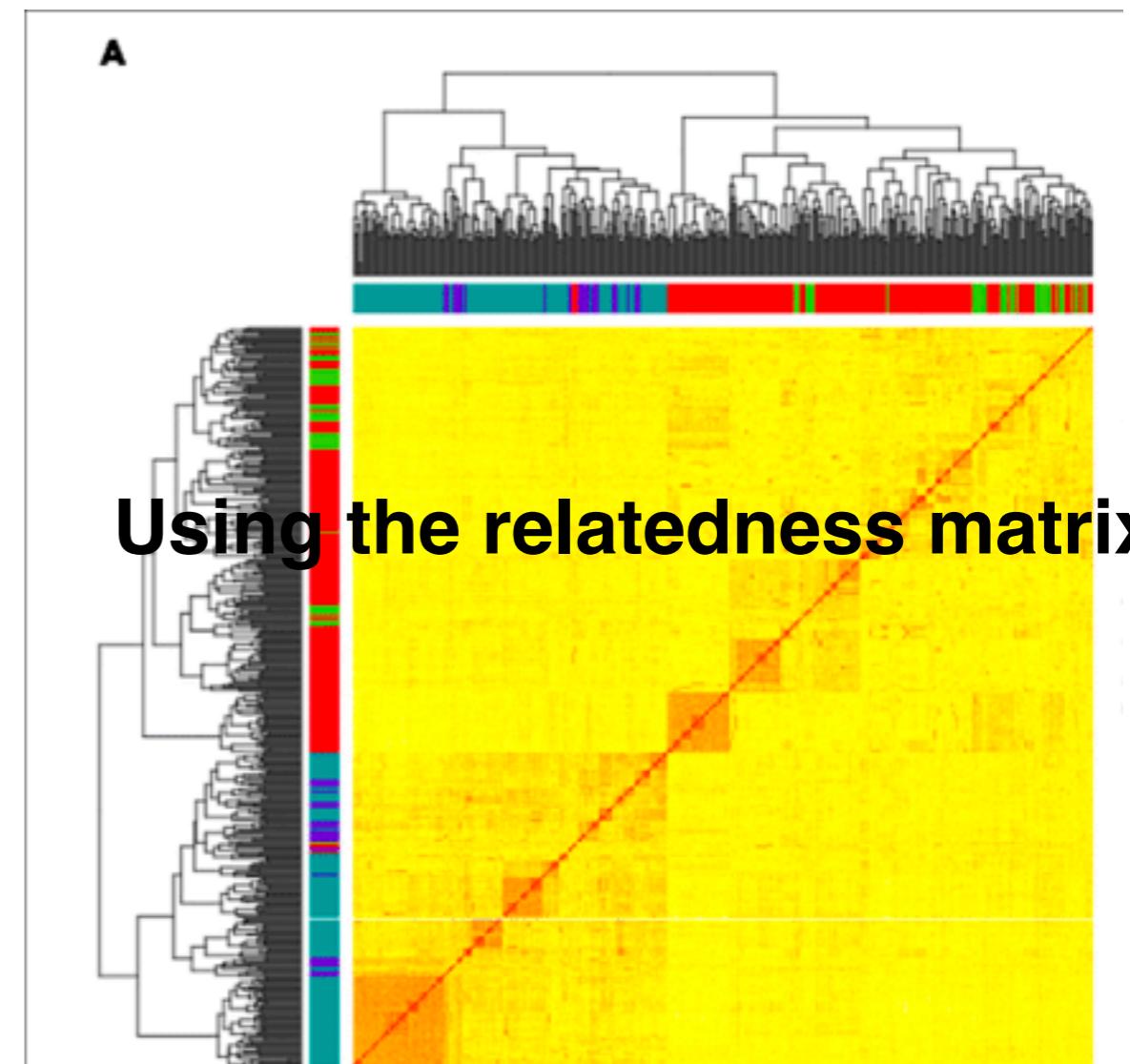
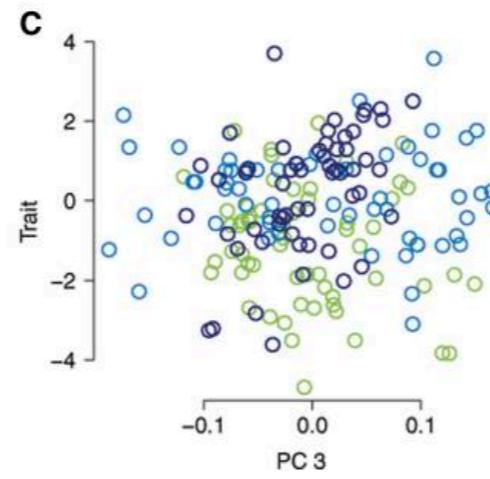
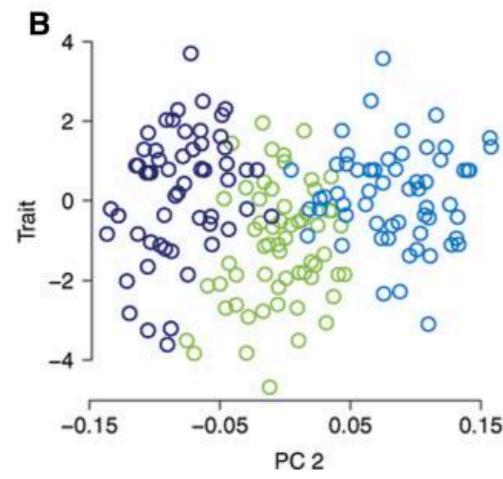
GWAS is a simple linear regression

- and can easily be extended

... to control for population structure!



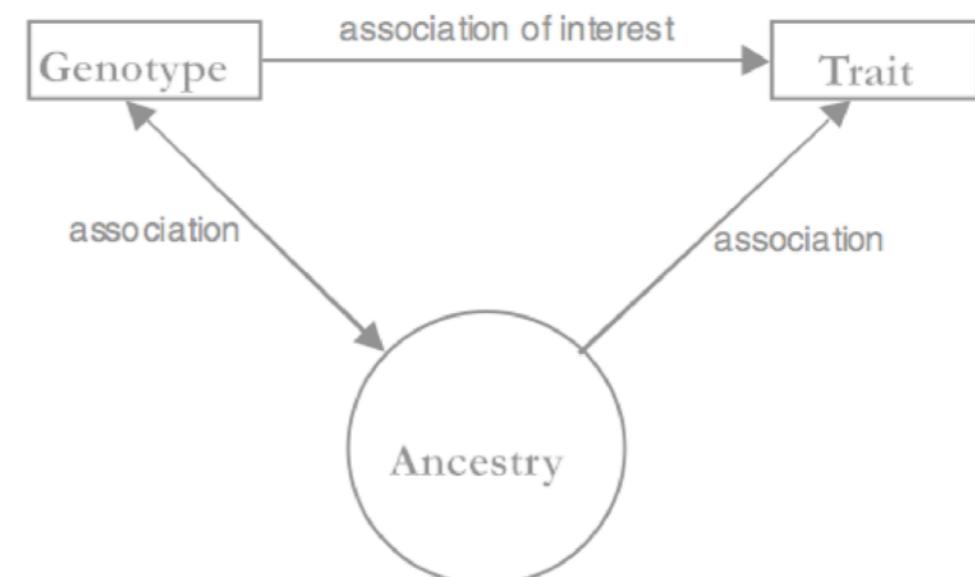
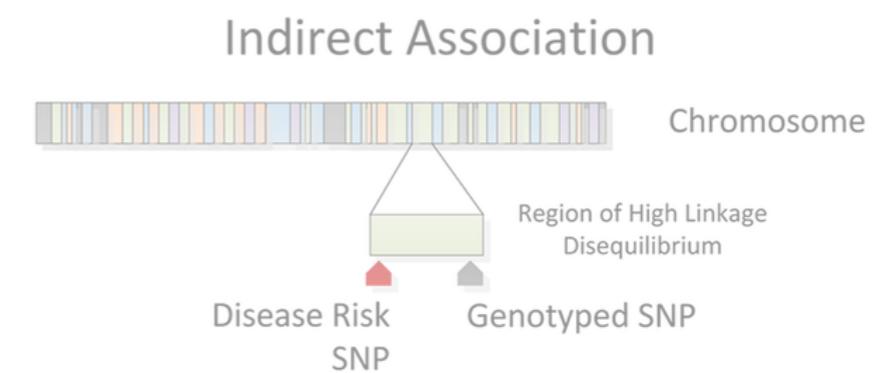
Using PCs as covariates



Using the relatedness matrix

GWAS issues:

- Correlation != Causation
- Population structure
- **Multiple testing**



Multiple test correction: Bonferroni

- A Bonferroni correction sets the Type I error for the entire set of N tests using the following approach: for a desired type I error α set the Bonferroni Type I error α_B to the following:

$$\alpha_B = \frac{\alpha}{N}$$

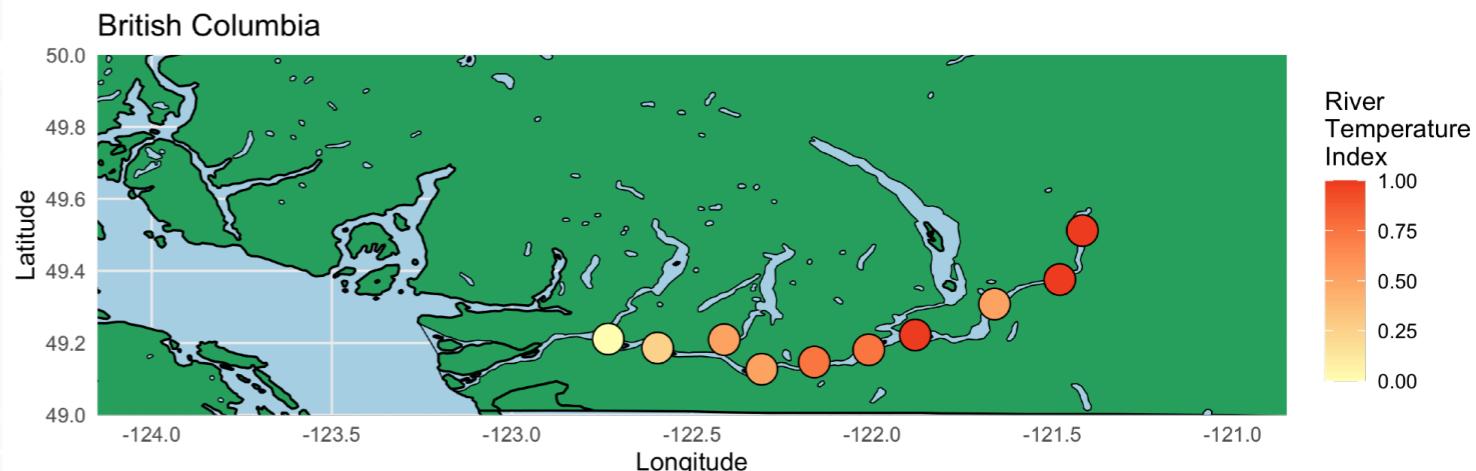
- We therefore use the Bonferroni Type I error to assess EACH of our N tests
- For example, if we have $N=100$ and we want an overall Type I error of 0.05, we require a test to have a p-value less than 0.0005 to be considered significant

Multiple test correction: FDR

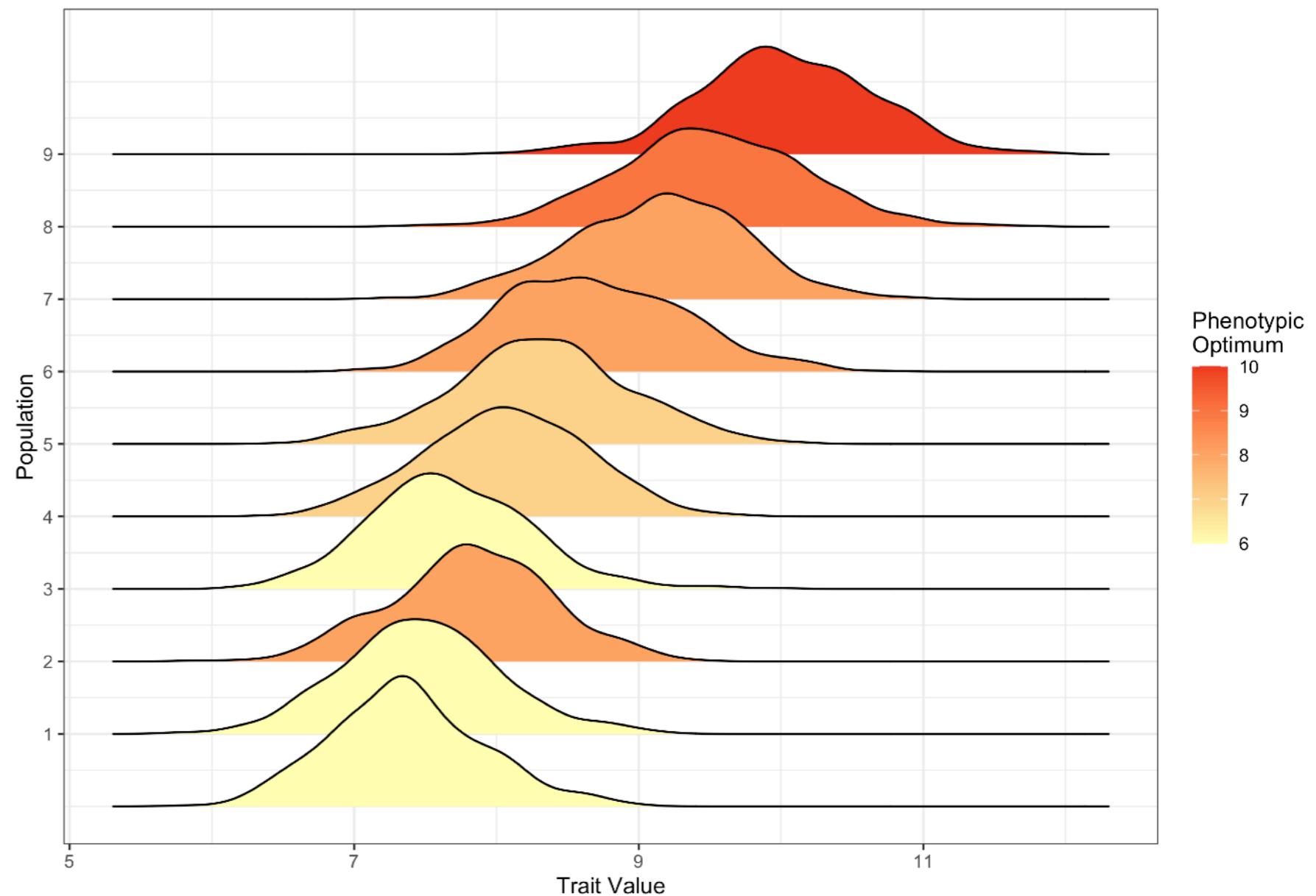
- For N tests and a specified Type I error, the FDR is defined in terms of the number of cases where the null hypothesis is rejected R :

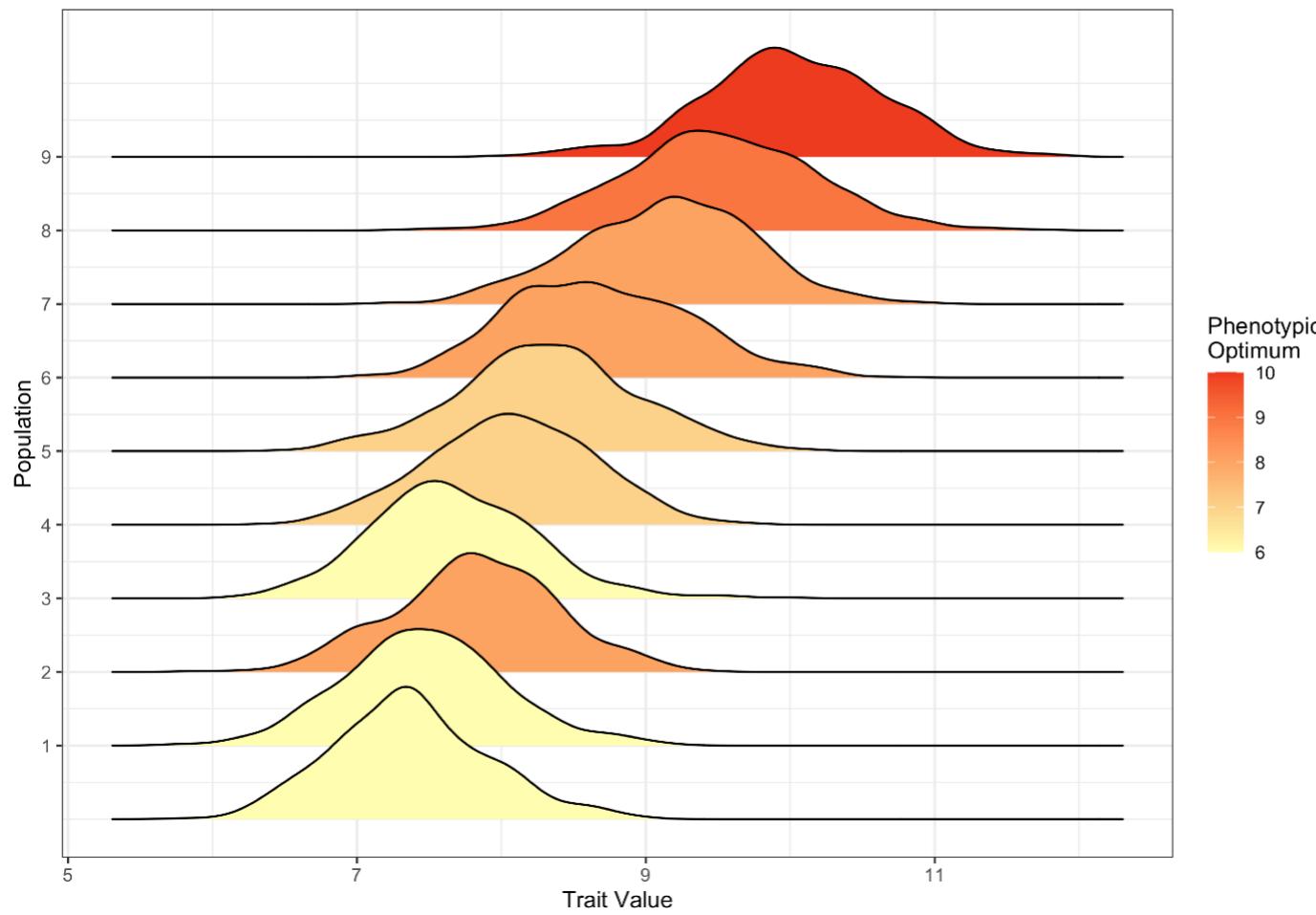
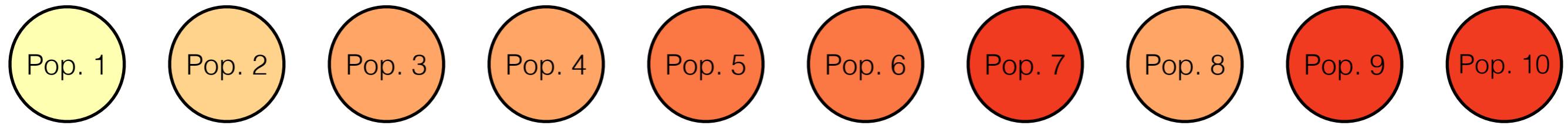
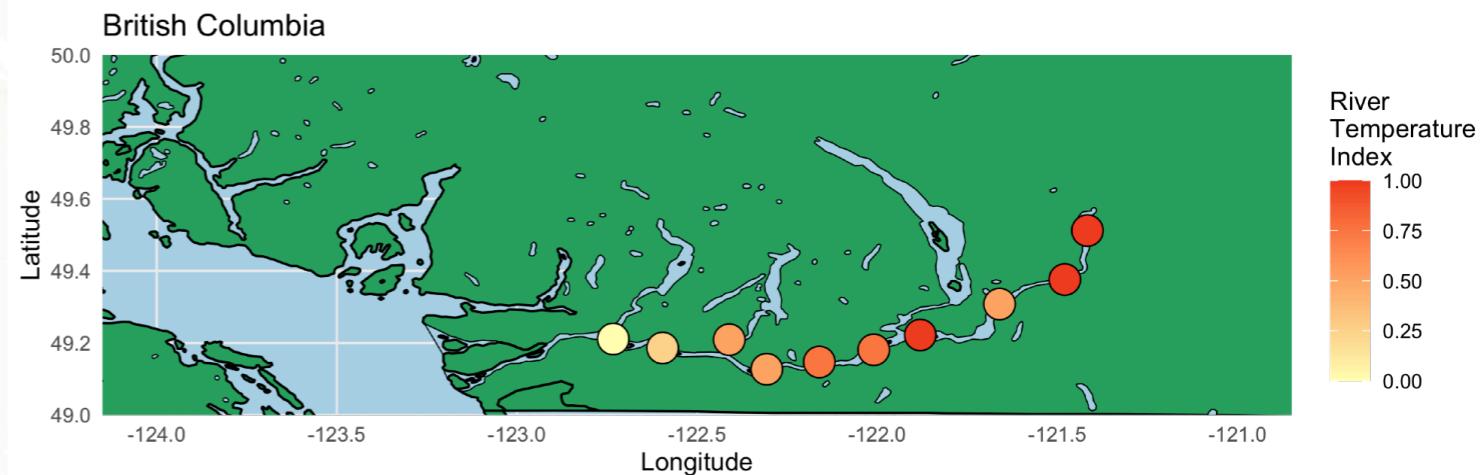
$$FDR = \frac{N * \alpha}{R}$$

- Intuitively, the FDR is the proportion of cases where we reject the null hypothesis that are false positives
- We can estimate the FDR, e.g. say for $N=100,000$ tests and a Type I error of 0.05, we reject the null hypothesis 10,000 times, the $FDR = 0.5$



Which alleles and genes across the genome correlate with temperature tolerance?





What might be a difficulty here?

GWAS programs

- Tassel
- ANGSD
- GWAStools (R)
- GenABEL (R)
- **PLINK/GEMMA**
- GCTA

Plotting

- dplyr + base R for data manipulation
- ggplot2 for plotting

