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

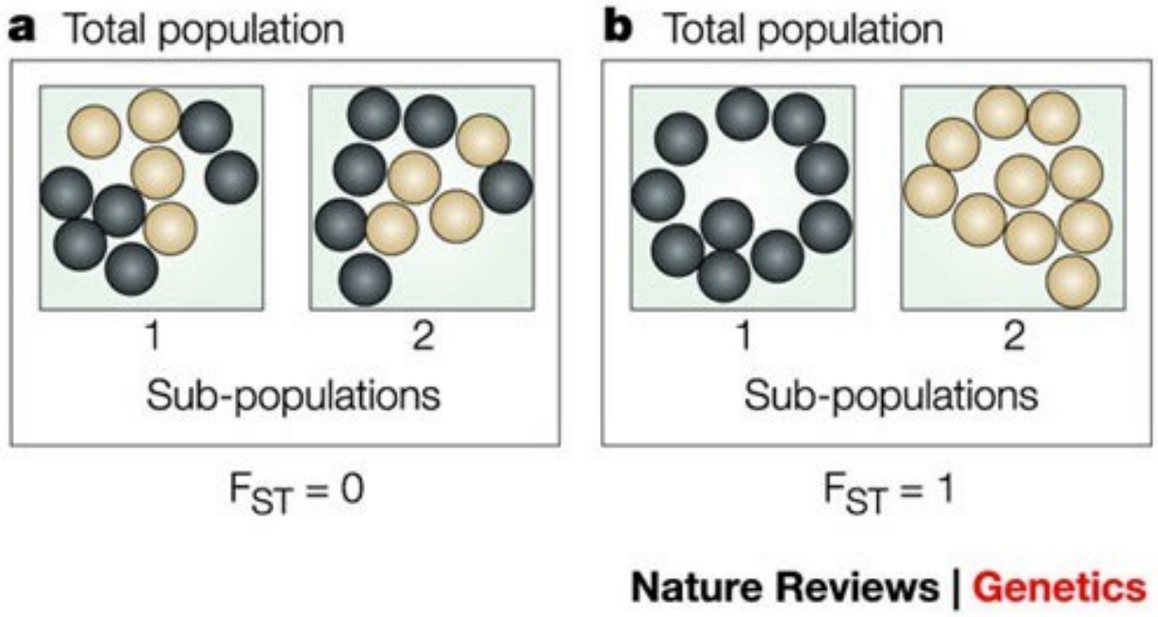
Considerations for SNPs

- 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}



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

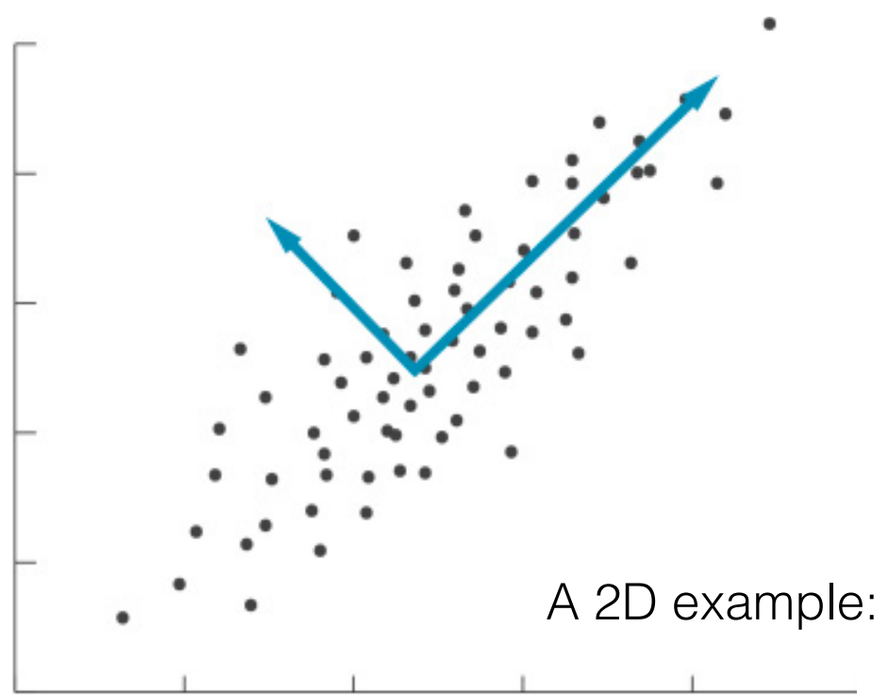
F_{ST} Programs

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

Many different Fst estimators:
Wright, Weir and Cockerham, and Hudson et al.

Principal Component Analysis

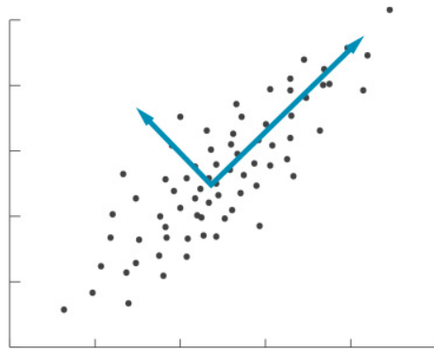
- Model-free approach to assaying populations structure
- Converts a set of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components.

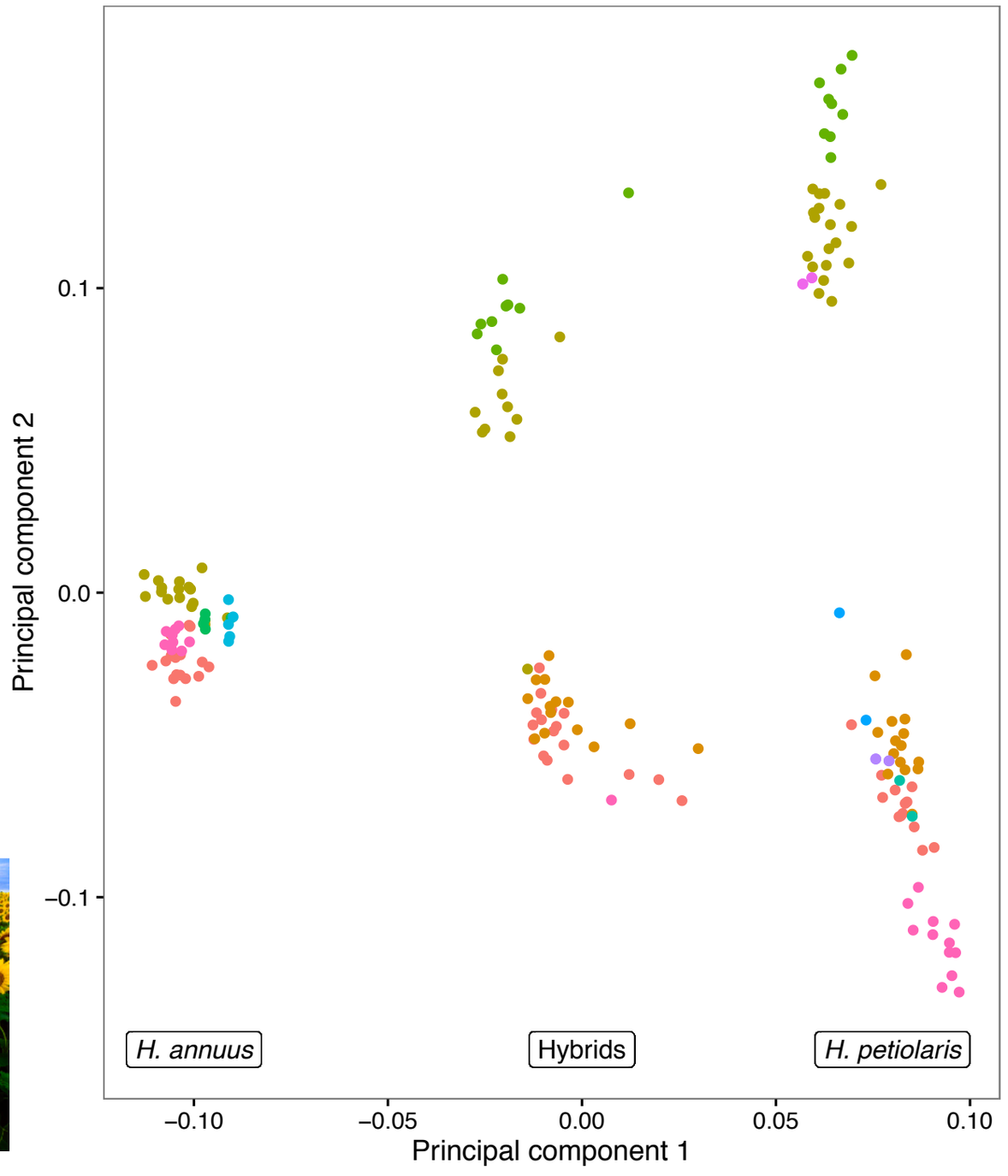


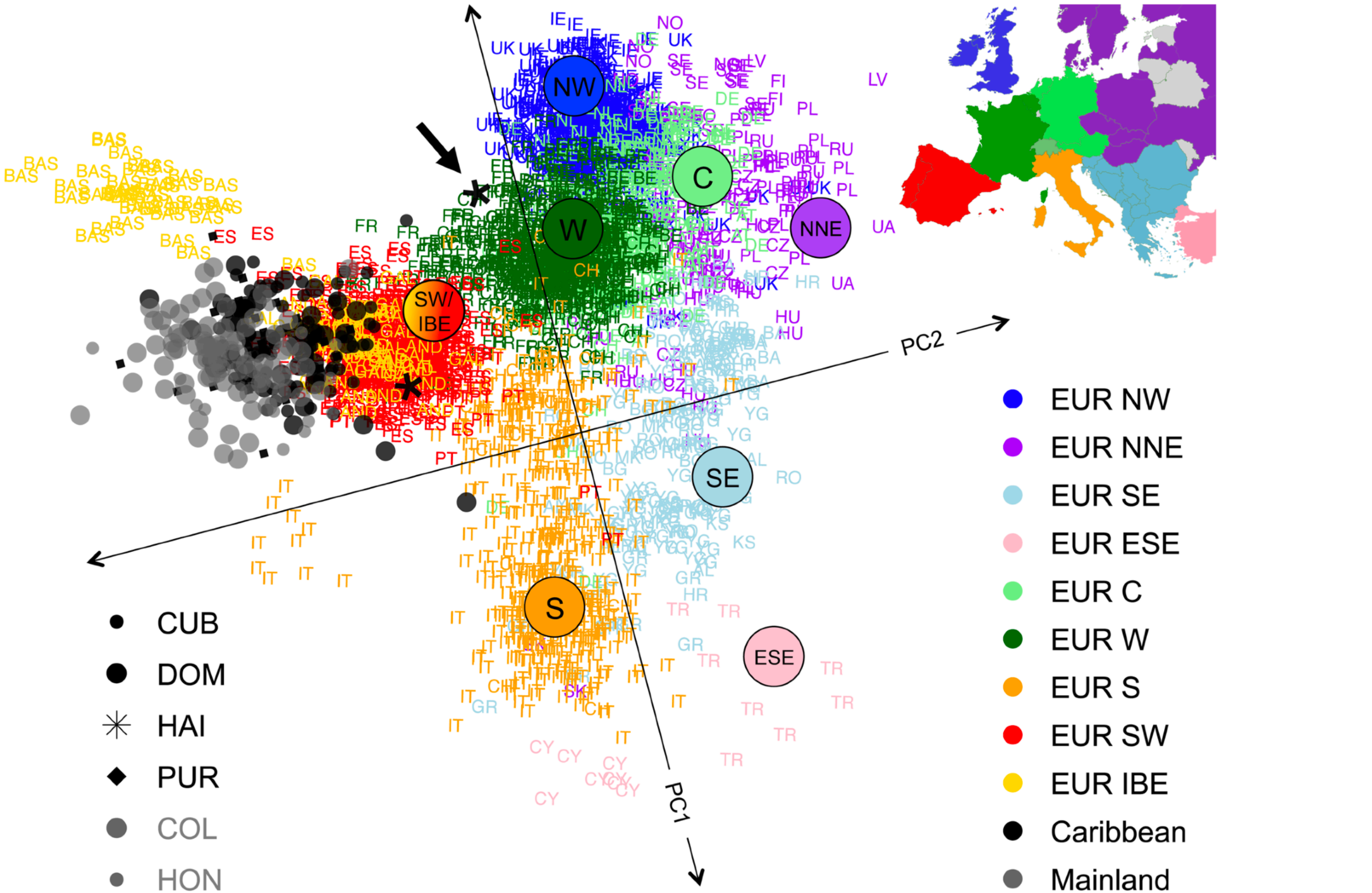
A 2D example:

5396 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5395 2 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 -1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
5394 3 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
5385 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 1 0 1 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5381 4 0 -1 -1 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5378 5 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5377 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 -1 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5375 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
5369 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5359 7 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5357 8 0 0 0 0 0 0 0 1 1 0 -1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
5348 9 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5346 8 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5079 1 0 0 0 0 0 0 0 2 -1 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4886 10 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5393 8 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
5053 6 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5223 5 0 0 0 0 0 0 0 1 1 0 -1 0 0 0 0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
5040 11 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
556 12 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
503 13 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
5333 14 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4964 15 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5015 14 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
5002 16 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4989 17 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4968 6 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0
5197 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4966 8 0 0 0 0 0 0 0 0 0 0 0 0 -1 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5234 18 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4965 19 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4907 8 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
4862 20 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4958 5 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4955 14 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
4935 4 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 1
4928 10 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
5343 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0
5294 16 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
5031 2 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 2 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
4828 21 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
4925 22 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0
4921 23 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5323 8 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
5389 16 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

Principal Component Analysis






NW
C
W
NNE
SW/
IBE
SE
S
ESE
PC1
PC2
CUB
DOM
HAI
PUR
COL
HON
EUR NW
EUR NNE
EUR SE
EUR ESE
EUR C
EUR W
EUR S
EUR SW
EUR IBE
Caribbean
Mainland

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**

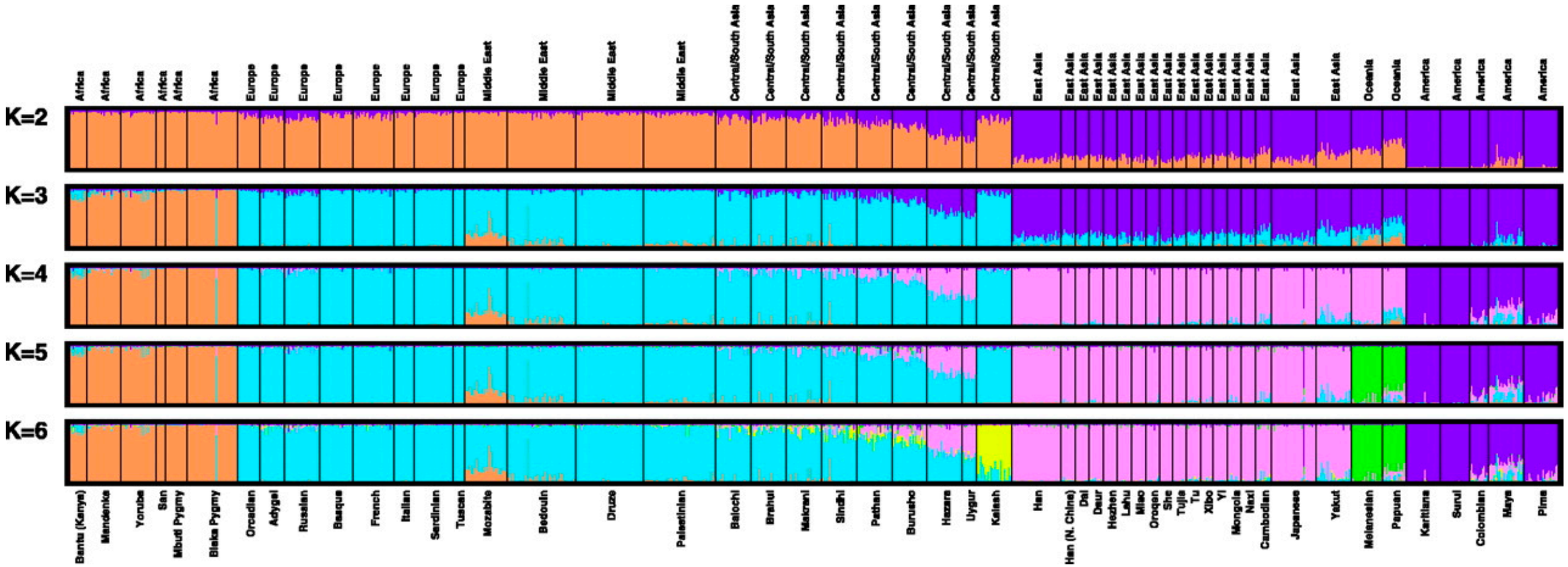
STRUCTURE

- **Models** *K* populations with a set of allele frequencies at each locus, *assumes hardy-weinberg*

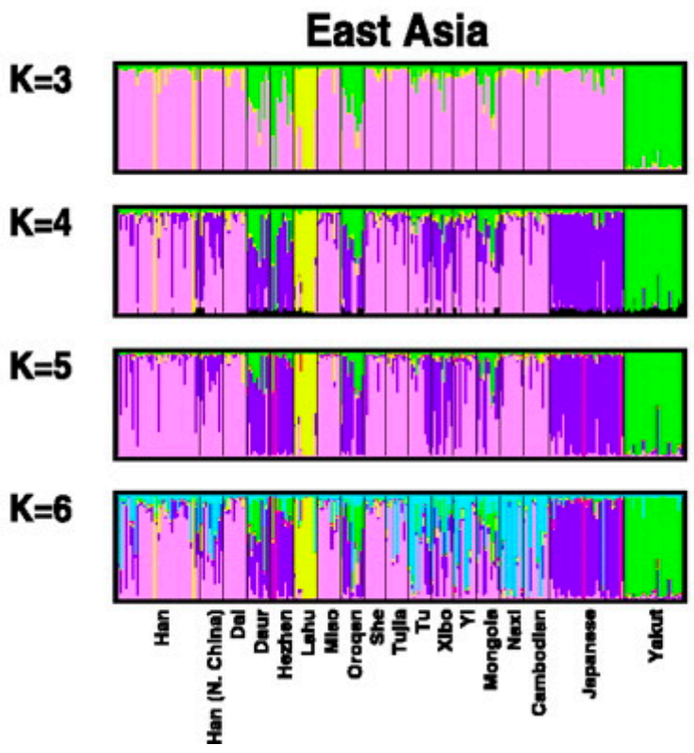
- Individuals are assigned to one or more populations based on their genotype

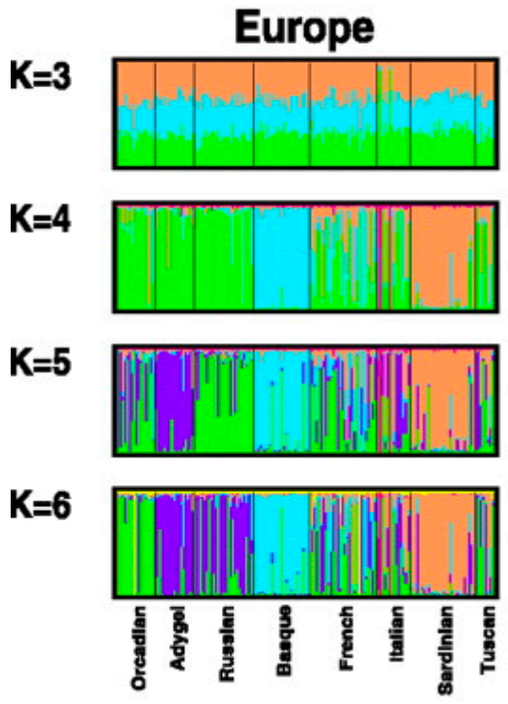
- Can pick the best *K* based on *the fit of the model to your data*

STRUCTURE



Rosenberg et al 2002

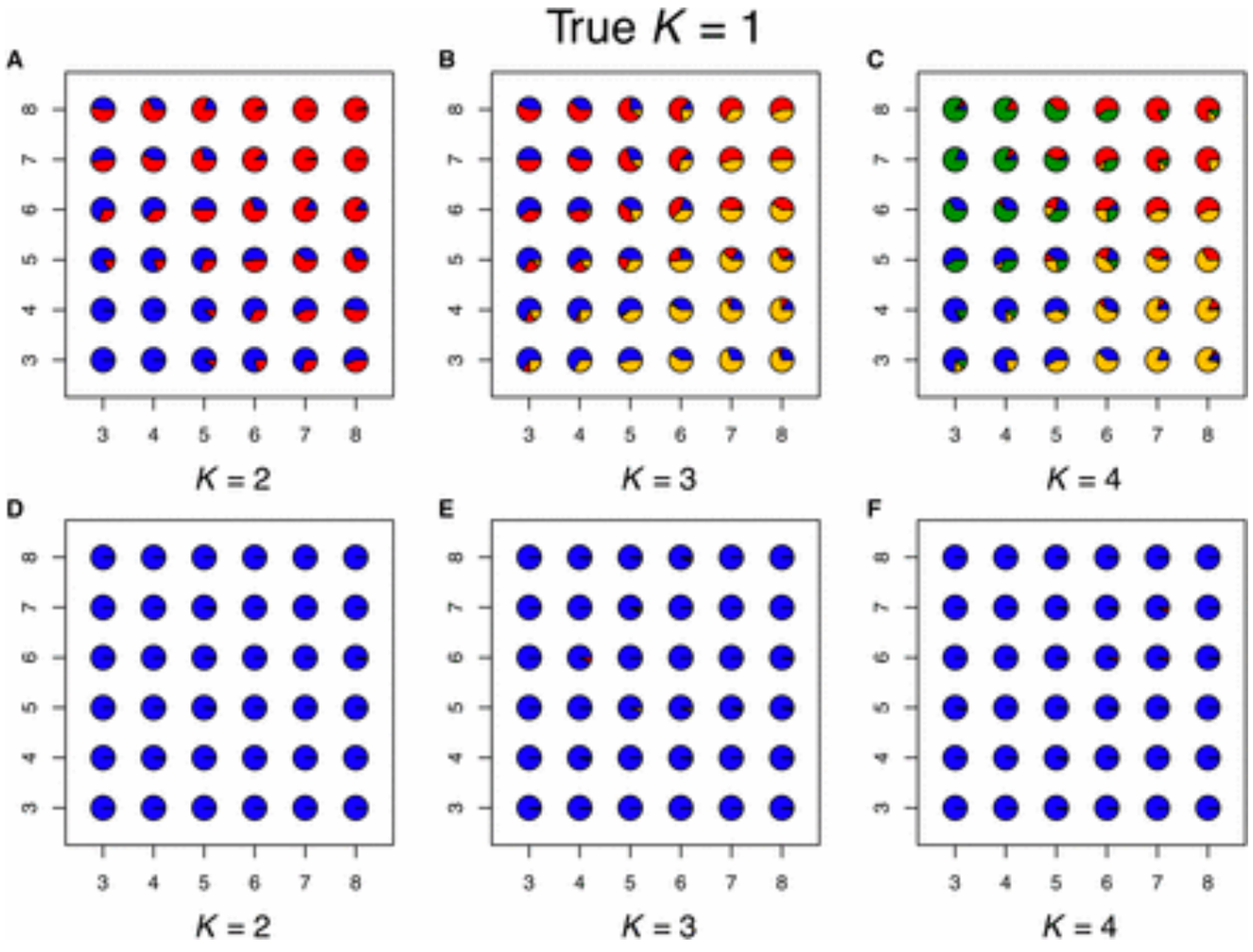




Inferences depend on the scale of analysis

STRUCTURE

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



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

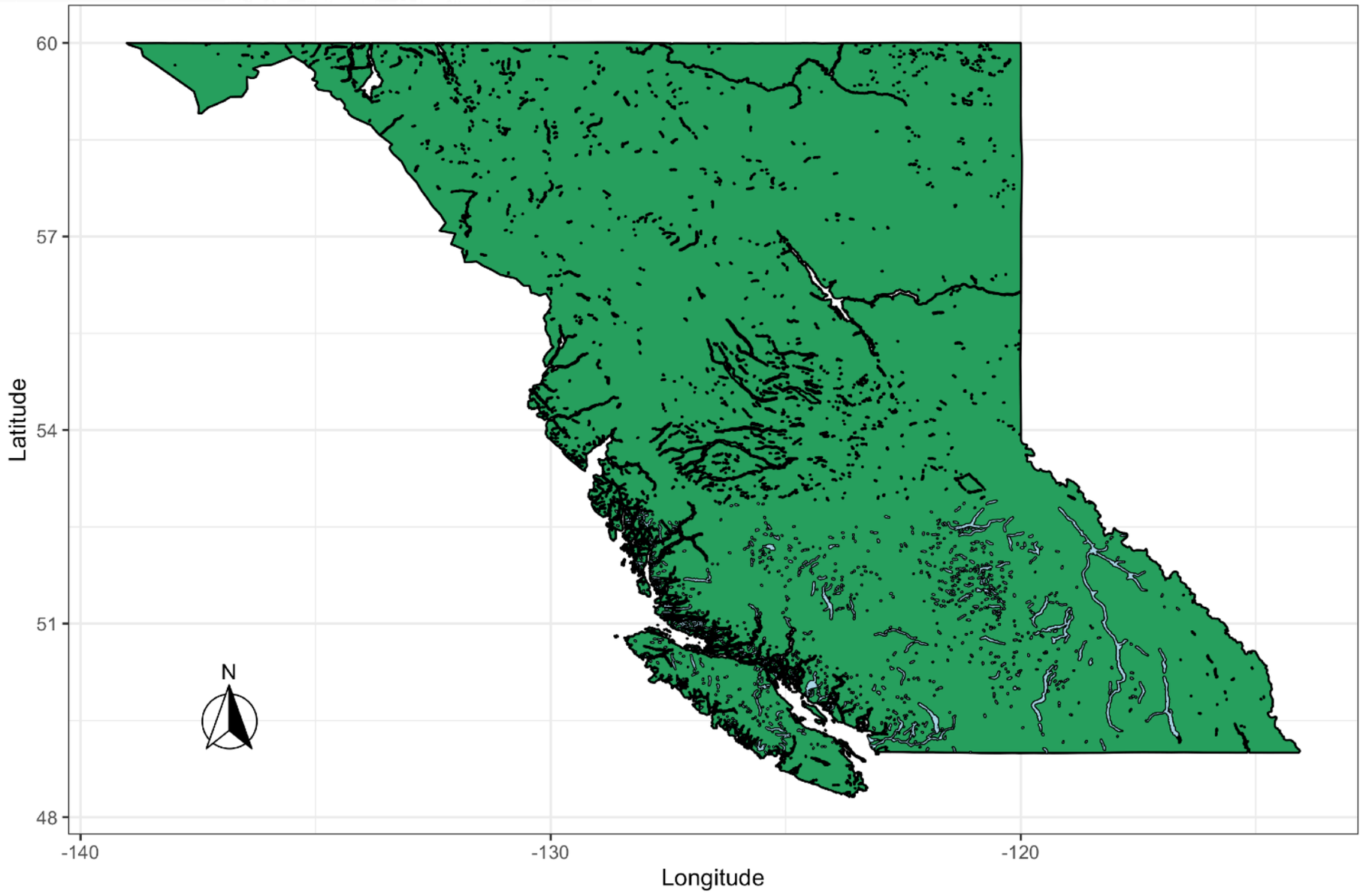
Accounting for non-random mating:
No groupings of excess covariance in allele frequencies

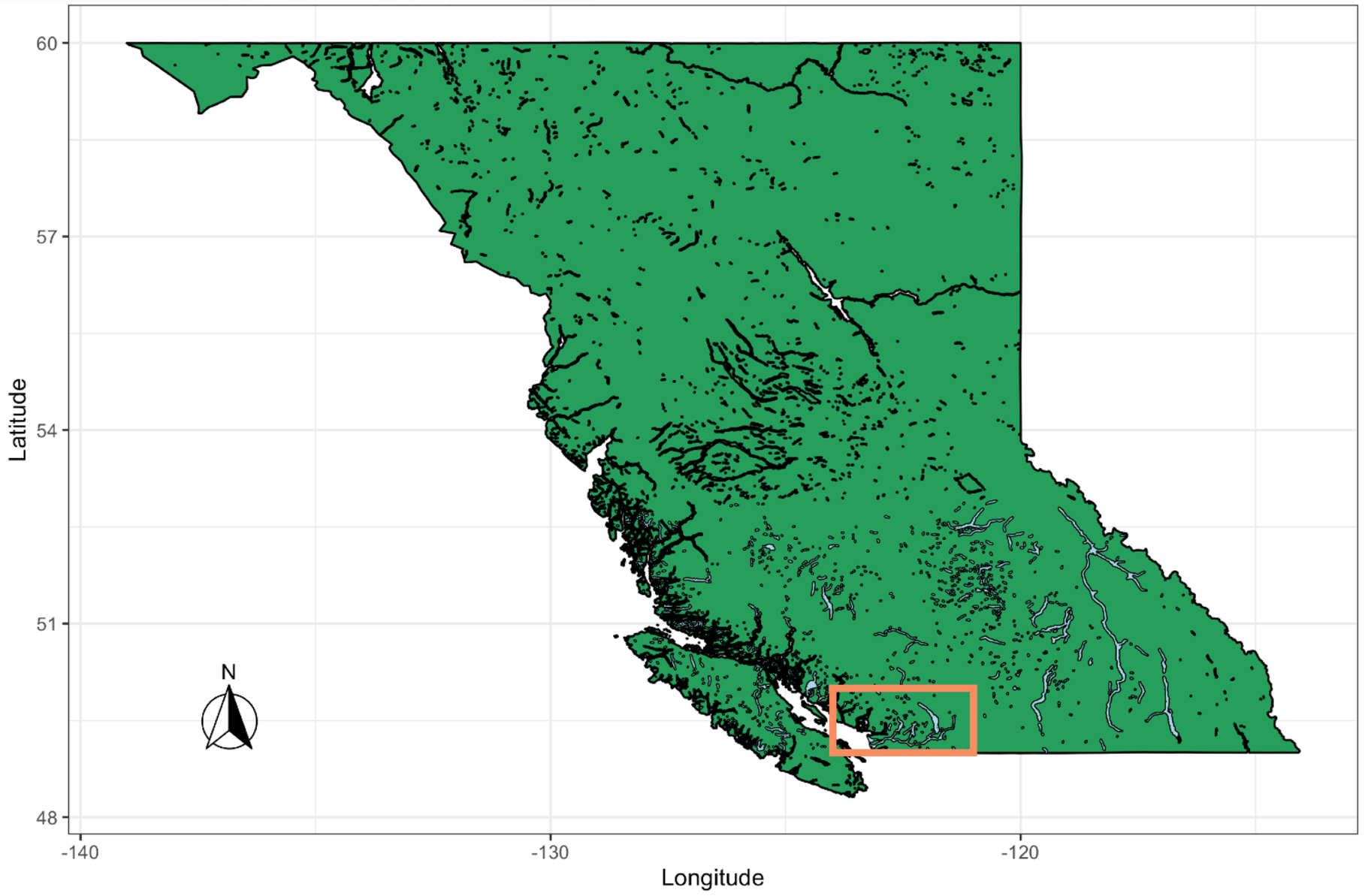
STRUCTURE programs

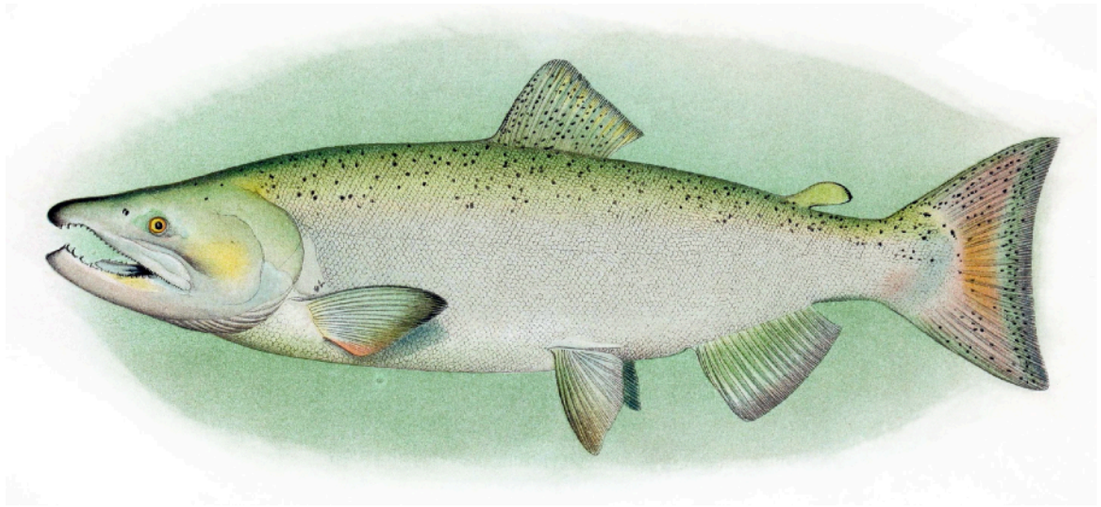
- STRUCTURE (original implementation, lots of model options, very slow & made for small genetic data sets)
- **Admixture** (fast - made for large genomic datasets)
- FASTstructure (fast - made for large genomic datasets)
- NGSadmix (ANGSD implementation, based on genotype likelihoods)

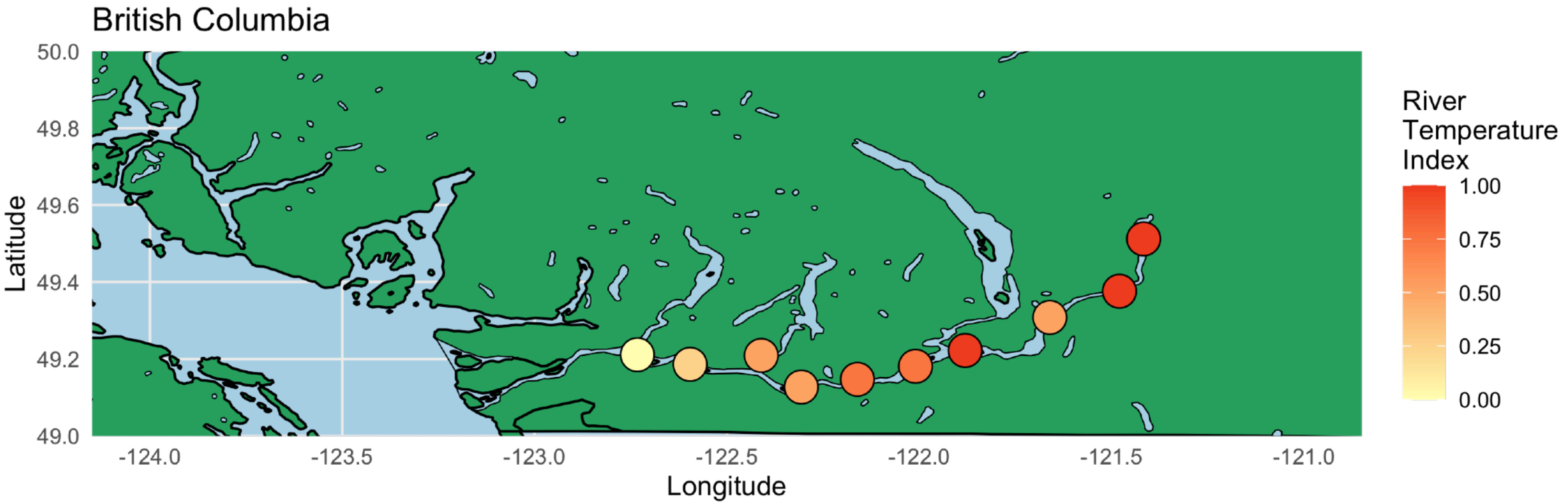
STRUCTURE

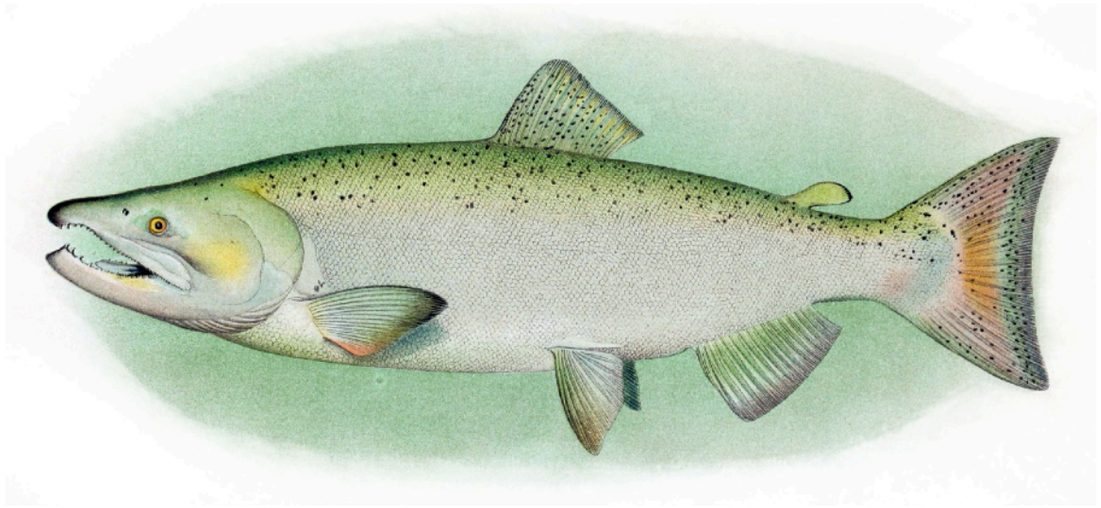
- You should prune dataset to unlinked SNPs
- Run multiple times to confirm consistency - “Cross validation”
- Admixture has a bootstrap option, to get CIs of admixture proportions

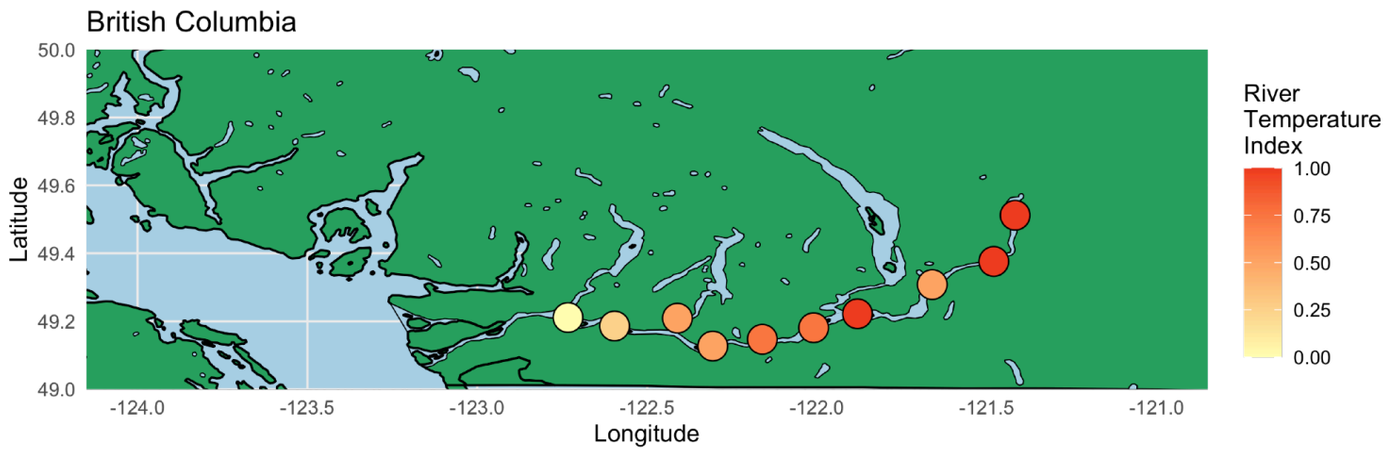

60
57
54
51
48
Latitude
-140
-130
-120
Longitude
N

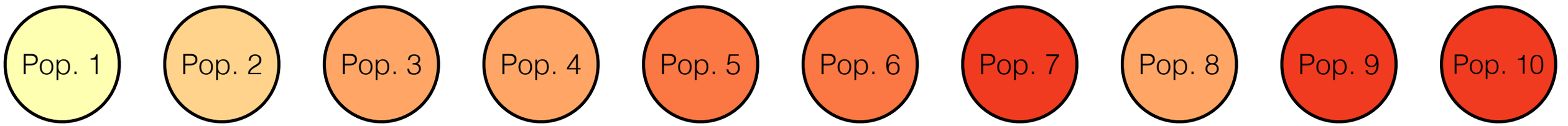

60
57
54
51
48
Latitude
-140
-130
-120
Longitude
N

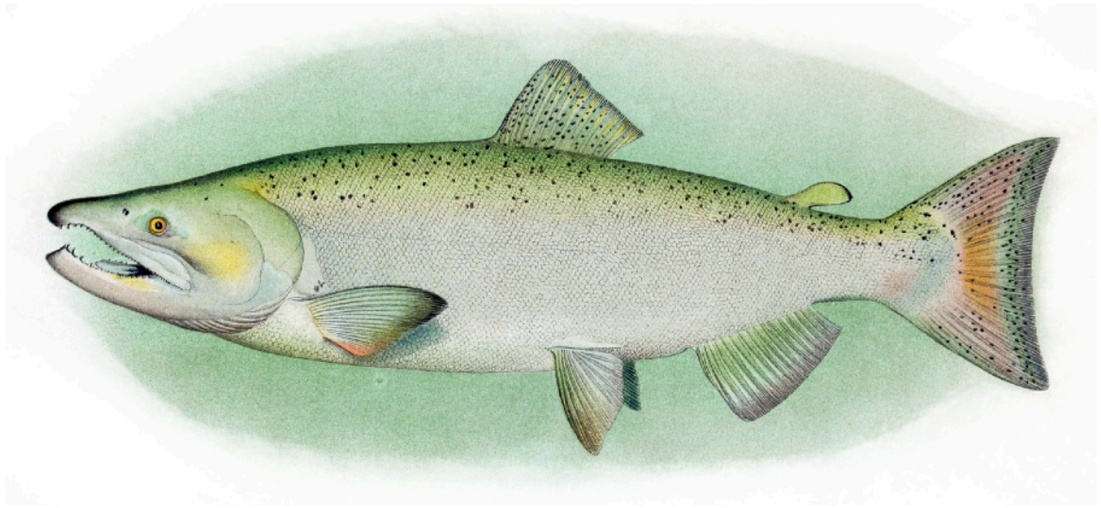


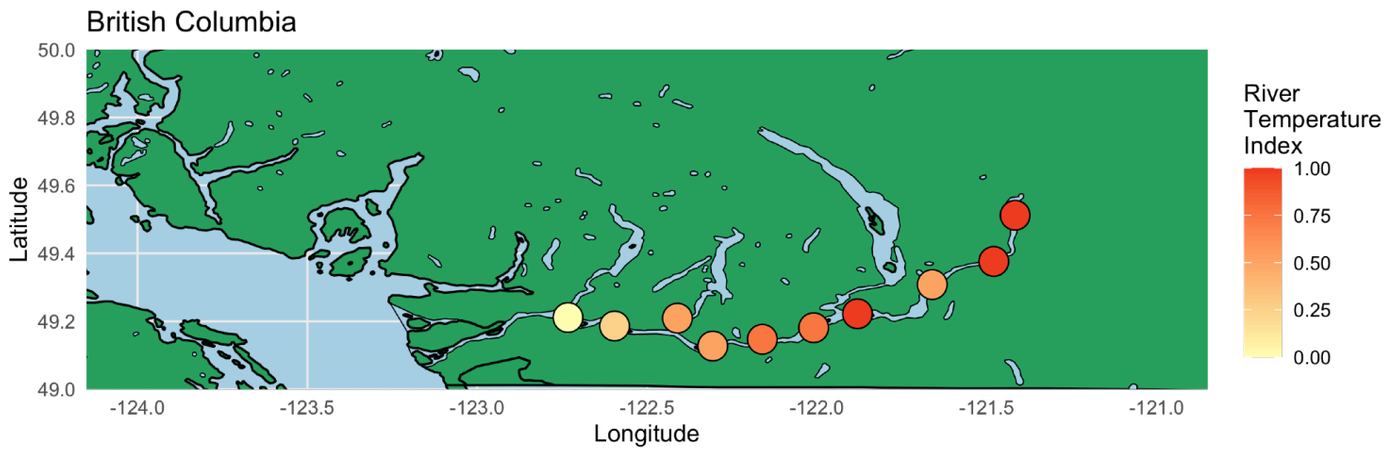

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

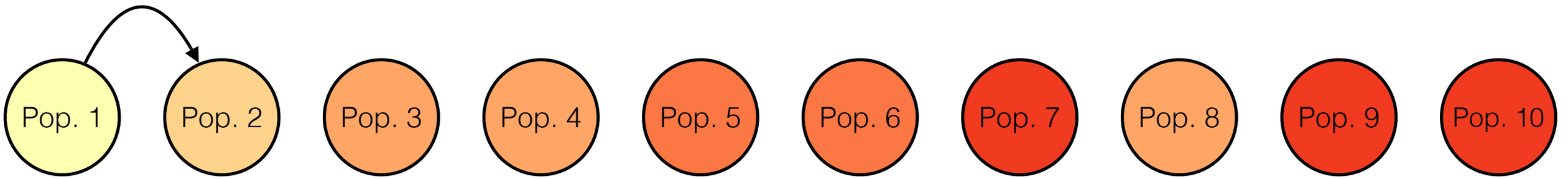


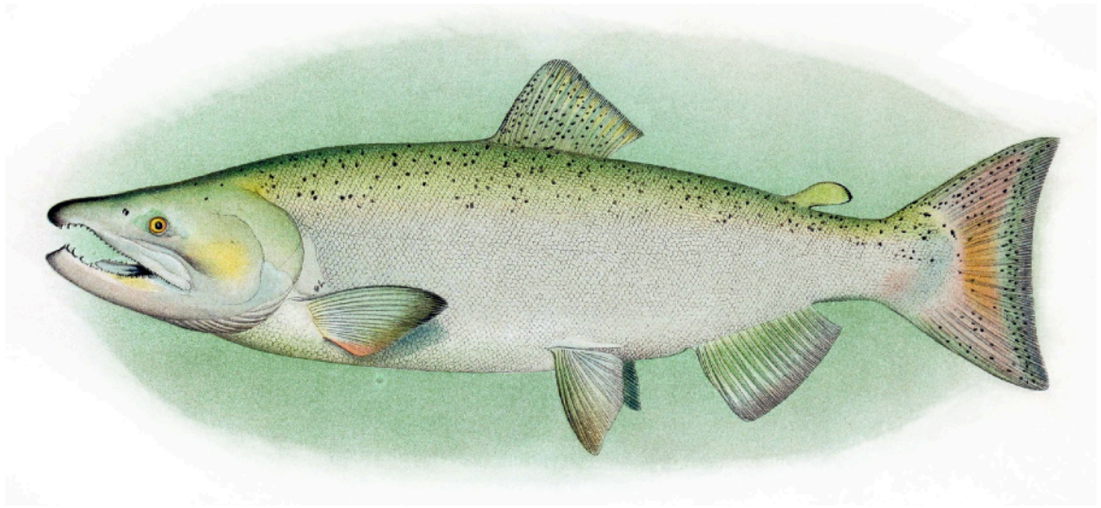

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

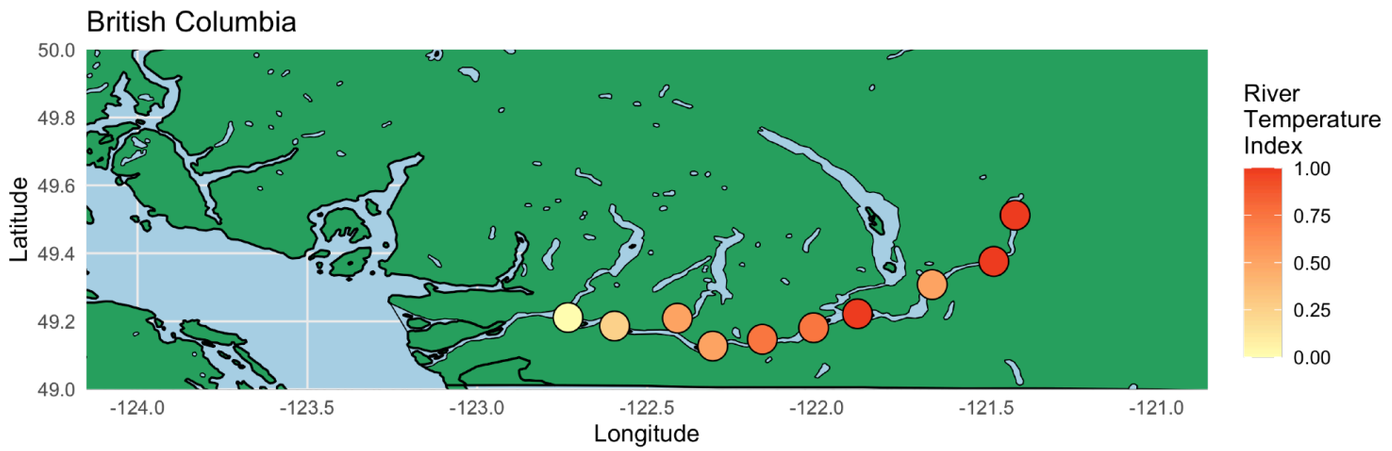

Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
Pop. 6
Pop. 7
Pop. 8
Pop. 9
Pop. 10

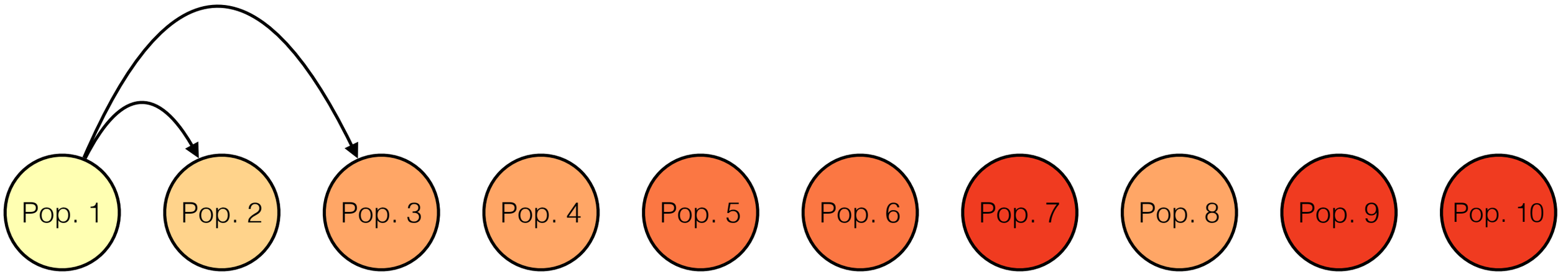


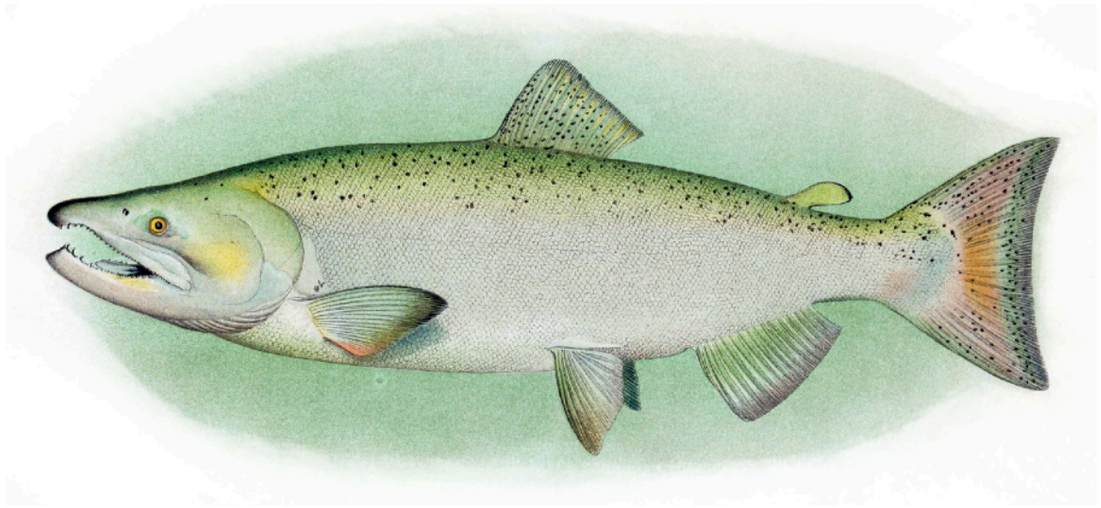

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

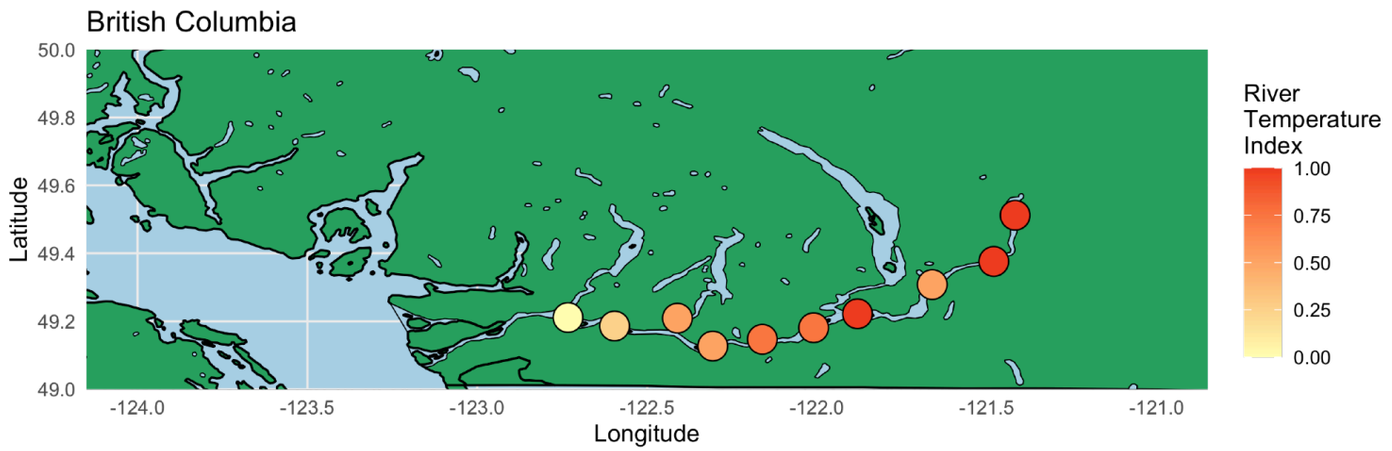

Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
Pop. 6
Pop. 7
Pop. 8
Pop. 9
Pop. 10

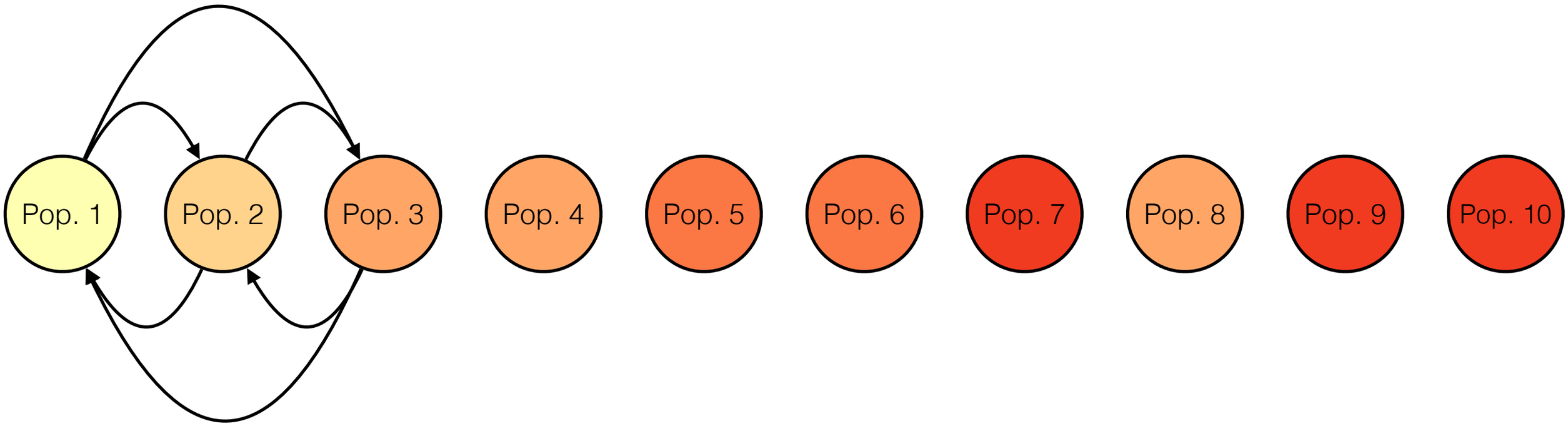


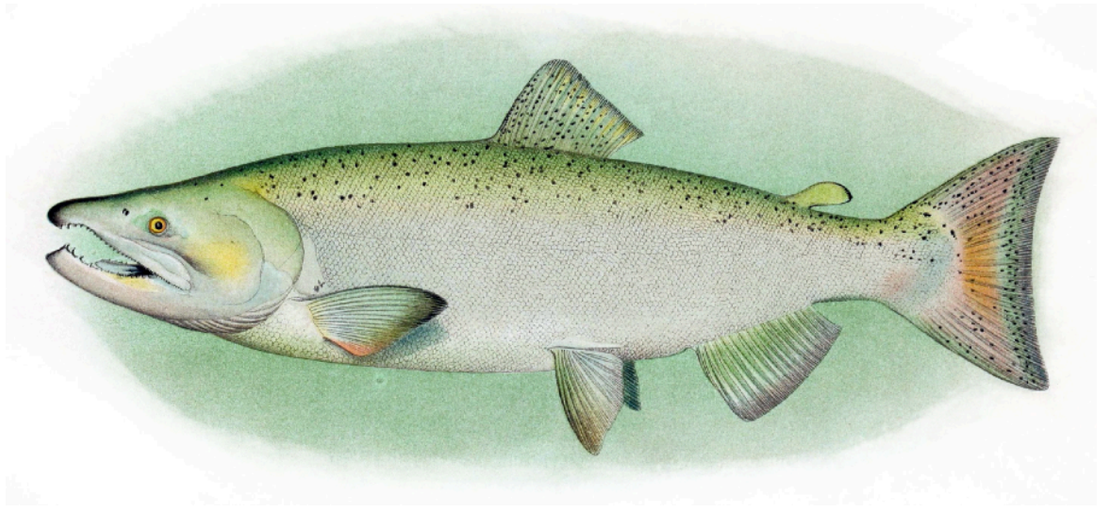

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

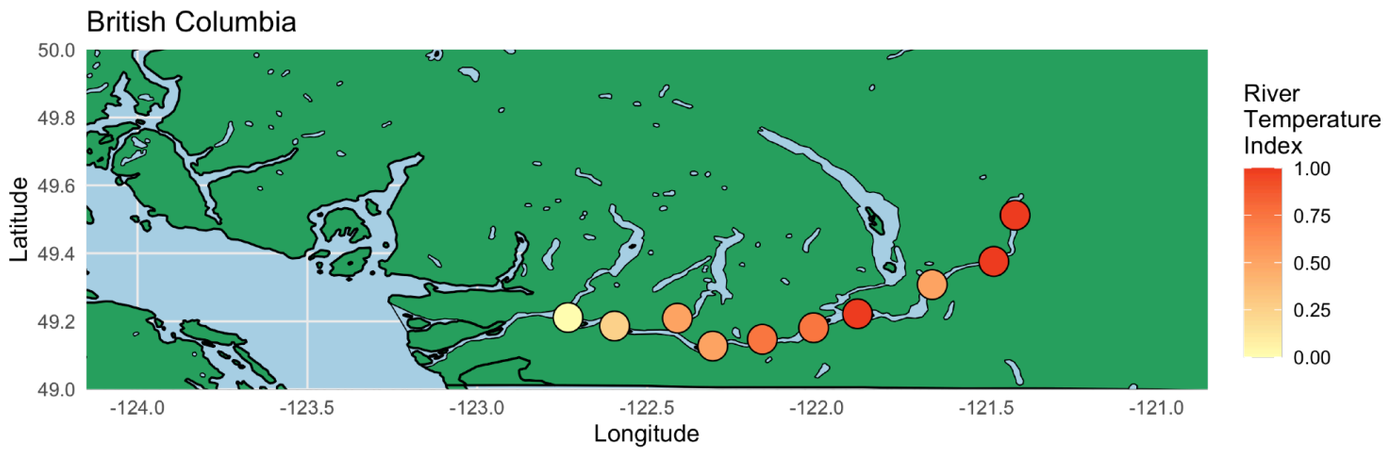

Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
Pop. 6
Pop. 7
Pop. 8
Pop. 9
Pop. 10

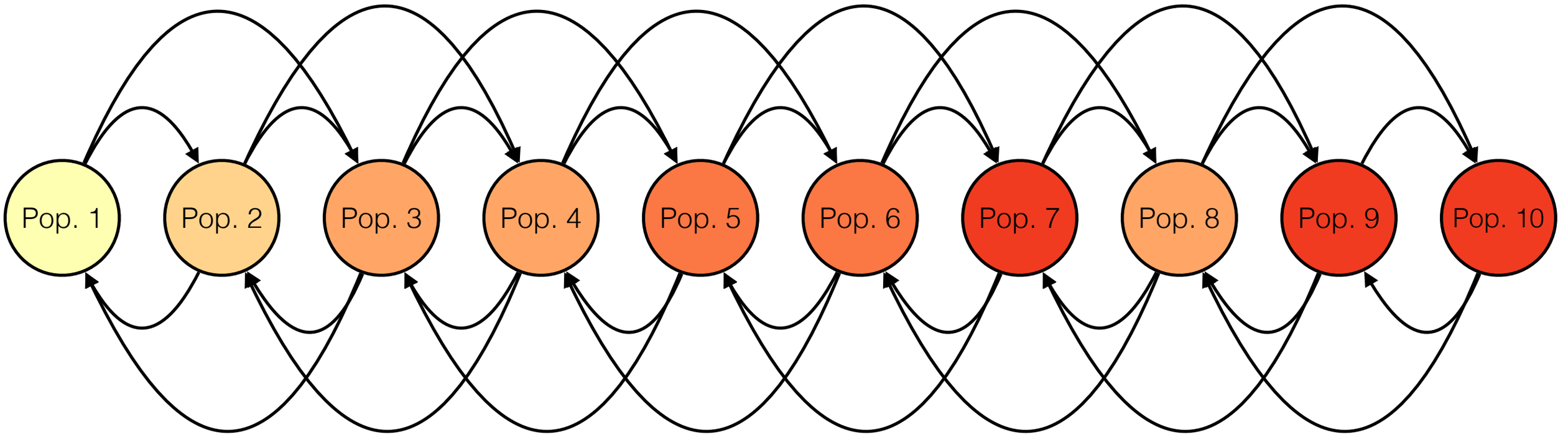


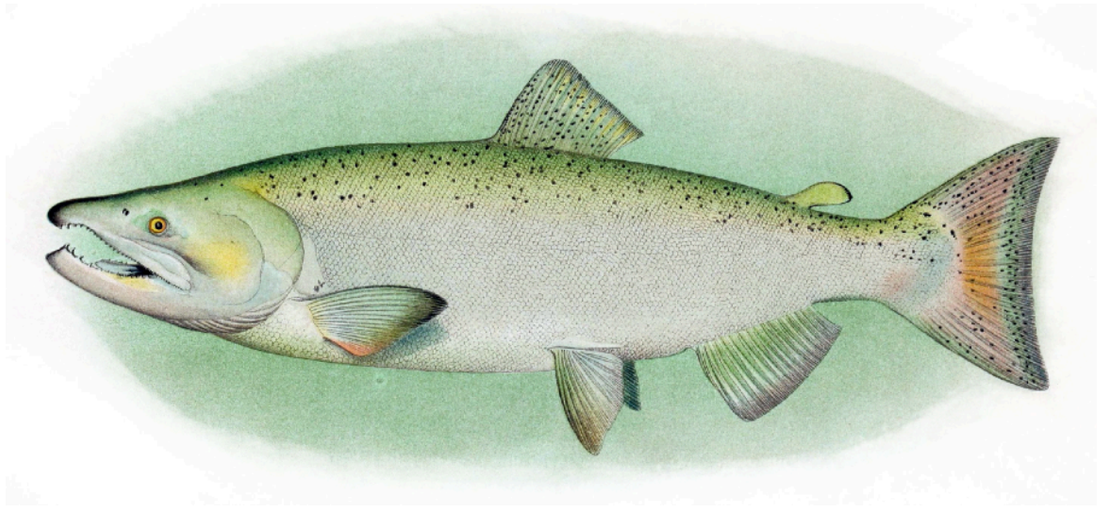

British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00

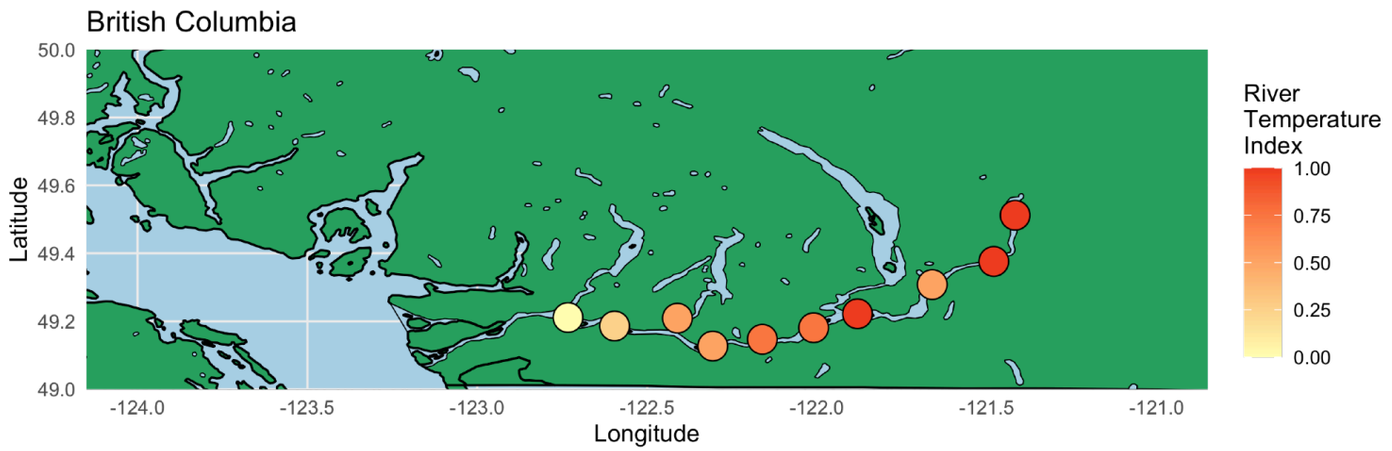

Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
Pop. 6
Pop. 7
Pop. 8
Pop. 9
Pop. 10

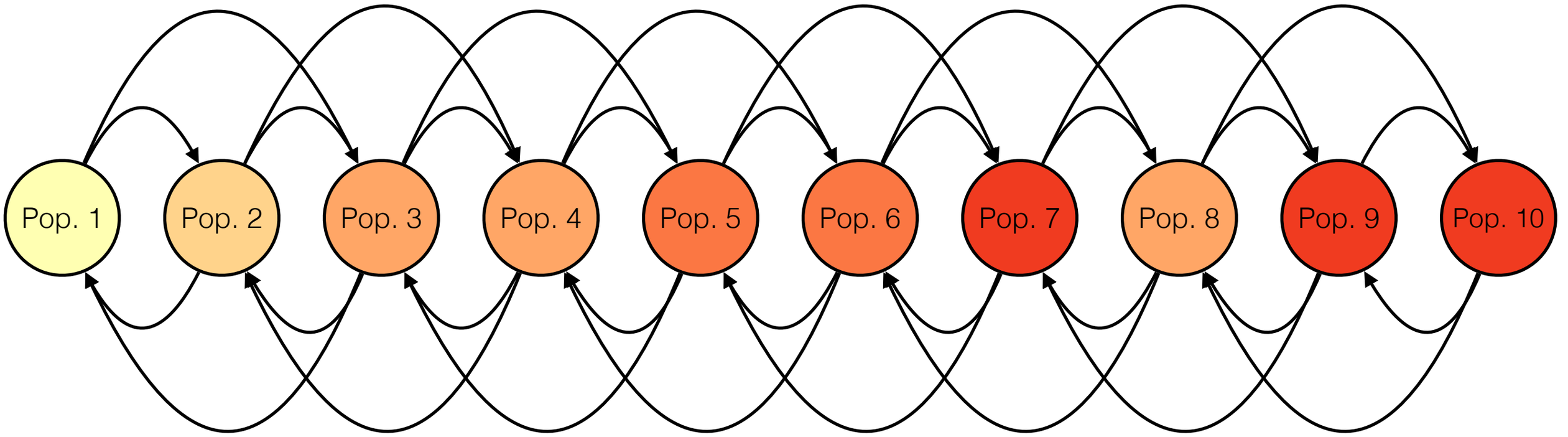



British Columbia
50.0
49.8
49.6
49.4
49.2
49.0
Latitude
-124.0
-123.5
-123.0
-122.5
-122.0
-121.5
-121.0
Longitude
River Temperature Index
1.00
0.75
0.50
0.25
0.00


Pop. 1
Pop. 2
Pop. 3
Pop. 4
Pop. 5
Pop. 6
Pop. 7
Pop. 8
Pop. 9
Pop. 10





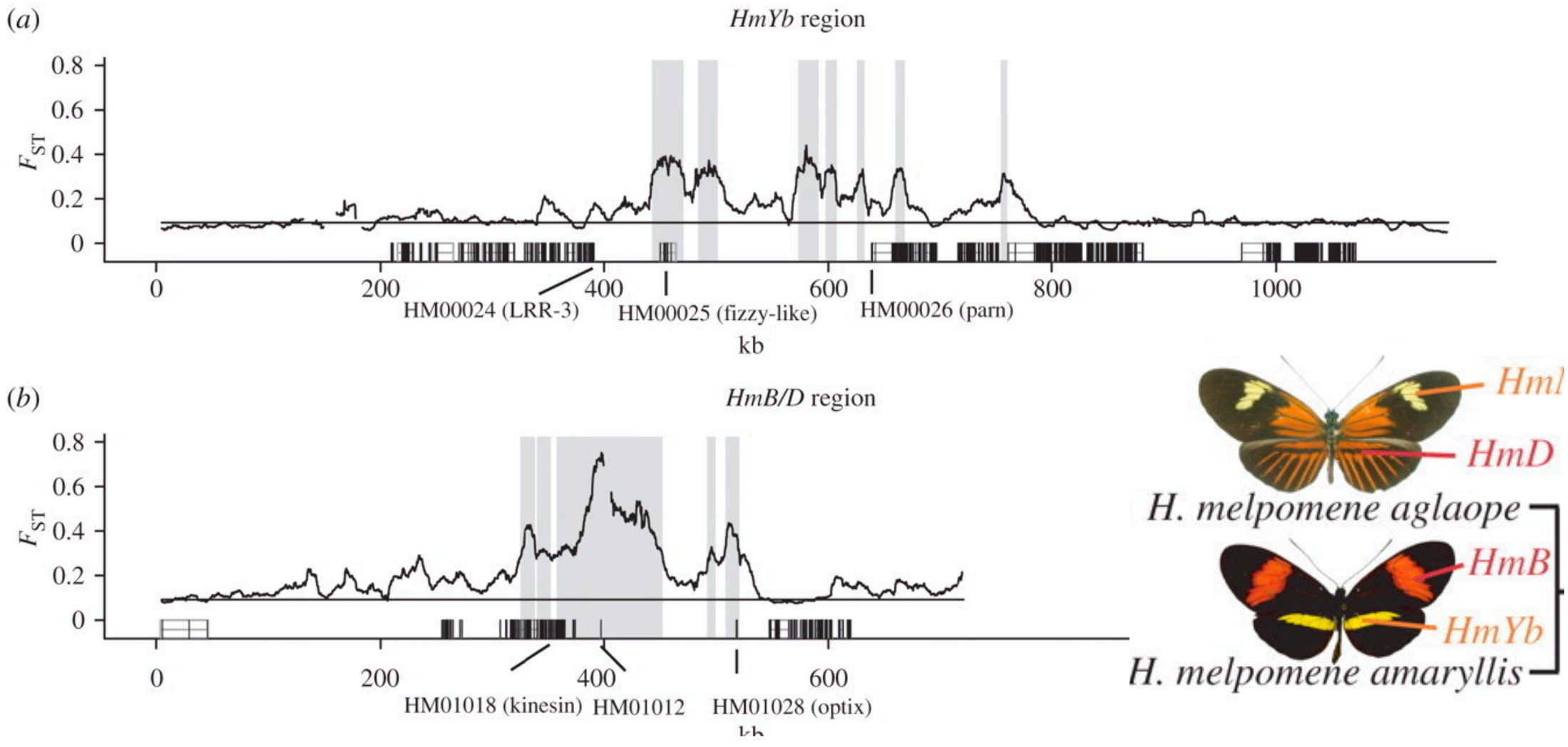


Given our knowledge of the system:
How would you examine patterns of population structure? What would you look out for?

Population genomics of adaptation (and phenotypes)

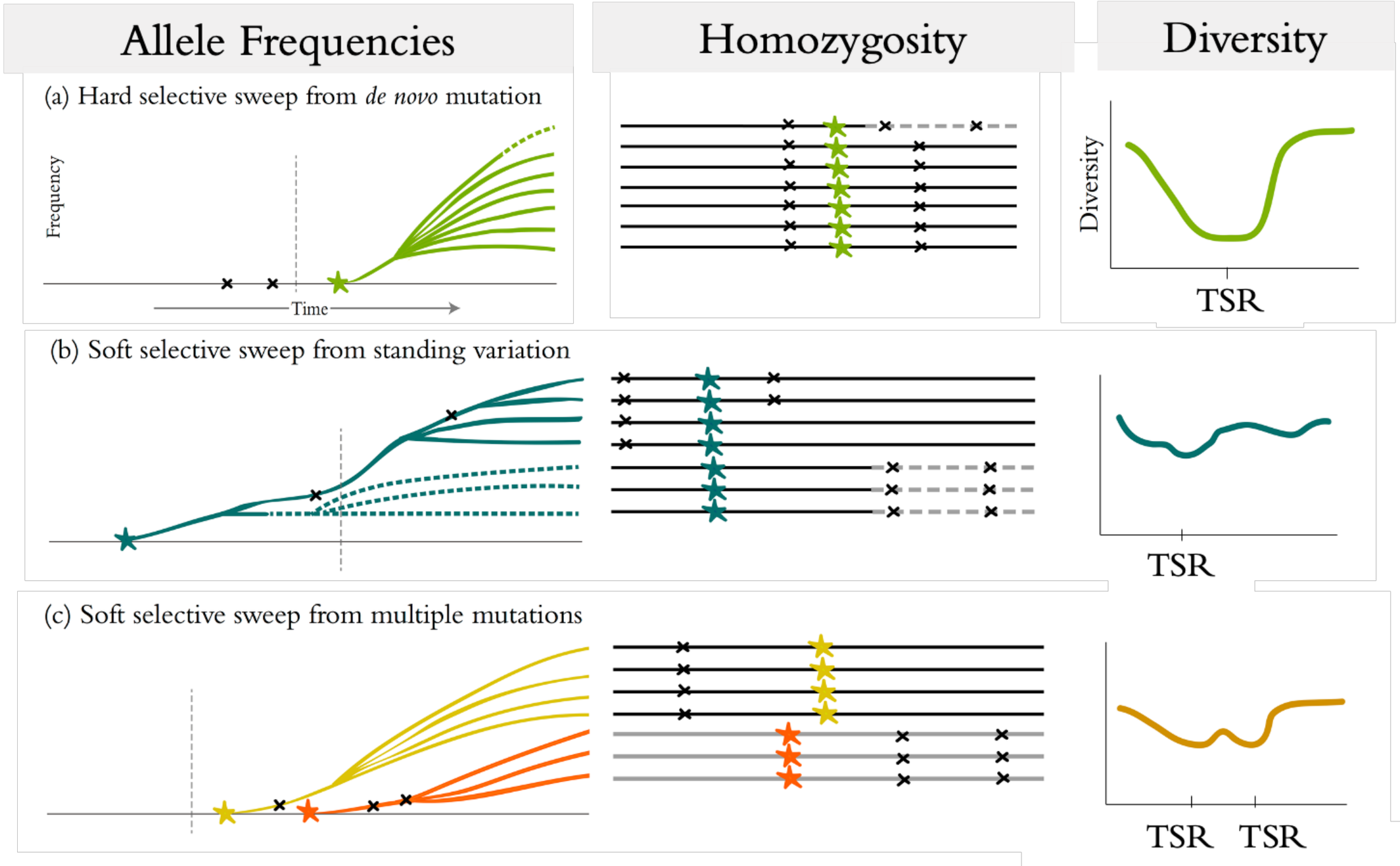
- Scans for selection
  - Fst revisited
  - Selective sweeps
- GWAS

F_{ST}: Scans for selection



Nandeau et al., 2012

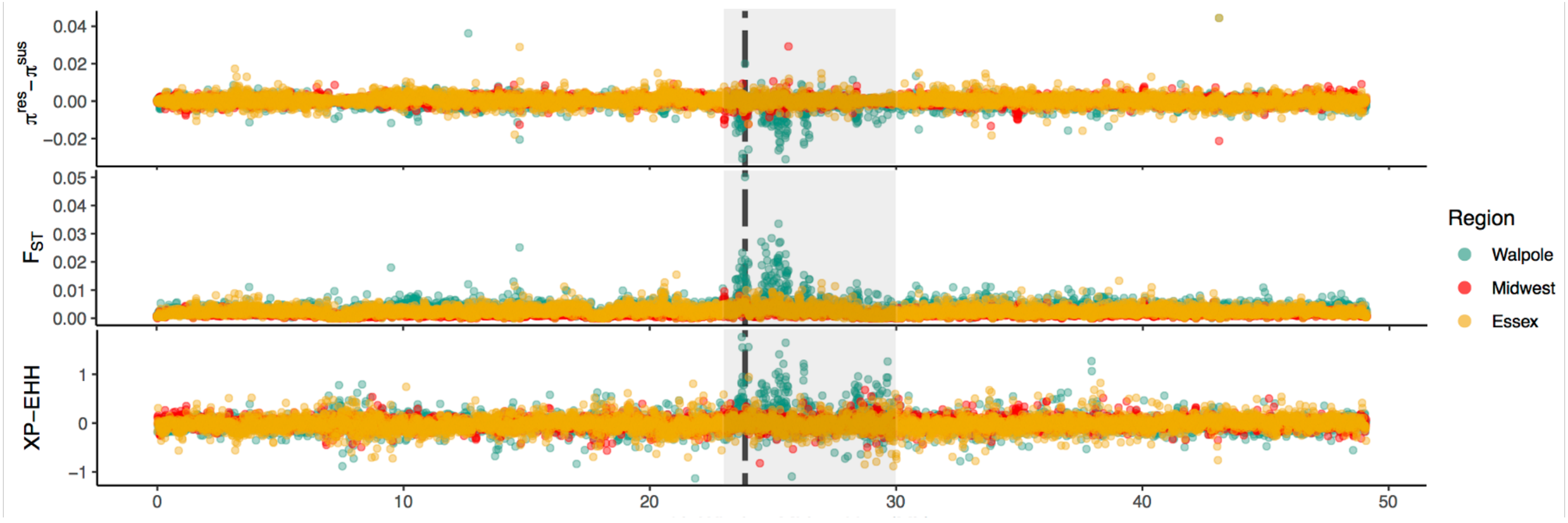
Scans for selection: Sweeps



Hermisson & Pennings 2018

Scans for selection: Sweeps

Herbicide resistance allele



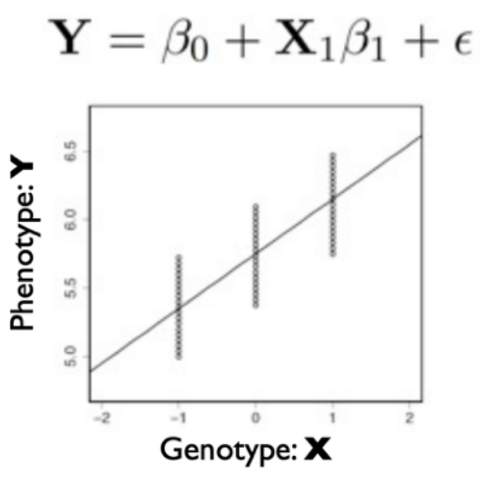
Loss of diversity, excess homozygosity, and allele frequency differentiation around locus in resistant haplotypes

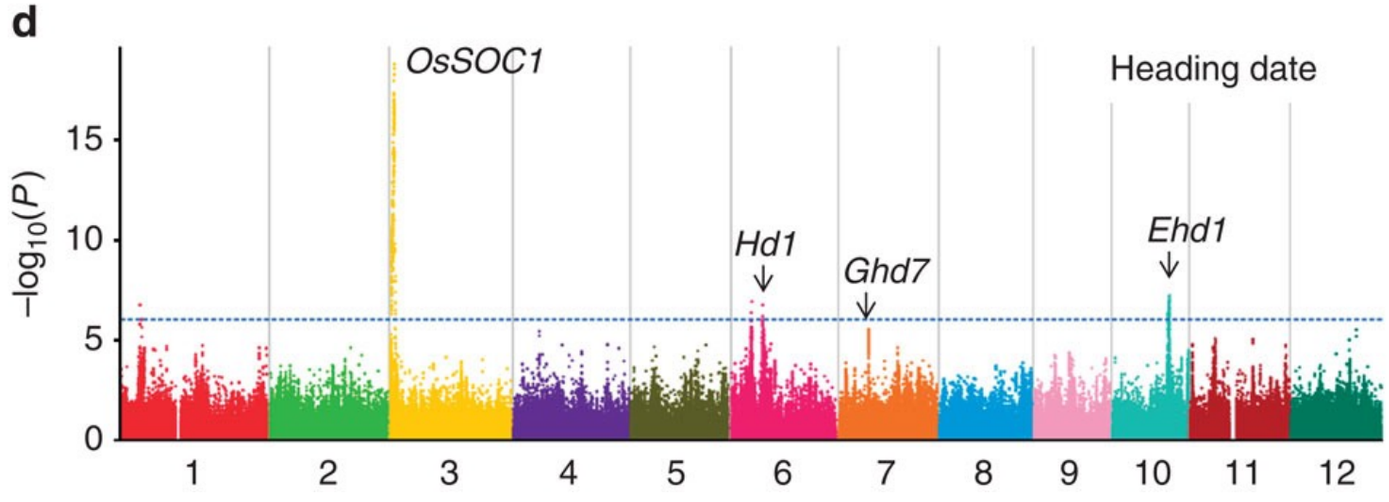
Population genomics of adaptation (and phenotypes)

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

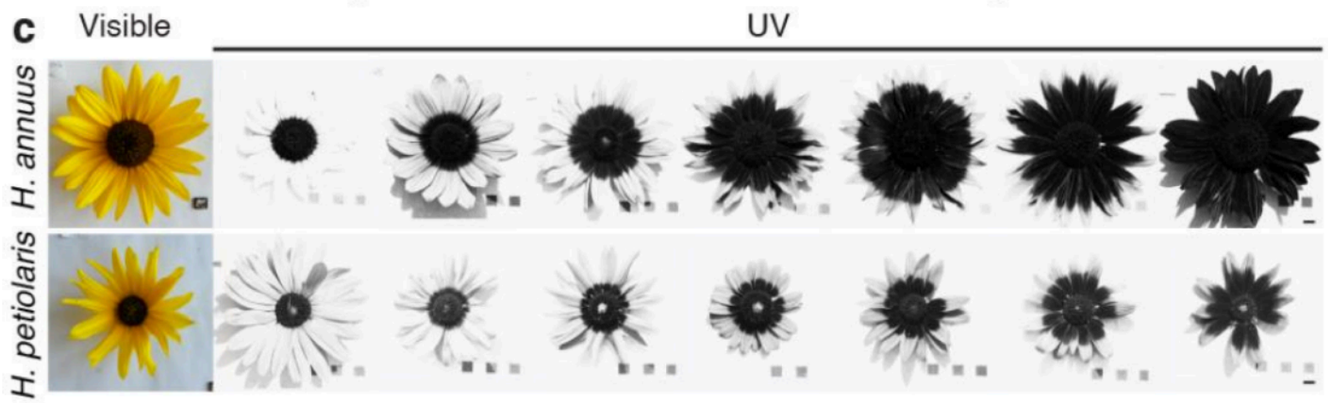
SNP-phenotype associations (GWAS): one allele at a time

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



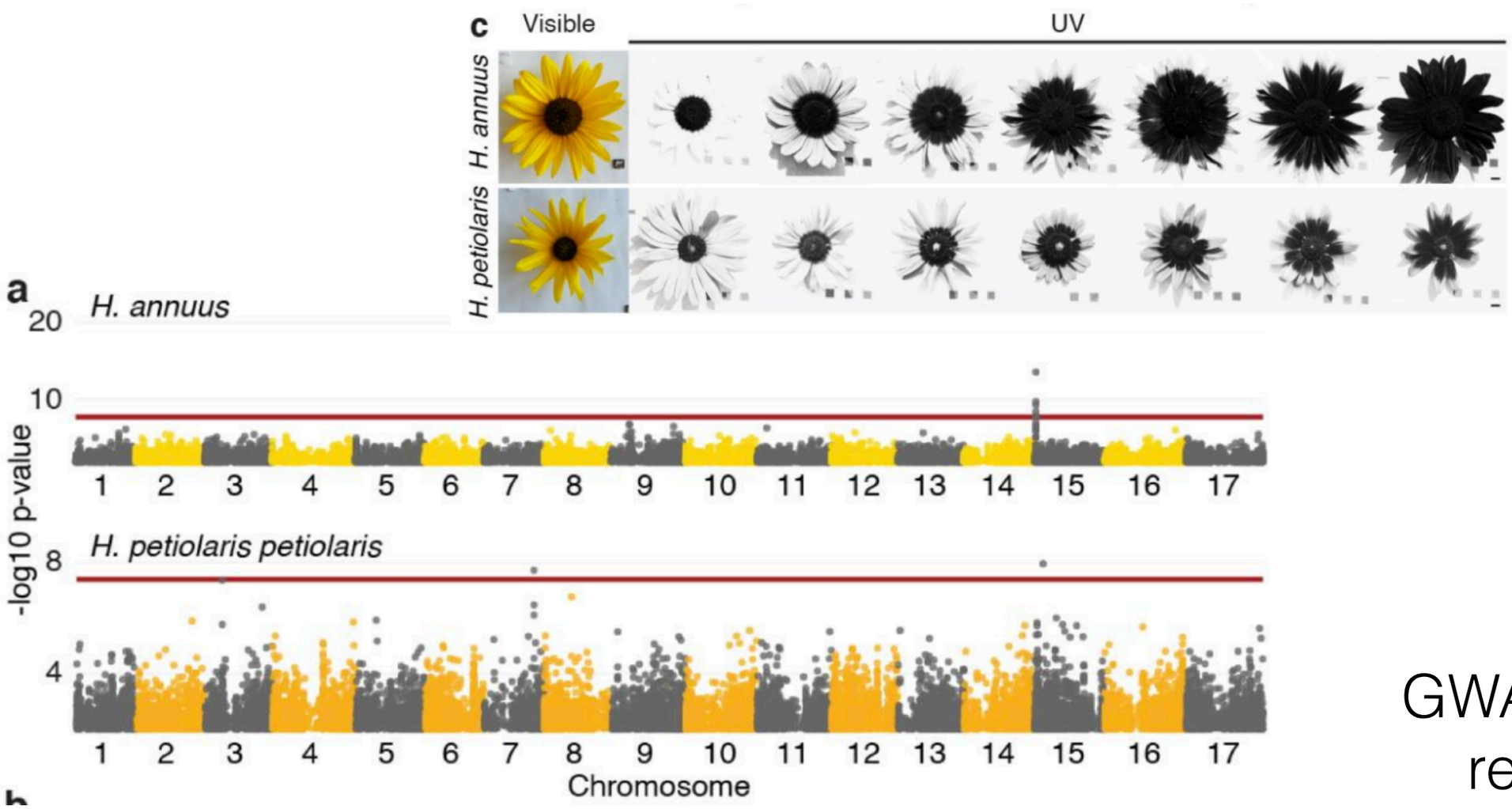


Huang et al., 2015; Nat. Com.



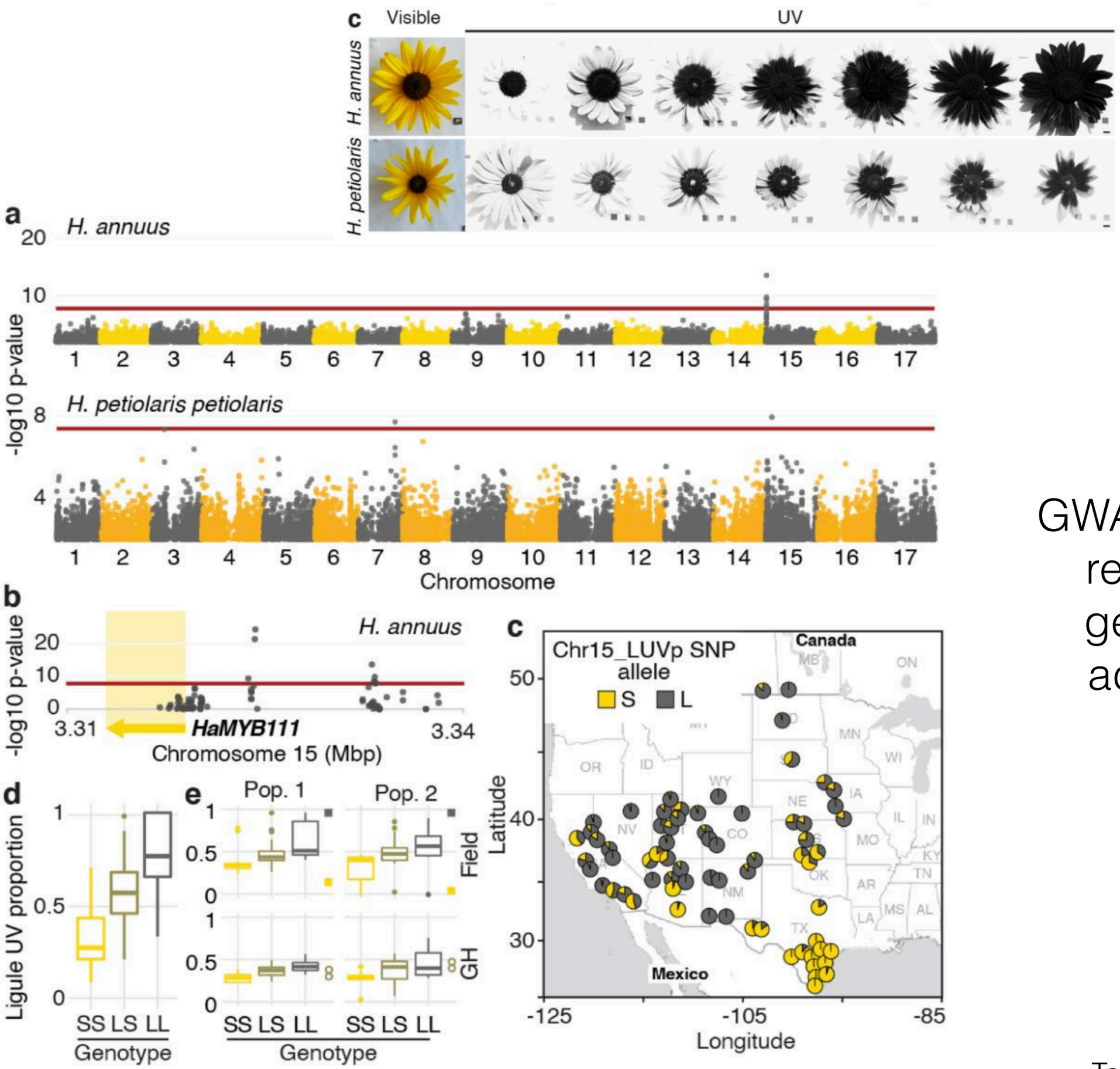
GWAS can help resolve the genetics of adaptation

Todesco et al., *bioRxiv*



GWAS can help resolve the genetics of adaptation

Todesco et al., *bioRxiv*

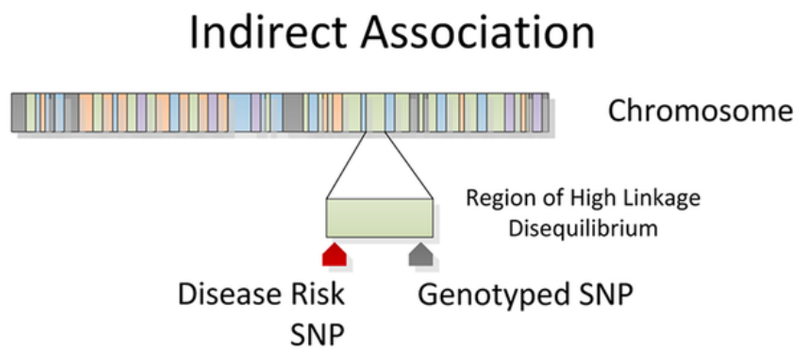


GWAS can help resolve the genetics of adaptation

Todesco et al., *bioRxiv*

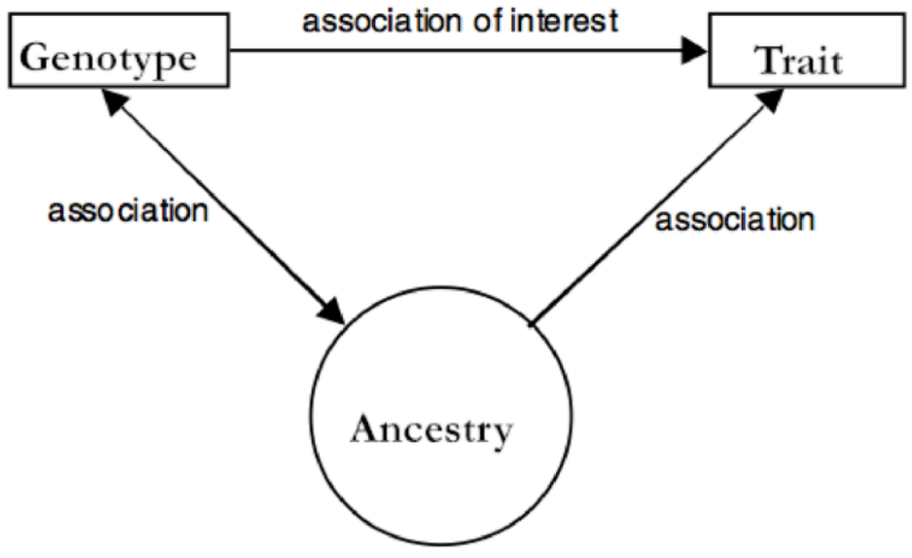
GWAS issues:

- Correlation != Causation



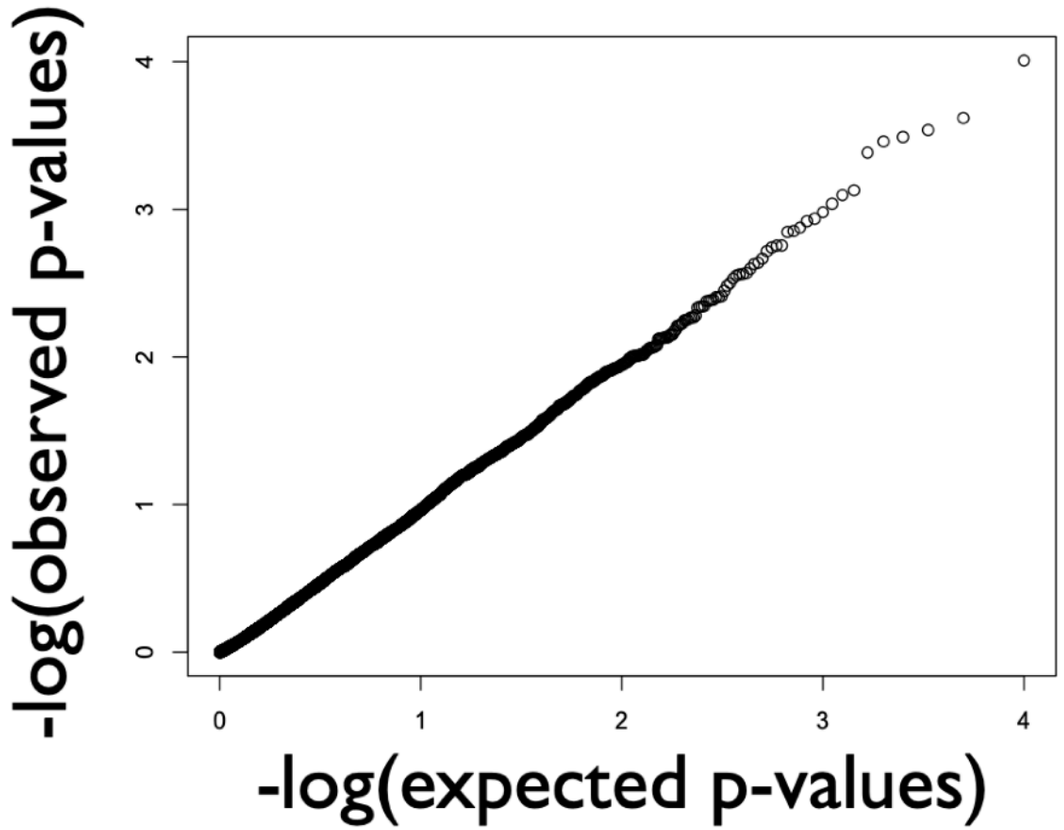
- Population structure

- Multiple testing



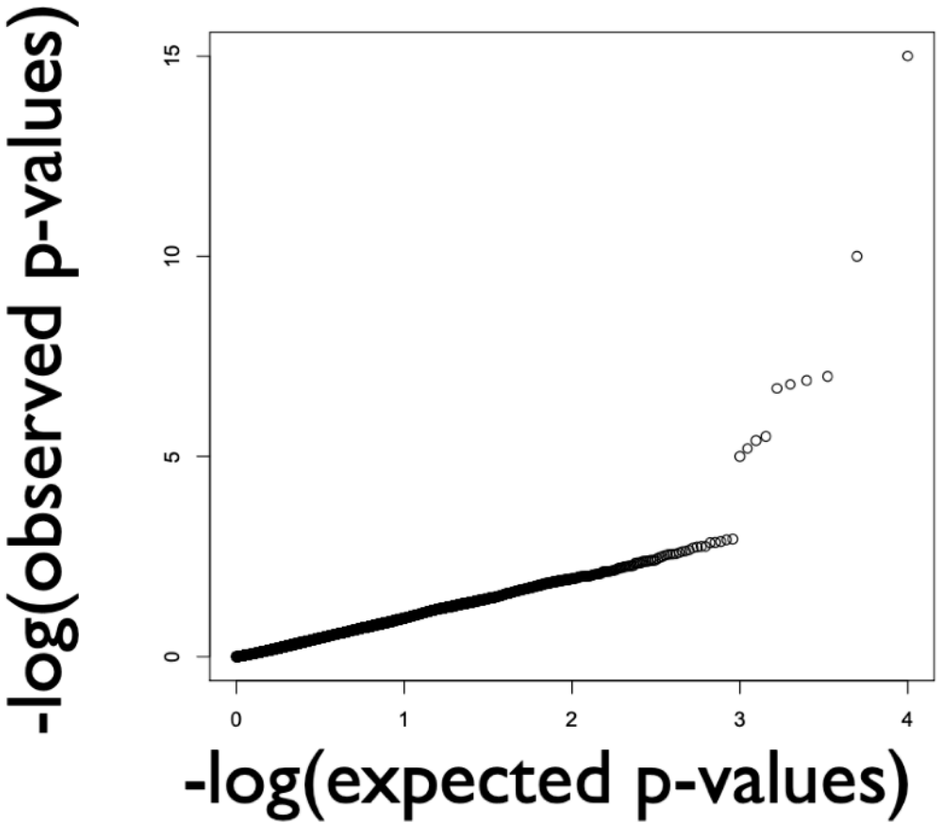
Evaluating your statistical GWA model: QQplots

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



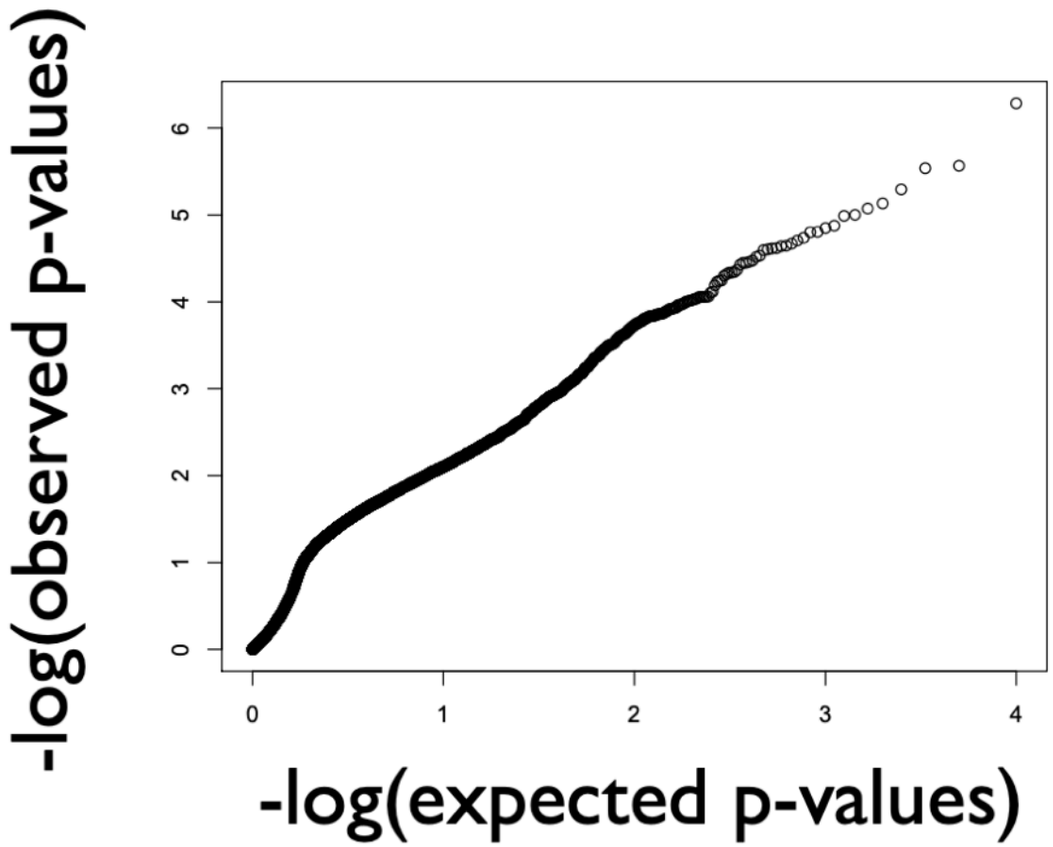
Evaluating your statistical GWA model: QQplots

- 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

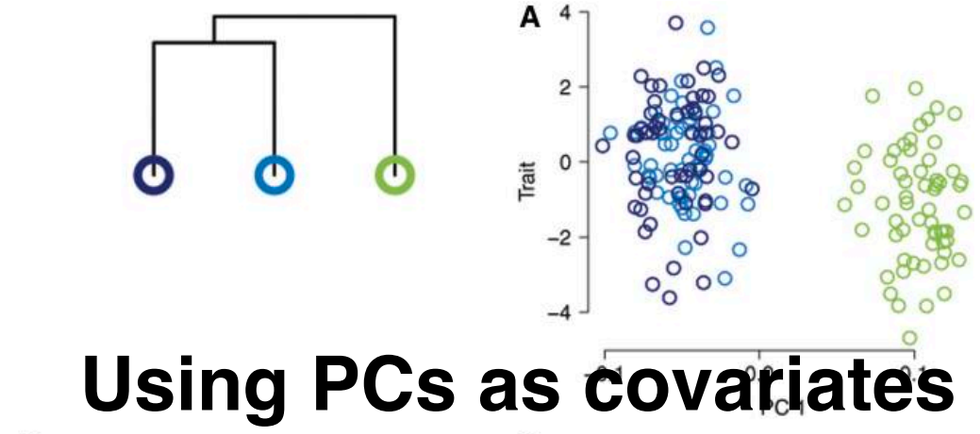
- In practice, you can find your QQ plot looks different than either the "null GWAS" case or the "ideal GWAS" case, for example:



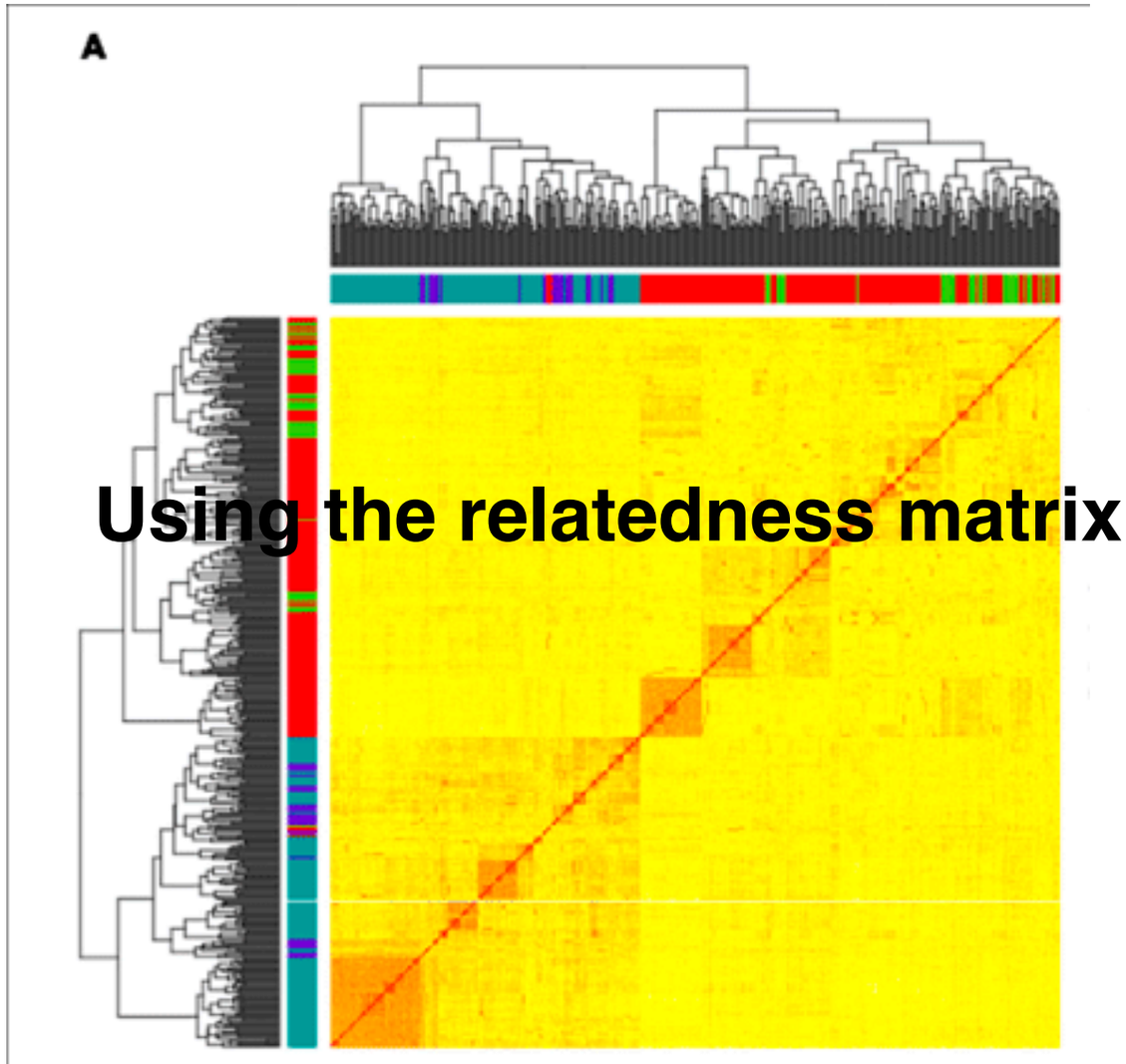
What might be driving this?

GWAS is a simple linear regression - and can easily be extended

… to control for population structure!

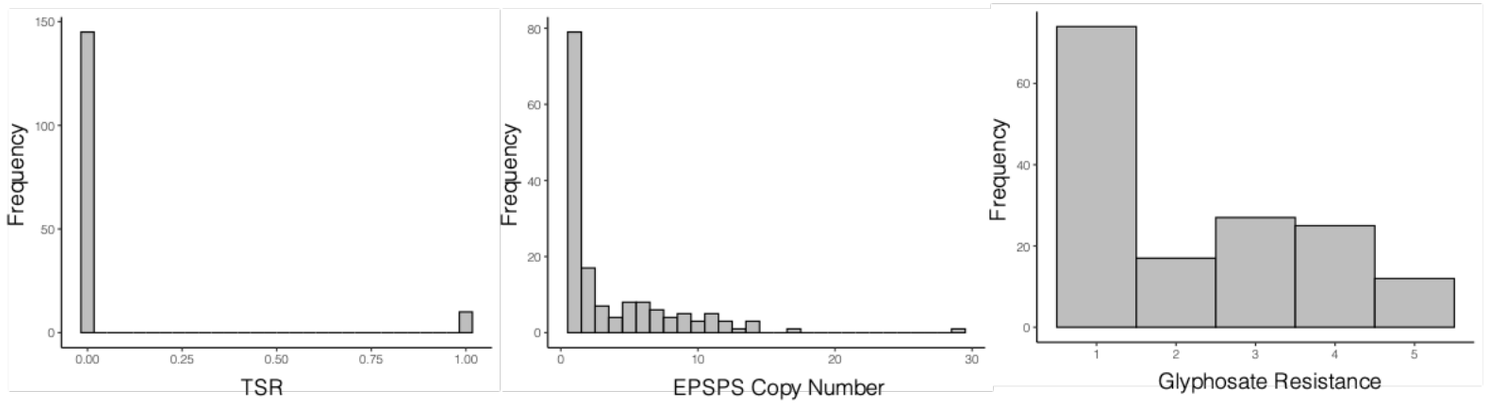


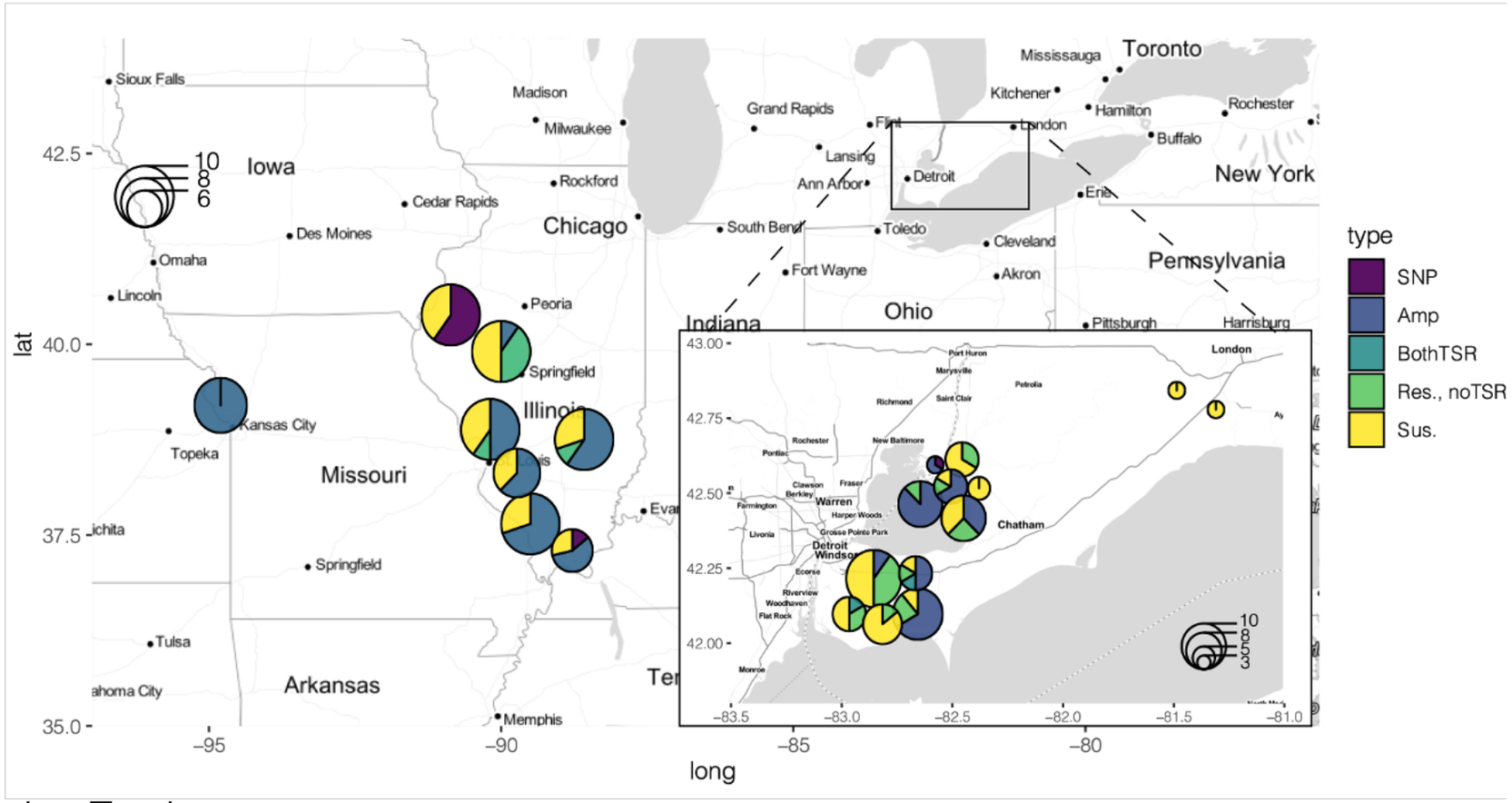
Using PCs as covariates



Using the relatedness matrix

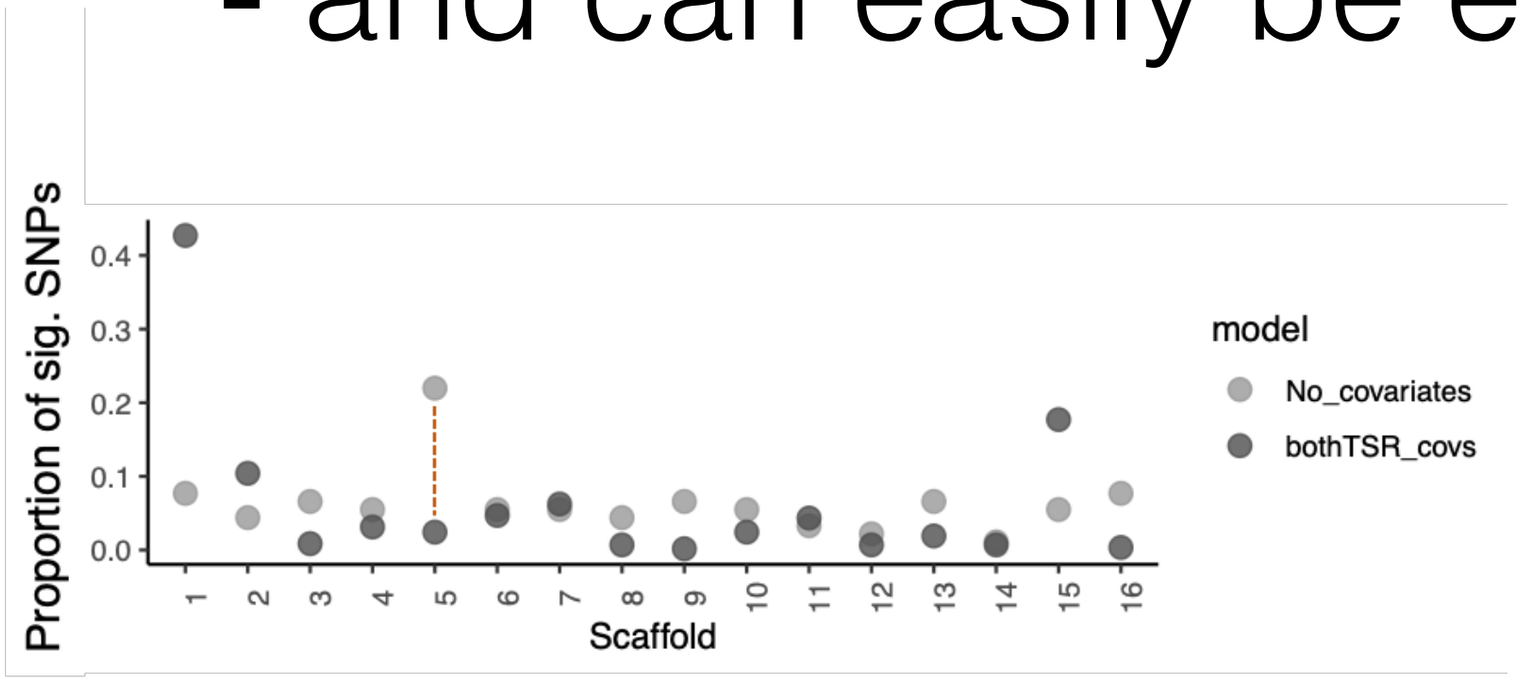
GWAS is a simple linear regression - and can easily be extended



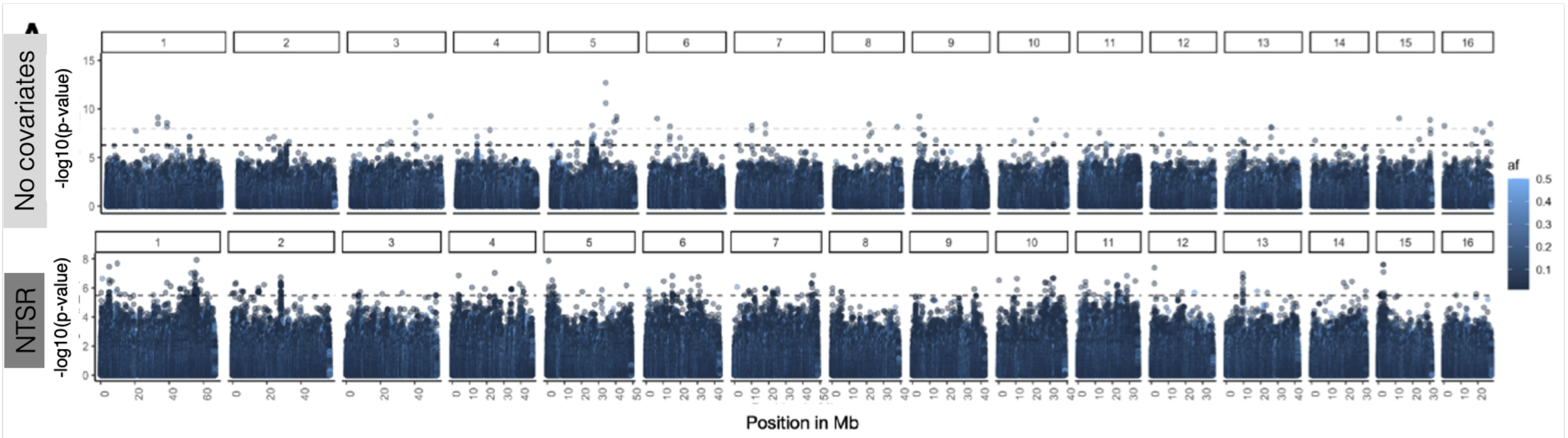


Kreiner 2021, *Molecular Ecology*

GWAS is a simple linear regression - and can easily be extended



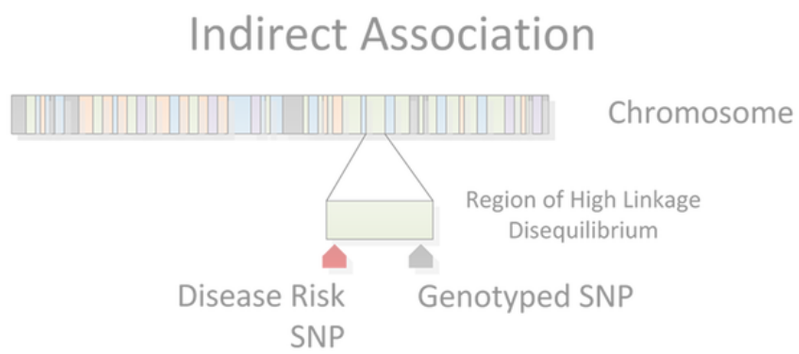
- i.e. Multiple regression
- Can also incorporate interaction effects etc.



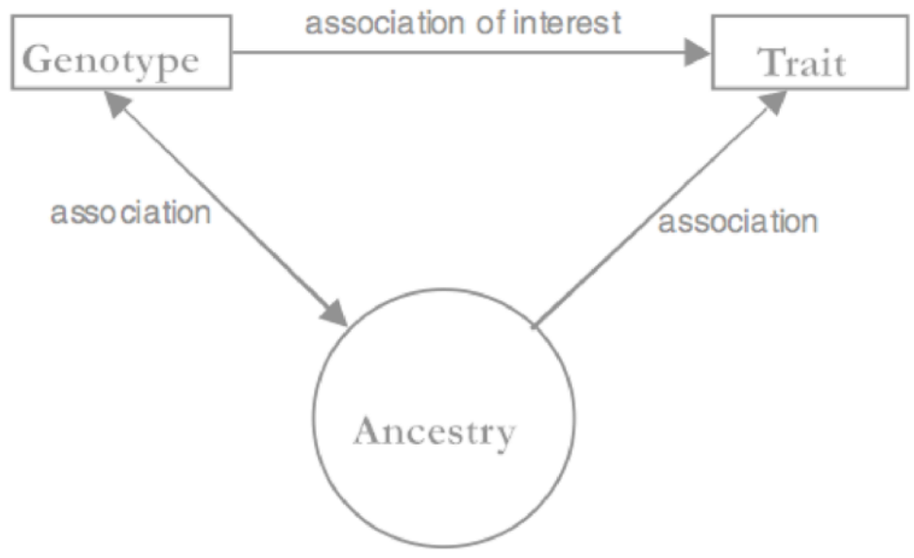
Kreiner 2021, *Molecular Ecology*

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:

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

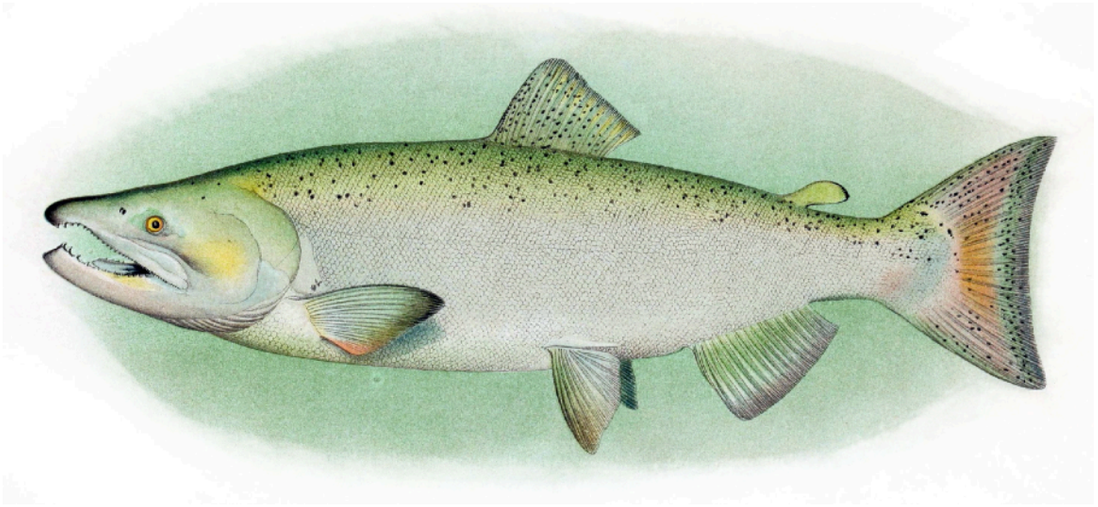
- We therefore use the Bonferroni Type I error to assess EACH of our *N* tests
- For example, if we have *N*=100 and we want an overall Type I error of 0.05, we require a test to have a p-value less than 0.0005 to be considered significant

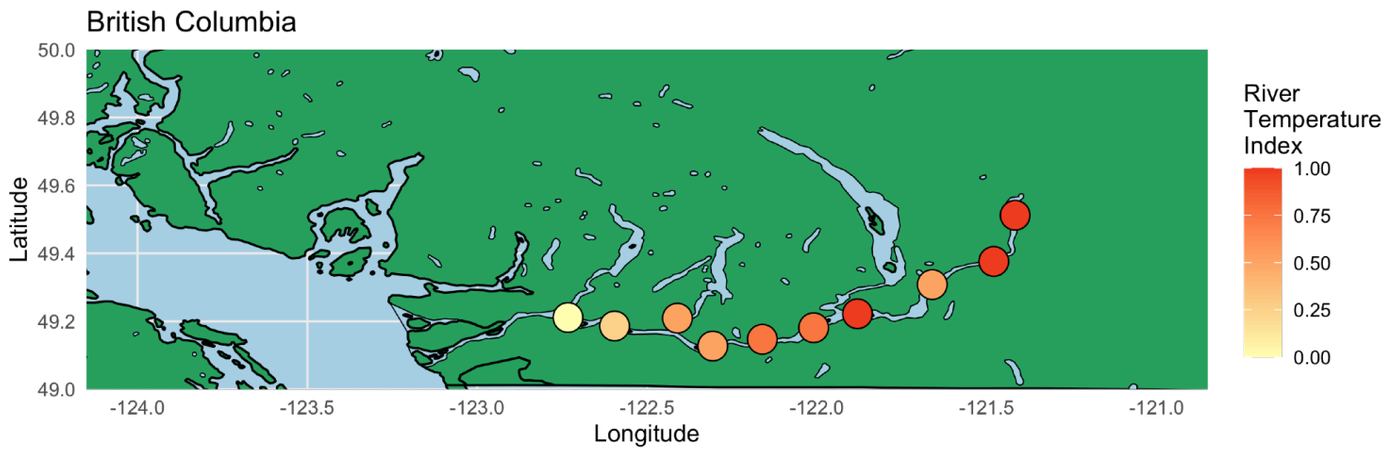
Multiple test correction: FDR

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

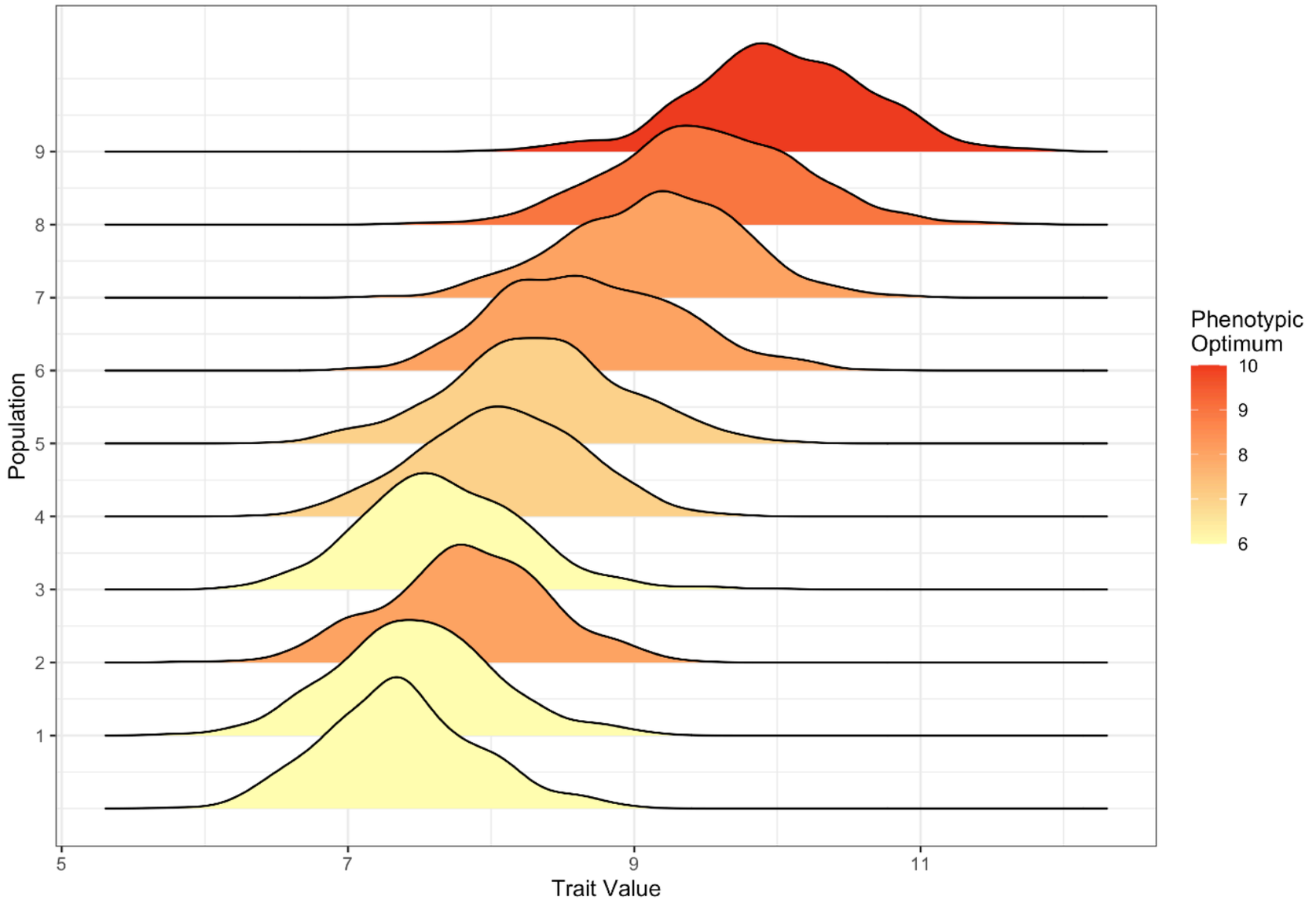
$$FDR = \frac{N * \alpha}{R}$$

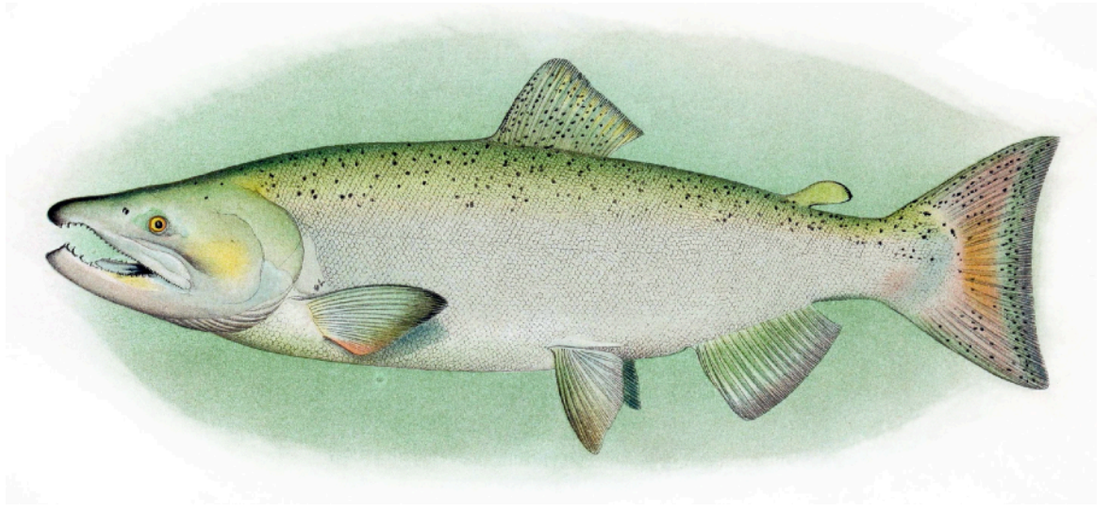
- Intuitively, the FDR is the proportion of cases where we reject the null hypothesis that are false positives
- We can estimate the FDR, e.g. say for *N*=100,000 tests and a Type I error of 0.05, we reject the null hypothesis 10,000 times, the FDR = 0.5

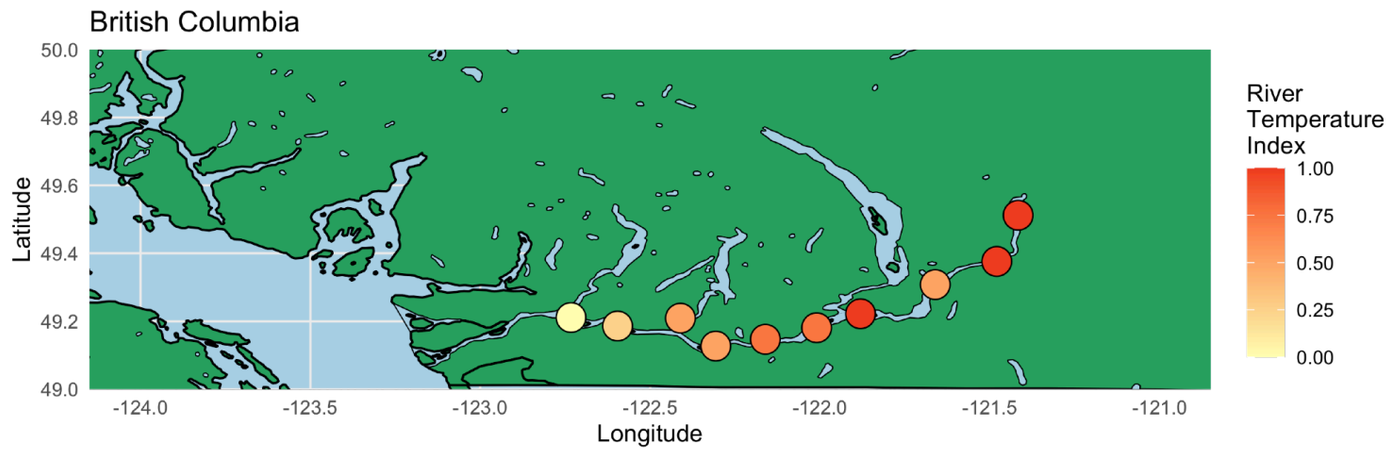




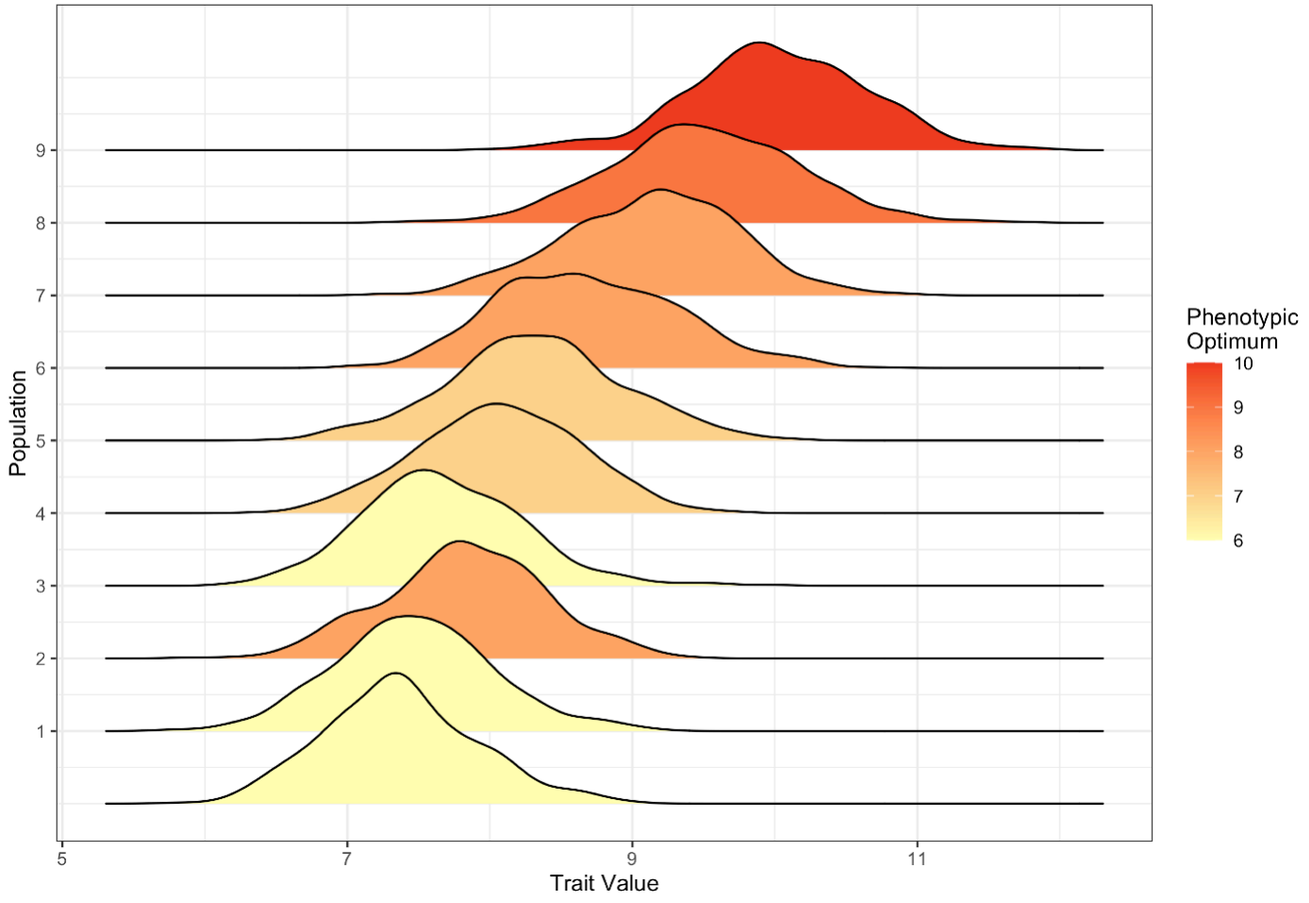
Which alleles and genes across the genome correlate with temperature tolerance?







Pop. 1 Pop. 2 Pop. 3 Pop. 4 Pop. 5 Pop. 6 Pop. 7 Pop. 8 Pop. 9 Pop. 10



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

Data formats

Wide format

ID	Product1	Product2	Product3	Product4
1	1	*NA*	1	1
2	1	1	*NA*	1
3	1	1	*NA*	*NA*
4	1	1	1	1

ggplot(data=…, aes(ID,product)

Long format

ID	Product	value
1	Product1	1
1	Product3	1
1	Product4	1
2	Product1	1
2	Product2	1
2	Product4	1
3	Product1	1
3	Product2	1
4	Product1	1
4	Product2	1
4	Product3	1
4	Product4	1

ggplot(data=…, aes(ID,value, group=Product)