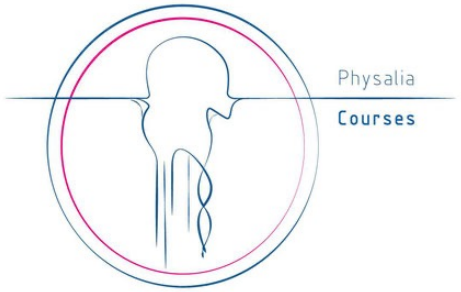


Population structure and demography

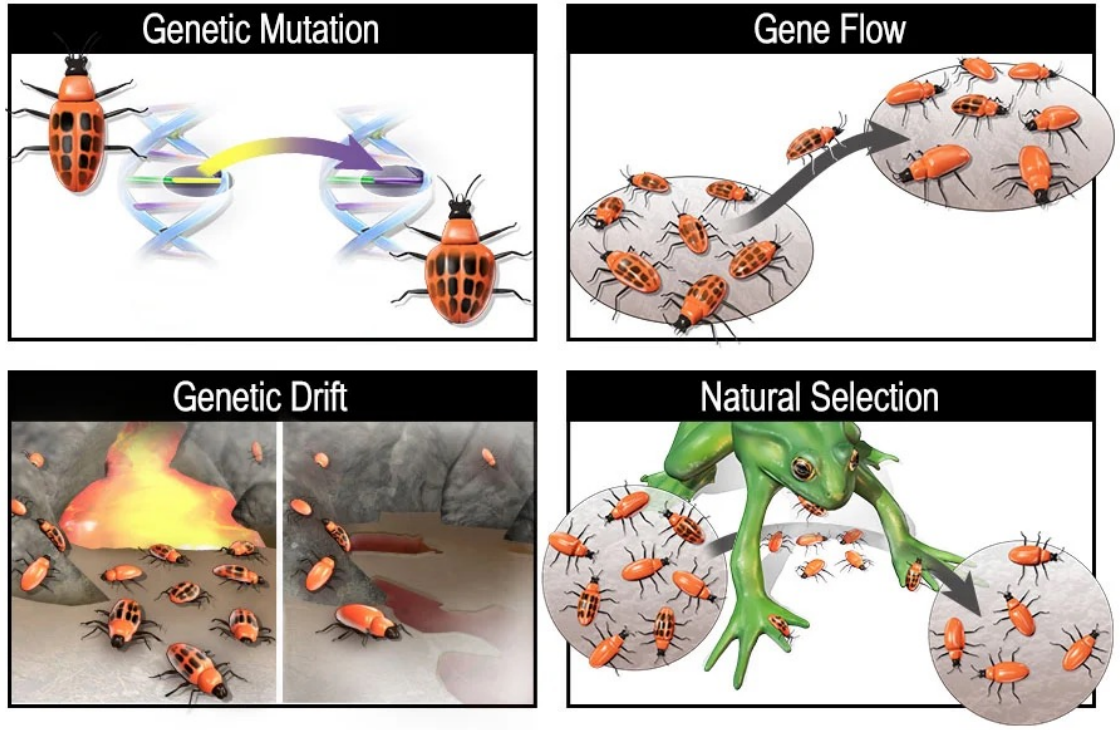
Day 2 – Lecture 1

(adapted from Claire Mérot & Anna Tigano's slides)

Why does population structure matter when studying adaptation?

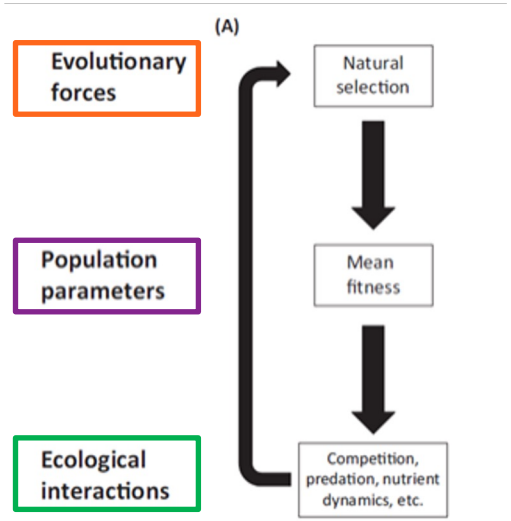
Why does population structure matter when studying adaptation?

- Evolution is the result of the interplay of different evolutionary forces



Why does population structure matter when studying adaptation?

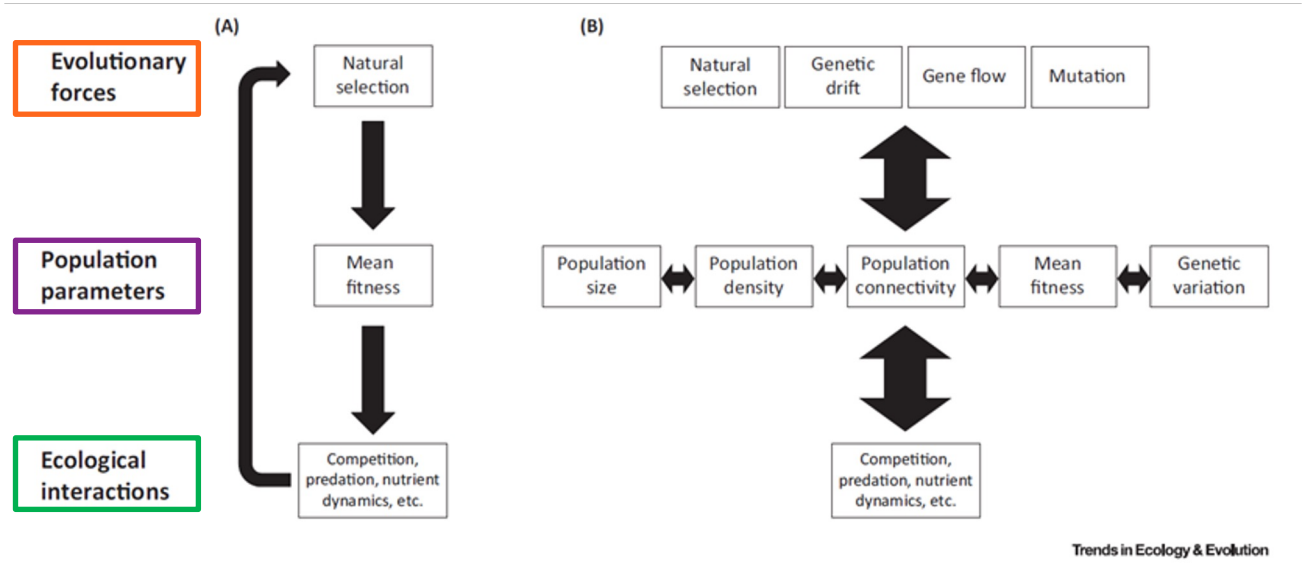
- Evolutionary, demographic and ecological processes are inseparable



Lowe, W. H., Kovach, R. P., & Allendorf, F. W. (2017). Population genetics and demography unite ecology and evolution. *Trends in Ecology & Evolution*, *32*(2), 141-152.

Why does population structure matter when studying adaptation?

- Evolutionary, demographic and ecological processes are inseparable



Lowe, W. H., Kovach, R. P., & Allendorf, F. W. (2017). Population genetics and demography unite ecology and evolution. *Trends in Ecology & Evolution*, *32*(2), 141-152.

Why does population structure matter when studying adaptation?

- Complementary objectives

Study	Selection and adaptation	Demographic history and population structure

Why does population structure matter when studying adaptation?

- Complementary objectives

Study	Selection and adaptation	Demographic history and population structure
Focus	• On (putatively) adaptive loci	

Why does population structure matter when studying adaptation?

- Complementary objectives

Study	Selection and adaptation	Demographic history and population structure
Focus	• On (putatively) adaptive loci	
Use	• Study ecological/functional diversity • Understand adaptative processes under divergent or balancing selection • Identify candidate genes	

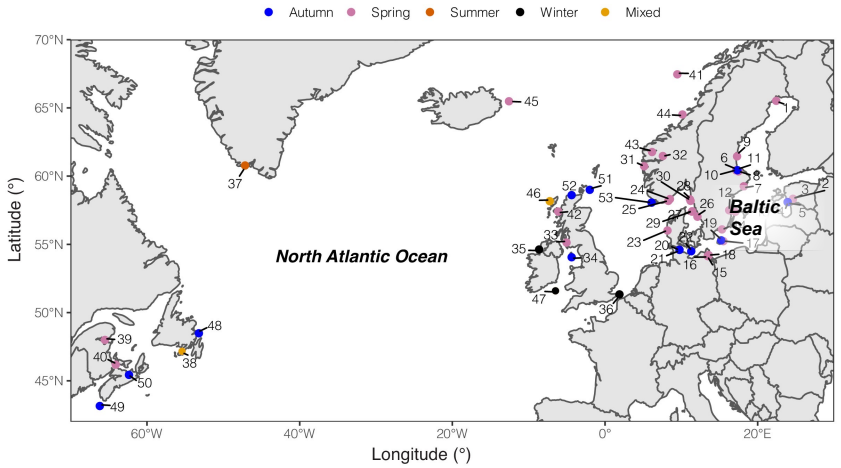
Why does population structure matter when studying adaptation?

- Complementary objectives

Study	Selection and adaptation	Demographic history and population structure
Focus	• On (putatively) adaptive loci	• On neutral loci
Use	• Study ecological/functional diversity • Understand adaptative processes under divergent or balancing selection • Identify candidate genes	• Understand the demographic history of populations • Describe population connectivity • Assess general genetic diversity

Different loci could tell a different story (neutral vs. outlier loci)

Atlantic herring
(*Clupea harengus*)

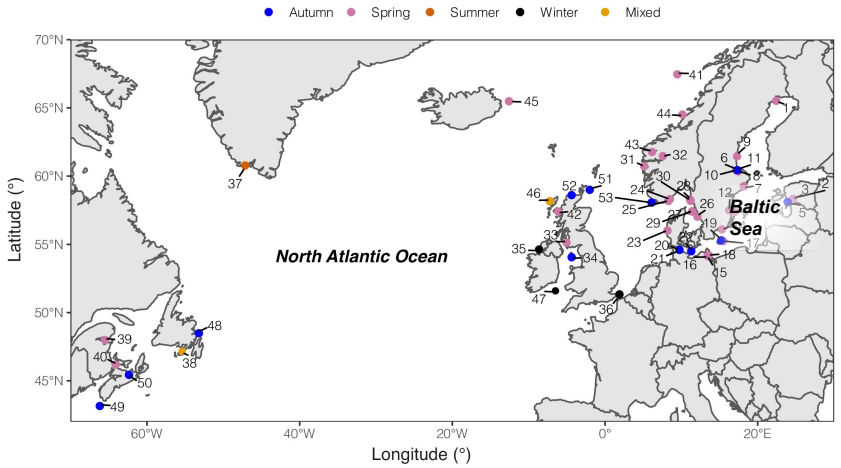


- Pool-seq data of 53 populations
- 11 million SNPs

Han et al. (2020) Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. *eLife* 9:e61076. https://doi.org/10.7554/eLife.61076

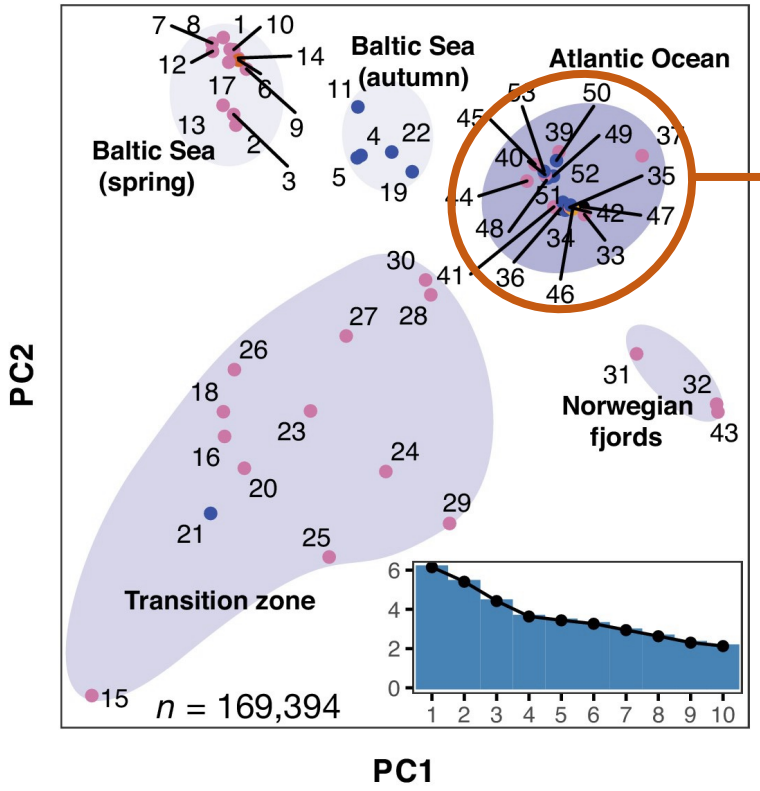
Different loci could tell a different story (neutral vs. outlier loci)

Atlantic herring
(*Clupea harengus*)



- Pool-seq data of 53 populations
- 11 million SNPs

Neutral markers

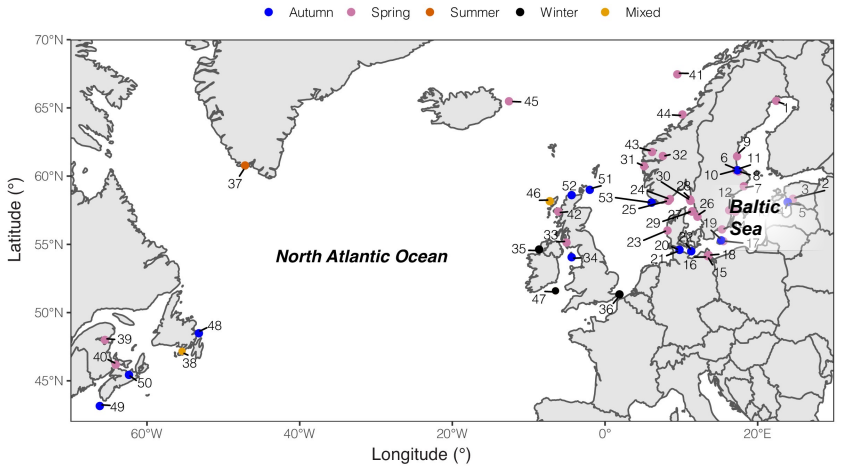


Atlantic populations (largest biomass) collapses into a single cluster

Han et al. (2020) Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. *eLife* 9:e61076. https://doi.org/10.7554/eLife.61076

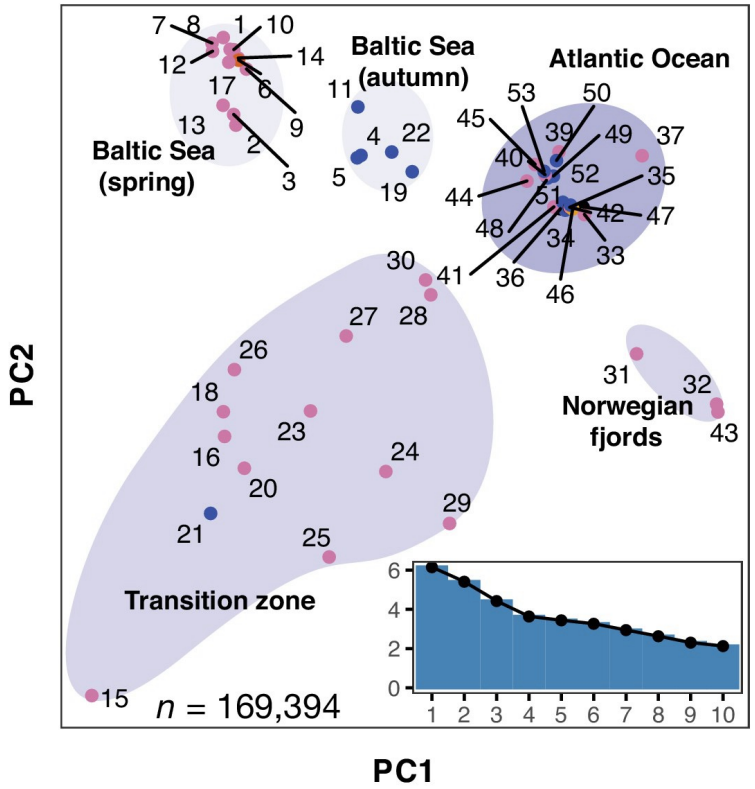
Different loci could tell a different story (neutral vs. outlier loci)

Atlantic herring
(*Clupea harengus*)

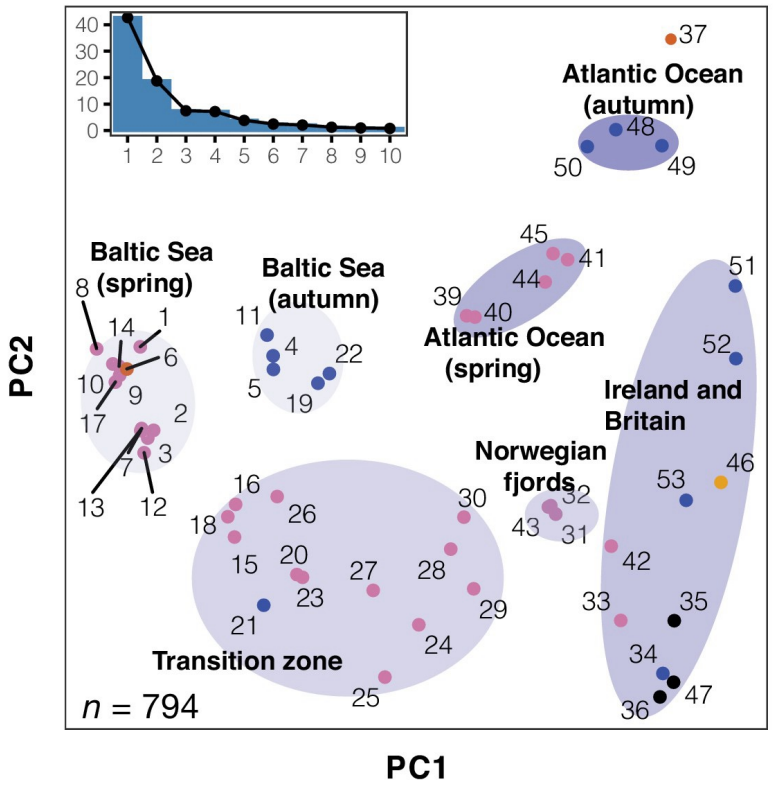


- Pool-seq data of 53 populations
- 11 million SNPs

Neutral markers



Outlier markers

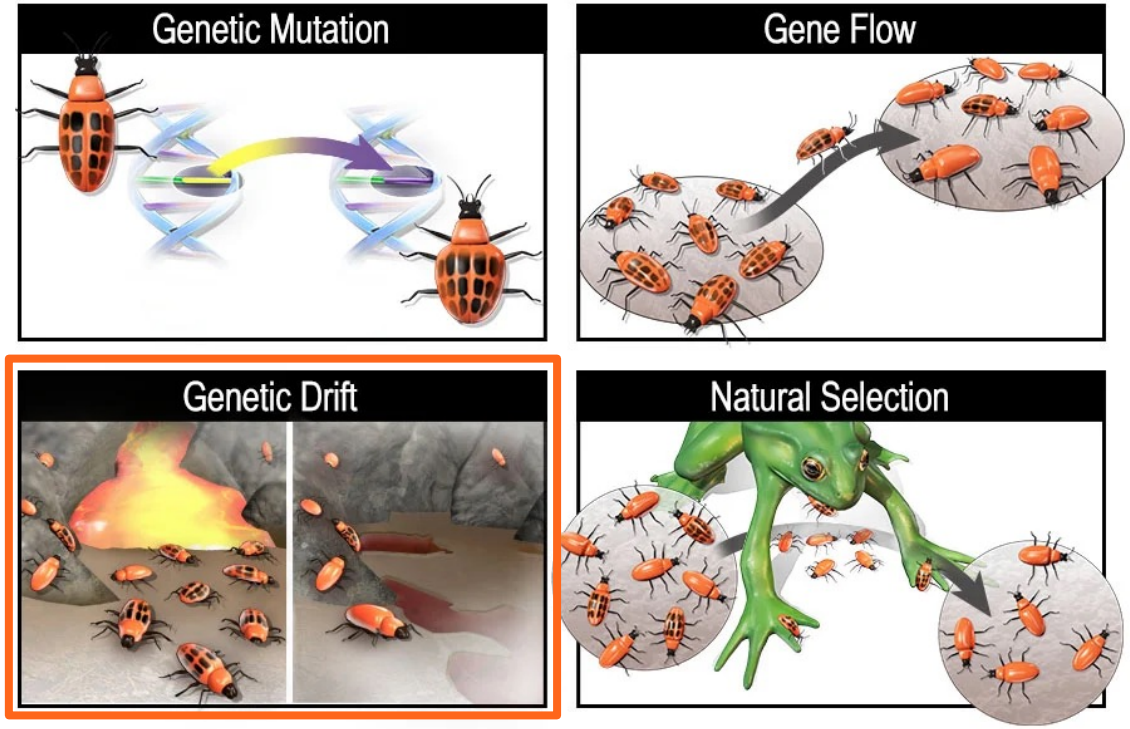


Greater segregation of populations. Clustering by differences in salinity and spawning season

Han et al. (2020) Ecological adaptation in Atlantic herring is associated with large shifts in allele frequencies at hundreds of loci. *eLife* 9:e61076. https://doi.org/10.7554/eLife.61076

Why does population structure matter when studying adaptation?

- Evolution is the result of the interplay of different evolutionary forces

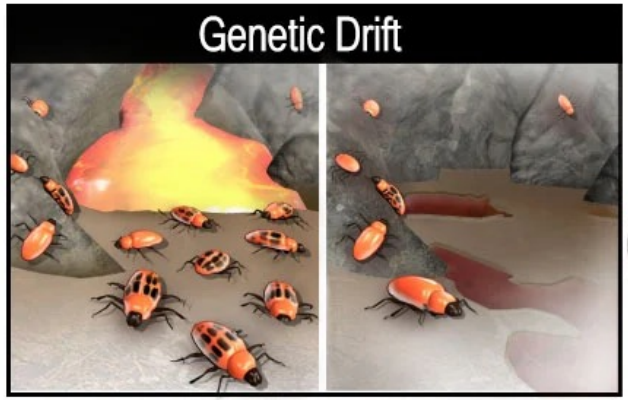


Genetic drift

- Variation in allele frequency due to random processes
- Stronger effect on smaller populations
- It can cause the loss or fixation of a genetic variant due to random sampling of alleles
- One of the main drivers of genetic population structure

↓

Genetic drift can generate a genetic footprint similar to that of selection!



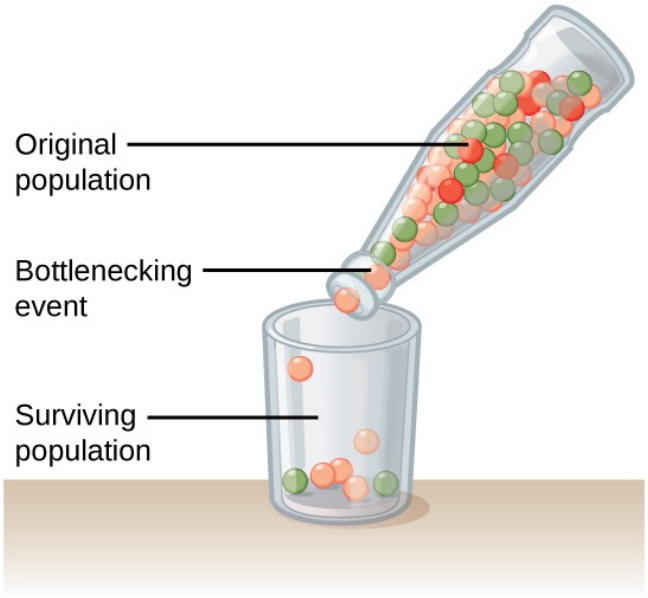
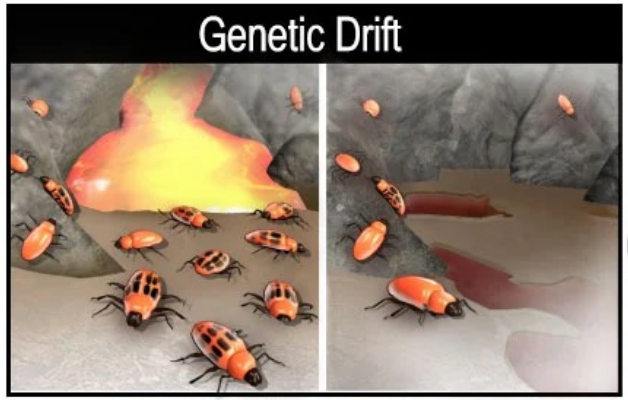
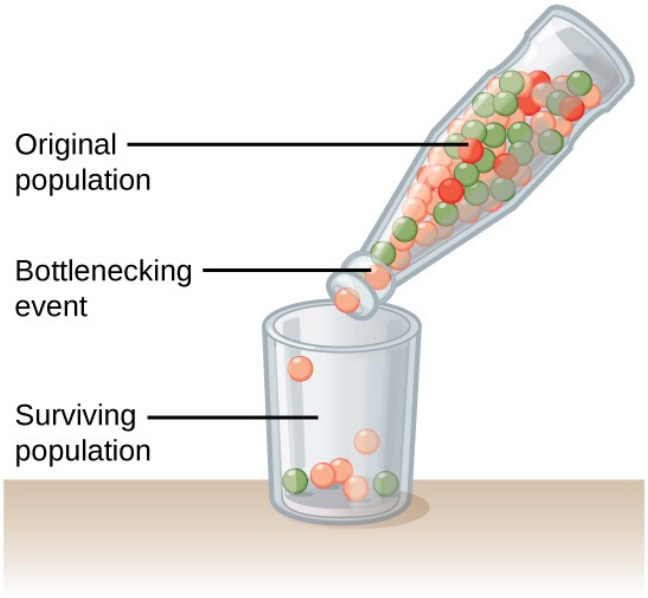


Image credit: "Population genetics: Figure 3," by OpenStax College, Biology, CC BY 3.0.

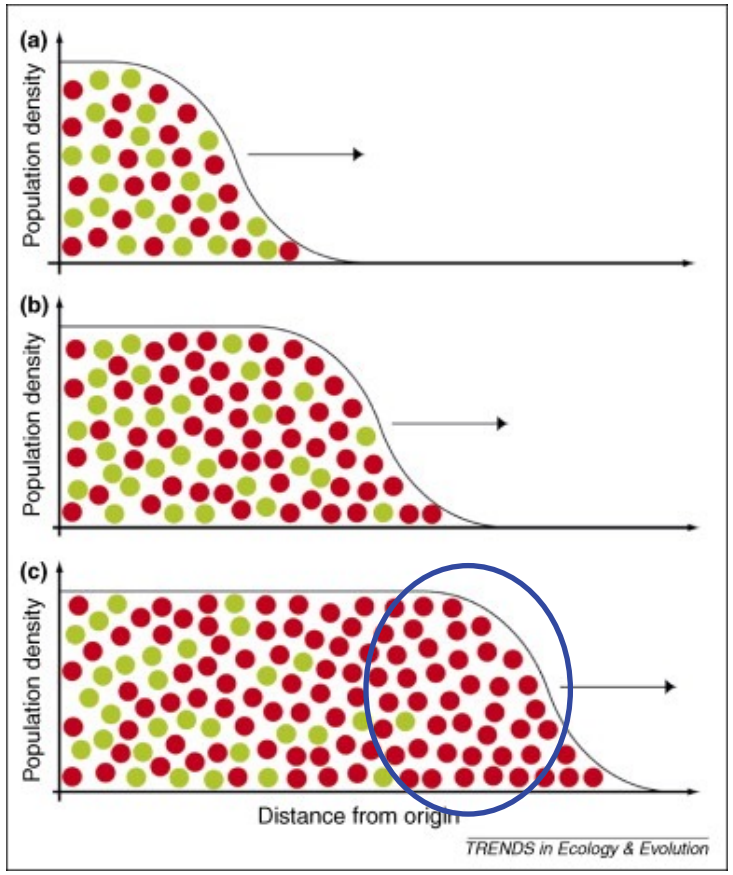
Demographic events to consider

- Allele surfing
- Spatial autocorrelation
- Isolation-by-distance (IBD)
- Secondary contact of different lineages





Allele surfing



Surfing during population expansions promotes genetic revolutions and structuration. Excoffier & Ray TREE 2008 https://doi.org/10.1016/j.tree.2008.04.004

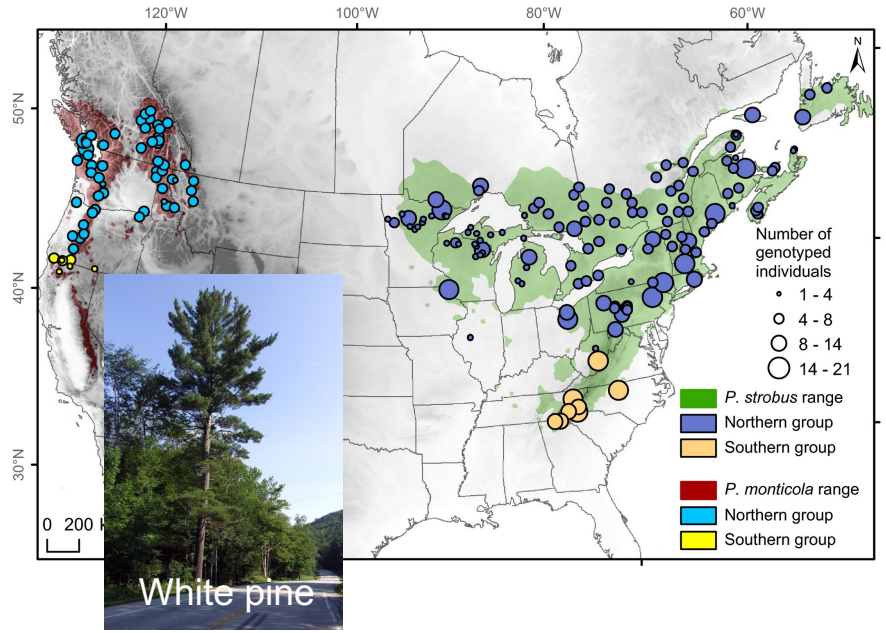
- Populations on the leading edge of a range expansion are often small
- Individuals in those populations may contribute disproportionately to the propagating wave of expansion

↓

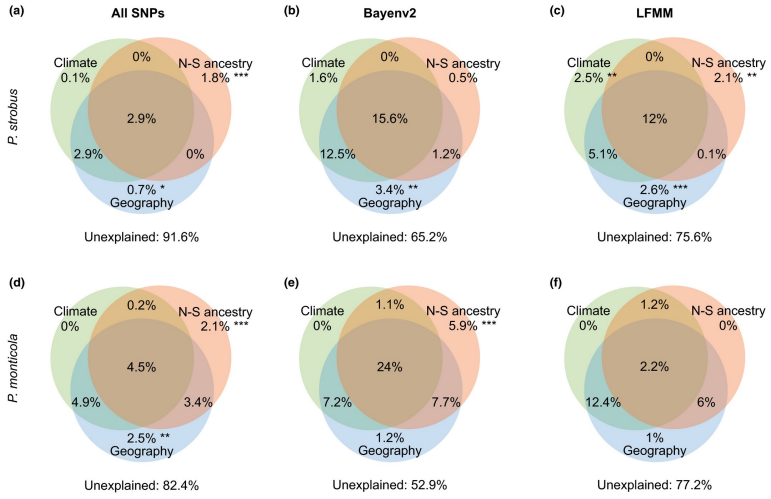
their alleles can easily reach high frequencies and spread over large areas, even in the absence of selection

Spatial autocorrelation

- Correlation between environmental variation & geographic distances (e.g. climatic clines)
- Nearby locations are not statistically independent
  - Strong correlation between neutral alleles and environmental variables are more likely to occur by chance than expected with some null models



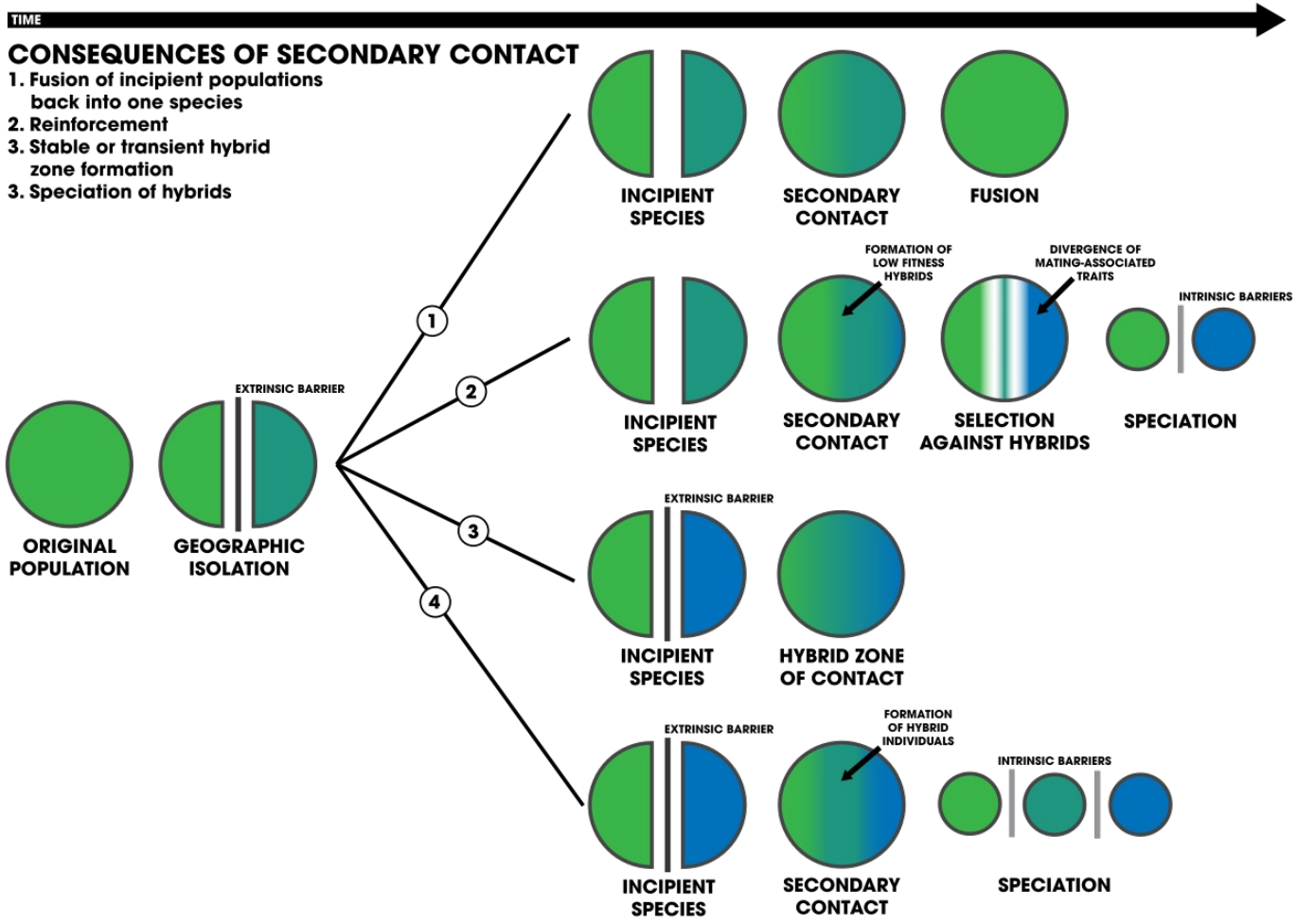
Isolation-by-distance or adaptation along a gradient, or both?



The challenge of separating signatures of local adaptation from those of isolation by distance and colonization history: The case of two white pines. Nadeau et al, 2016 https://doi.org/10.1002/ece3.2550

Contact between different lineages / hybridization

- Signatures of selection or of local adaptation are best detected in a context of (high) gene flow and large population sizes, when the effect of genetic drift is low
- Any substructure (lineages, secondary contact, admixed populations) should be taken into account



By Andrew Z. Colvin - Own work, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=69254578

Some “philosophical” questions…

Which population genetic patterns result from:

- the balance between selection / migration?
- the balance between genetic drift / migration?
- secondary contact of divergent lineages?
- past range expansion out of glacial refugees?

What is indicative of adaptation and what is the result of demographic history?

How to characterize population structure?

Unsupervised methods:

- PCA

Semi-supervised methods (*K* = number of expected genetic clusters)

- Bayesian clustering

Supervised methods (e.g., with location information)

- DAPC
- Pairwise F_{ST} (between pairs of populations)

How to characterize population structure?

Unsupervised methods:

- PCA

Principal Component Analysis (PCA)

- Statistical tool that reduces matrix complexity by identifying the eigenvectors and ordering them
- For genetic data, the fist PCs reflect axis of genetic variation along which individuals with same ancestry, or exchanging genetic material, are more similar to each other and appear closer in a PCA plot
- **Caution:** can be strongly driven by few loci in linkage disequilibrium (LD)

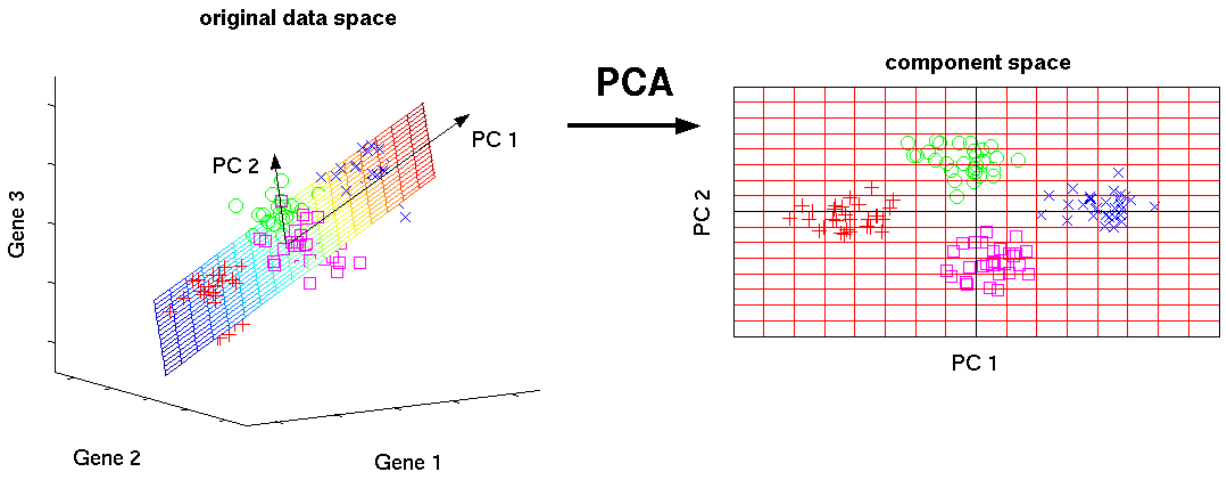


Image credit: http://phdthesis-bioinformatics-maxplanckinstitute-molecularplantphys.matthias-scholz.de/

Principal Component Analysis (PCA)

Recommendations for population structure purposes:

- Compare PCA on all SNPs vs. PCA on LD-pruned SNPs
- Look at loadings of the PCs (which fraction of the genome explains PC1? Explains PC2? Etc.)

Be careful when interpreting a PCA plot. When genetic variance is high, PC1 might explain less than 1% of the total variance, but in other cases PC1 can capture 20-50% of the total genetic variance, it depends on the dataset!

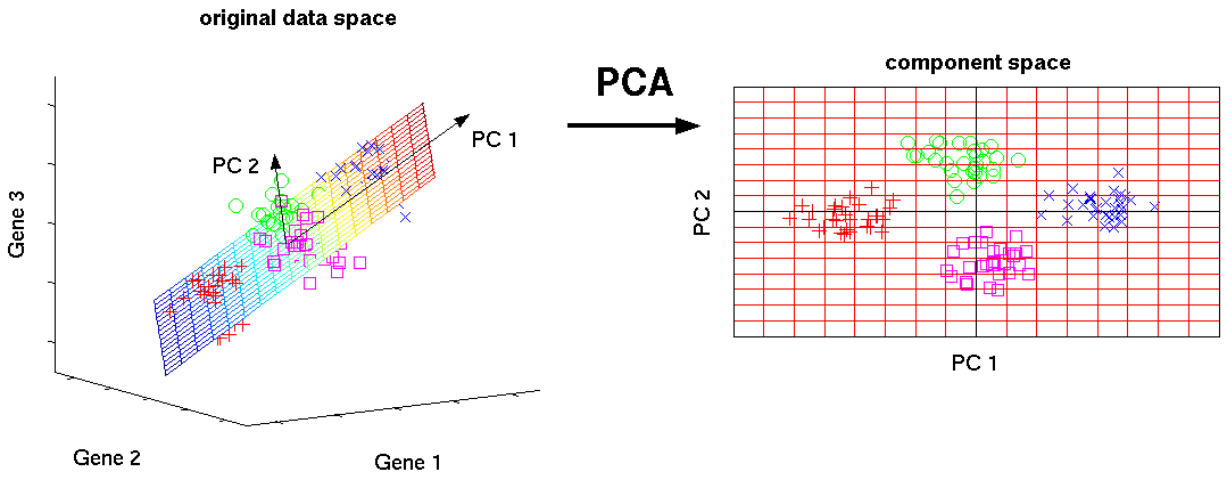
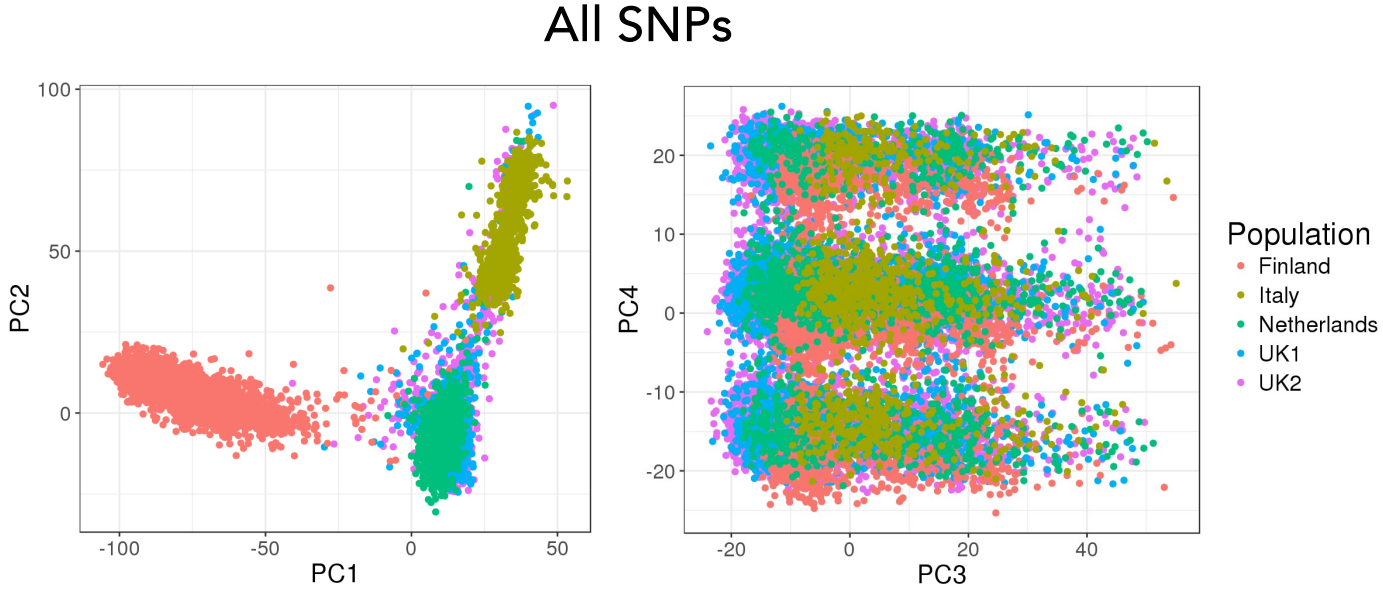
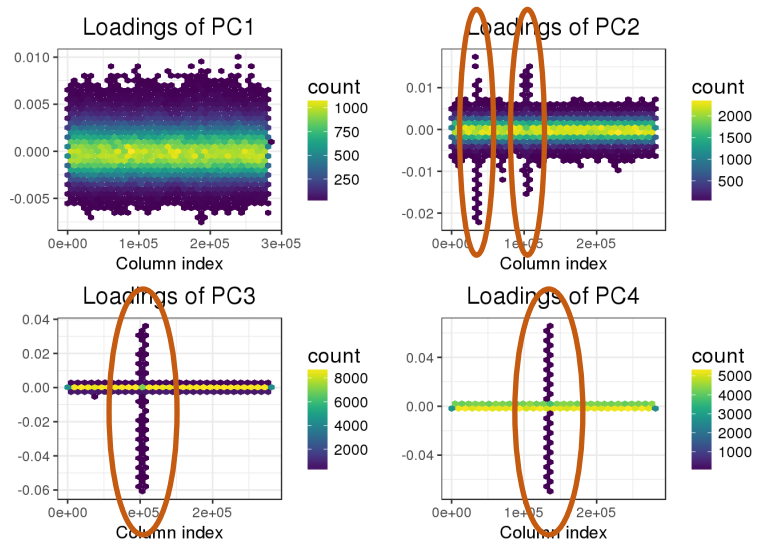


Image credit: http://phdthesis-bioinformatics-maxplanckinstitute-molecularplantphys.matthias-scholz.de/

Principal Component Analysis (PCA)

- Each individual is a point with coordinates along all PCs (scores)
- Each genetic marker contributes to all PCs with a different strength (loadings)

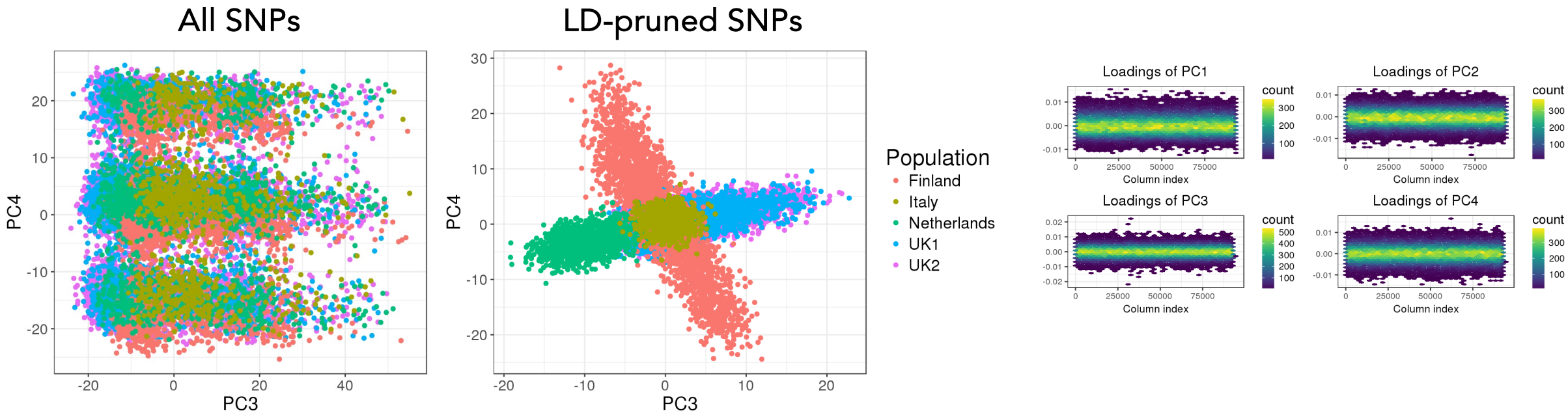




Images from tutorial on PCA for population genomics developed by Florian Privé
https://privefl.github.io/bigsnpr/articles/how-to-PCA.html

Principal Component Analysis (PCA)

- Each individual is a point with coordinates along all PCs (scores)
- Each genetic marker contributes to all PCs with a different strength (loadings)



Images from tutorial on PCA for population genomics developed by Florian Privé
https://privefl.github.io/bigsnpr/articles/how-to-PCA.html

How to characterize population structure?

Unsupervised methods:

- PCA

Semi-supervised methods (*K* = number of expected genetic clusters)

- Bayesian clustering

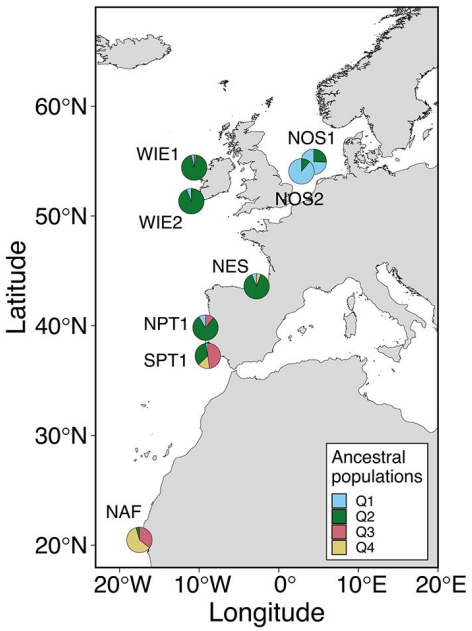
Bayesian clustering (STRUCTURE, etc..)

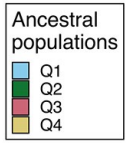
Atlantic horse Mackerel
(*Trachurus trachurus*)

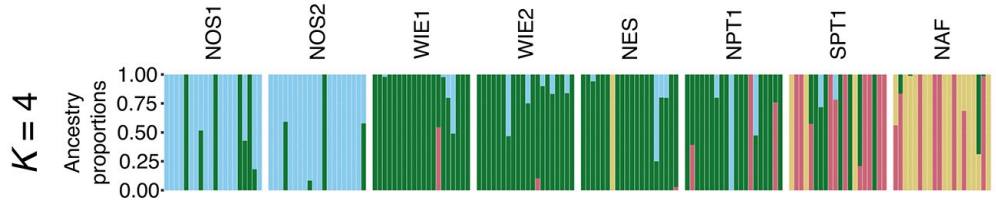
- Aims to sort individuals into *K* clusters so as to minimize departures from Hardy-Weinberg equilibrium and linkage equilibrium
- Caution: can be strongly driven by few loci in linkage disequilibrium
- Admixture or FastSTRUCTURE replaces STRUCTURE for genome-wide data
- Evaluate the fit of the model

Recommendations for population structure purpose:

- Compare results on all SNPs vs. on LD-pruned SNPs
- Explore many values of *K* and report likelihood

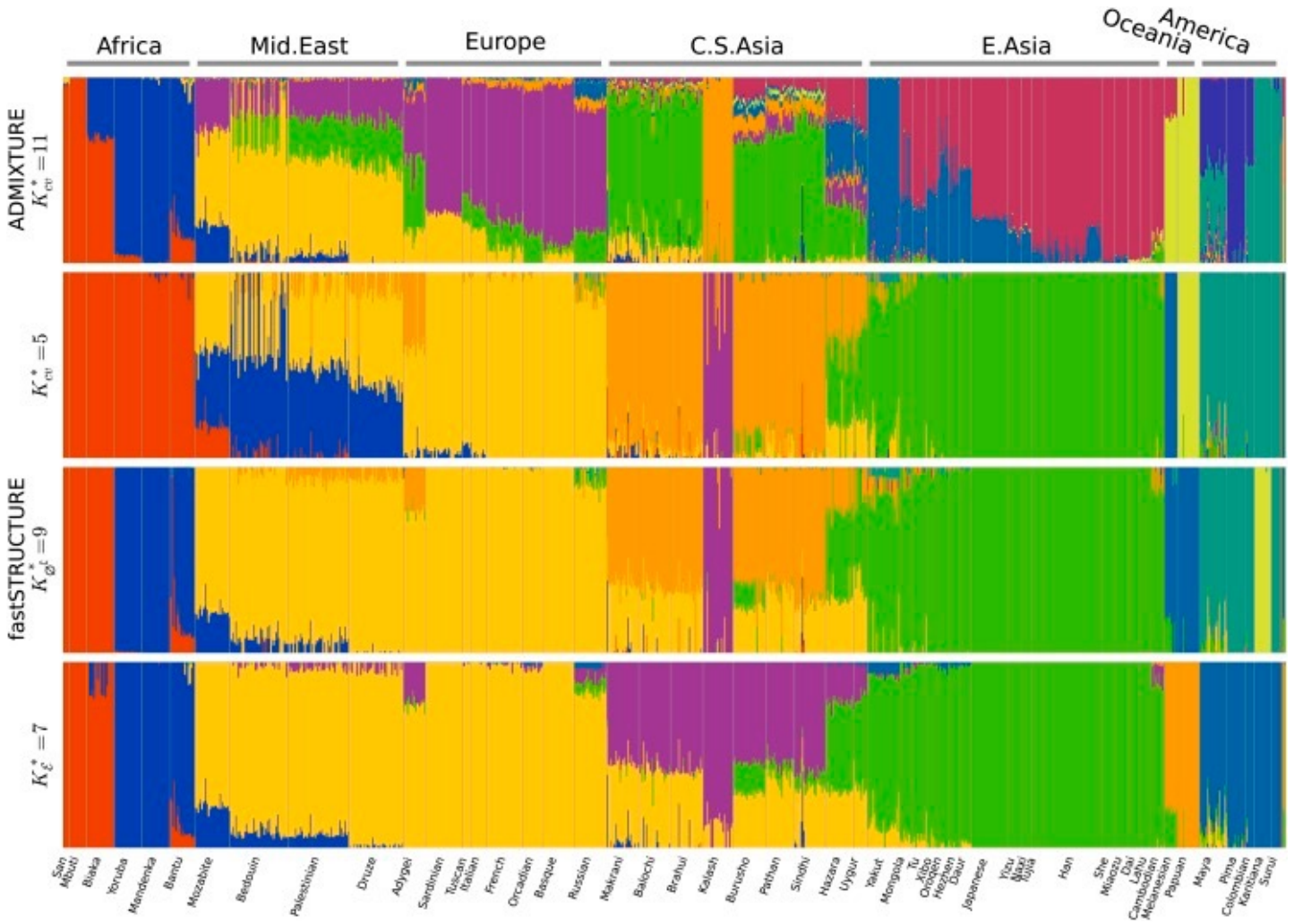






Fuentes-Pardo, A. P., Farrell, E. D., Pettersson, M. E., Sprehn, C. G., & Andersson, L. (2023). The genomic basis and environmental correlates of local adaptation in the Atlantic horse mackerel (*Trachurus trachurus*). *Evolutionary Applications*, 16, 1201–1219. https://doi.org/10.1111/eva.13559

Bayesian clustering (STRUCTURE, etc..)



Each individual is represented by a thin vertical line that is partitioned into *K* colored segments according to its membership coefficients in *K* clusters.

fastSTRUCTURE: variational inference of population structure in large SNP data sets 2014 Genetics

Anil Raj[1], Matthew Stephens[2], Jonathan K Pritchard[3]

10.1534/genetics.114.164350

How to characterize population structure?

Unsupervised methods:

- PCA

Semi-supervised methods (*K* = number of expected genetic clusters)

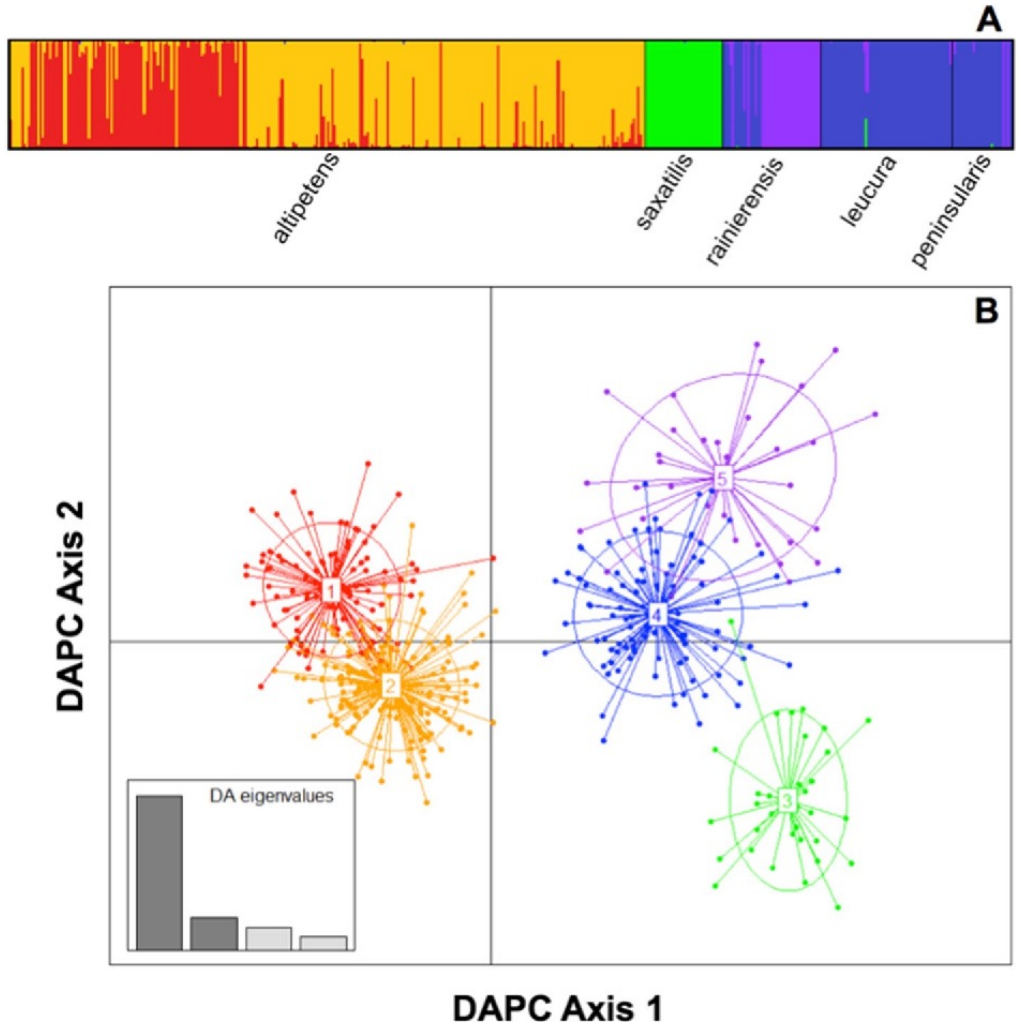
- Bayesian clustering

Supervised methods (e.g., with location information)

- DAPC
- Pairwise F_{ST} (between pairs of populations)

DAPC (discriminant PCA)

- A mix of a discriminant analysis and a PCA
- It will try very hard to find axis of variation that discriminate the groups given *a priori*
- Caution: Biased analysis when there is a much larger number of markers (SNPs) than groups (populations). Try to avoid over-fitting and don't overinterpret the output



Languin et al, 2018. Conservation genomics

Miller, J.M., Cullingham, C.I. & Peery, R.M. The influence of a priori grouping on inference of genetic clusters: simulation study and literature review of the DAPC method. *Heredity* (2020). https://doi.org/10.1038/s41437-020-0348-2

Pairwise F_{ST}

- F_{ST} is the proportion of the total genetic variance contained in a subpopulation relative to the total genetic variance. Values can range from 0 (low) to 1 (high) differentiation
- Pairwise F_{ST} is a measure of genetic distance between all pairs of populations
- To infer neutral population structure, likely better on LD-pruned SNPs

	WIE1	WIE2	NOS1	NOS2	NPT1	SPT1	NAF	NES	MED
WIE1	NA	0.004	0.0044	0.0048	0.0053	0.0065	0.0066	0.0034	0.0109
WIE2	0.004	NA	0.0042	0.0046	0.005	0.0063	0.0062	0.0035	0.0105
NOS1	0.0044	0.0042	NA	0.0012	0.0054	0.0068	0.0068	0.0038	0.0112
NOS2	0.0048	0.0046	0.0012	NA	0.0058	0.0072	0.007	0.0041	0.0116
NPT1	0.0053	0.005	0.0054	0.0058	NA	0.0074	0.0071	0.0047	0.0118
SPT1	0.0065	0.0063	0.0068	0.0072	0.0074	NA	0.0062	0.0054	0.0126
NAF	0.0066	0.0062	0.0068	0.007	0.0071	0.0062	NA	0.005	0.012
NES	0.0034	0.0035	0.0038	0.0041	0.0047	0.0054	0.005	NA	0.0101
MED	0.0109	0.0105	0.0112	0.0116	0.0118	0.0126	0.012	0.0101	NA

Atlantic horse Mackerel
(*Trachurus trachurus*)

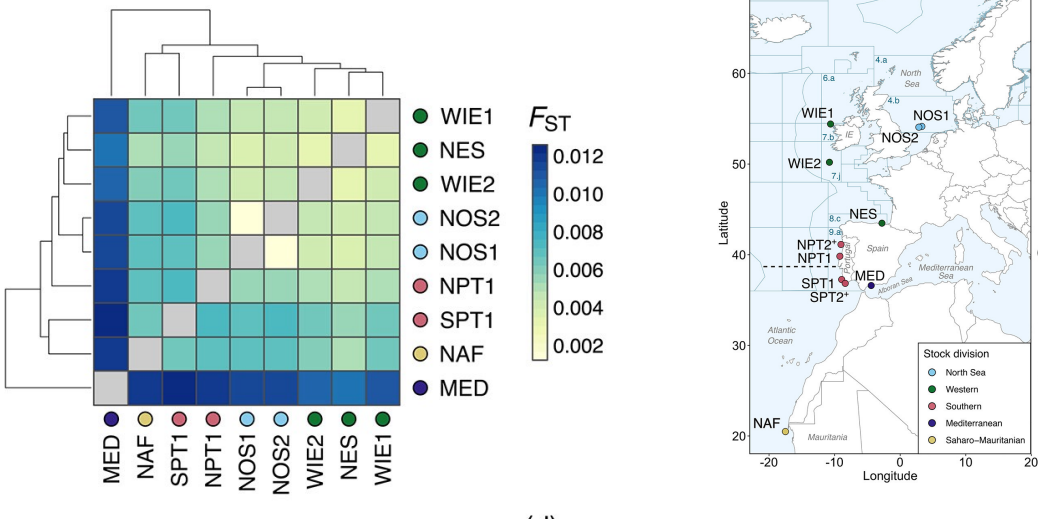
Pairwise F_{ST}

Absolute values are informative, traditionally researchers have used these ranks for interpretation:

- 0.000x = high gene flow (don't bother too much about looking for structure)
- 0.01-0.1 = consider population structure
- >0.1 = do you really have one species?

Caution: F_{ST} varies depending on the type of data you have (RRS, or WGS, number of markers, composition –neutral and/or outliers) and the life history of your species, so be careful about these value ranges, don't take them as golden run

	WIE1	WIE2	NOS1	NOS2	NPT1	SPT1	NAF	NES	MED
WIE1	NA	0.004	0.0044	0.0048	0.0053	0.0065	0.0066	0.0034	0.0109
WIE2	0.004	NA	0.0042	0.0046	0.005	0.0063	0.0062	0.0035	0.0105
NOS1	0.0044	0.0042	NA	0.0012	0.0054	0.0068	0.0068	0.0038	0.0112
NOS2	0.0048	0.0046	0.0012	NA	0.0058	0.0072	0.007	0.0041	0.0116
NPT1	0.0053	0.005	0.0054	0.0058	NA	0.0074	0.0071	0.0047	0.0118
SPT1	0.0065	0.0063	0.0068	0.0072	0.0074	NA	0.0062	0.0054	0.0126
NAF	0.0066	0.0062	0.0068	0.007	0.0071	0.0062	NA	0.005	0.012
NES	0.0034	0.0035	0.0038	0.0041	0.0047	0.0054	0.005	NA	0.0101
MED	0.0109	0.0105	0.0112	0.0116	0.0118	0.0126	0.012	0.0101	NA



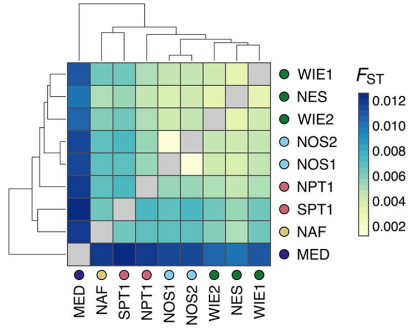
Fuentes-Pardo, A. P., Farrell, E. D., Pettersson, M. E., Sprehn, C. G., & Andersson, L. (2023). The genomic basis and environmental correlates of local adaptation in the Atlantic horse mackerel (*Trachurus trachurus*). *Evolutionary Applications*, 16, 1201–1219. https://doi.org/10.1111/eva.13559

Pairwise F_{ST}

Absolute values are informative, traditionally researchers have used these ranks for interpretation:

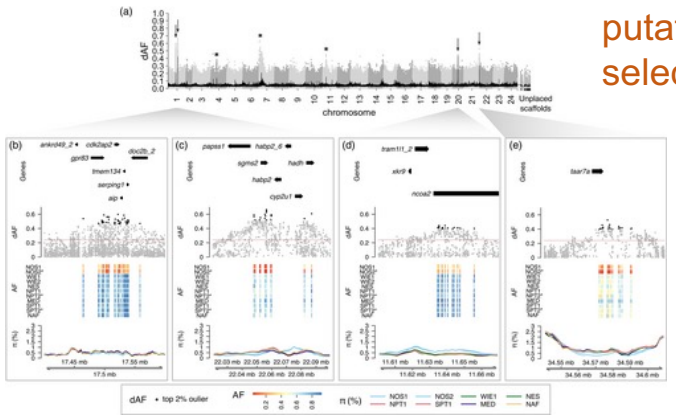
- 0.000x = high gene flow (don't bother too much about looking for structure)
- 0.01-0.1 = consider population structure
- >0.1 = do you really have one species?

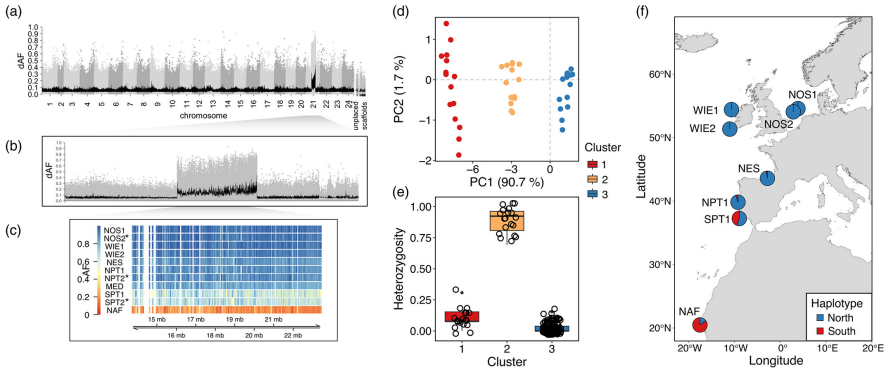
Caution: F_{ST} varies depending on the type of data you have (RRS, or WGS, number of markers, composition –neutral and/or outliers) and the life history of your species, so be careful about these value ranges, don't take them as golden run



Atlantic horse Mackerel
(*Trachurus trachurus*)

Genome scans revealed regions putatively under selection



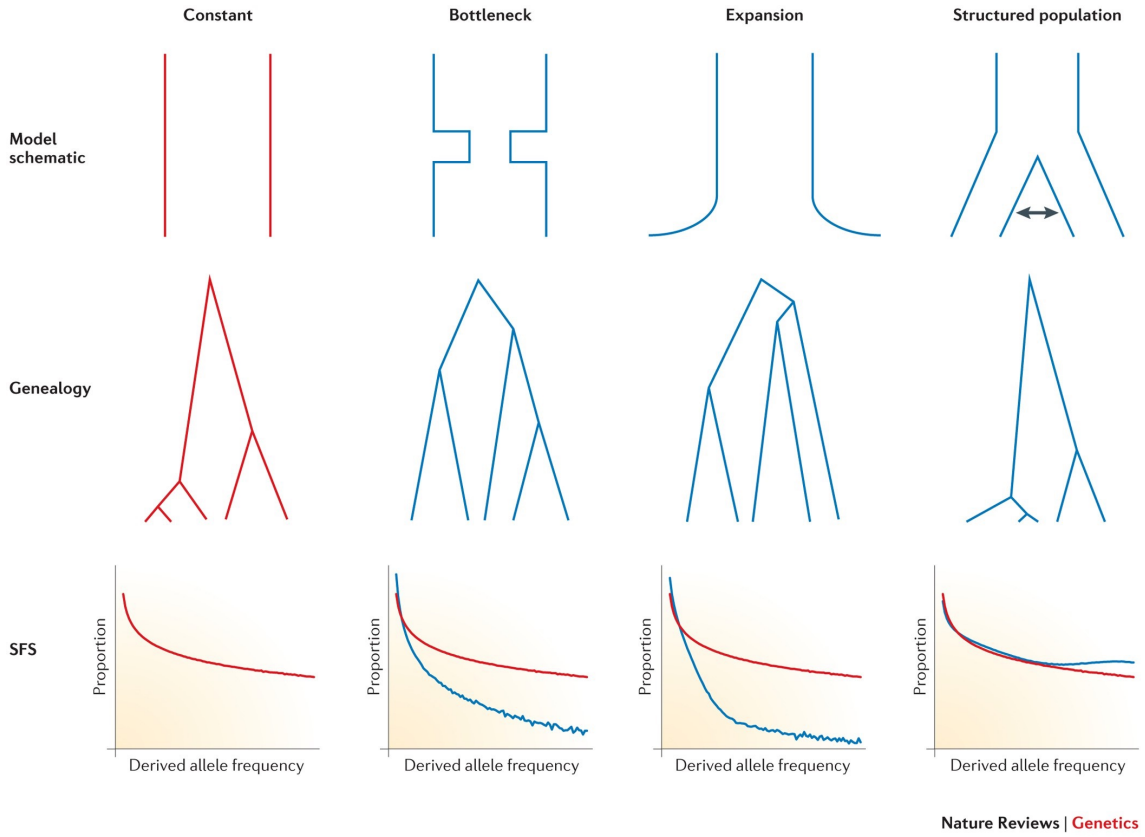


Fuentes-Pardo, A. P., Farrell, E. D., Pettersson, M. E., Sprehn, C. G., & Andersson, L. (2023). The genomic basis and environmental correlates of local adaptation in the Atlantic horse mackerel (*Trachurus trachurus*). *Evolutionary Applications*, 16, 1201–1219. https://doi.org/10.1111/eva.13559

Beyond present structure :
How to study population history and demography?

Use simulations and models:

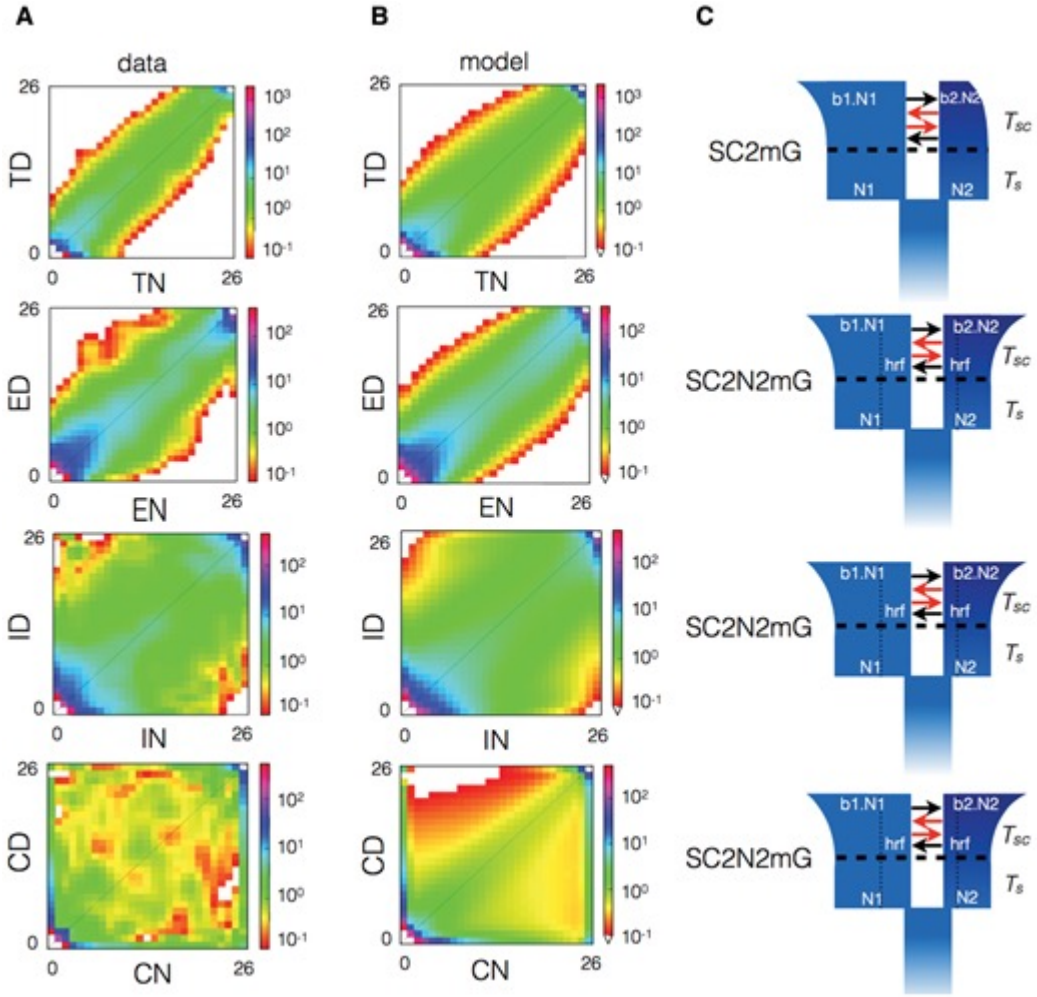
- To understand population history, bottleneck, gene flow, etc.
- Demography can set a null model against which one can look for the effect of selection



Schraiber & Akey 2015

Beyond present structure :
How to study population history and demography?

- Based on coalescence theory
- Compare SFS (site frequency spectrum) between real data and modelled data under different evolutionary scenarios
- Common tools: dadi, FastSimCoal, ABC, etc.



Rougeux et al 2017 GBE

Population structure and demography

- A good overview of this topic :

Schraiber, J., Akey, J. Methods and models for unravelling human evolutionary history. Nat Rev Genet 16, 727–740 (2015). https://doi.org/10.1038/nrg4005