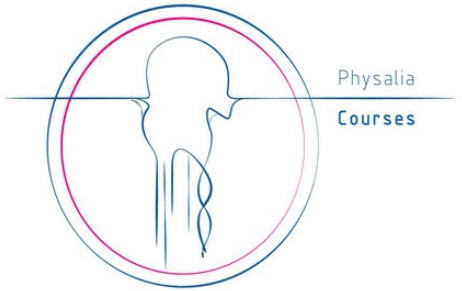


Genomic signatures of selection and adaptation

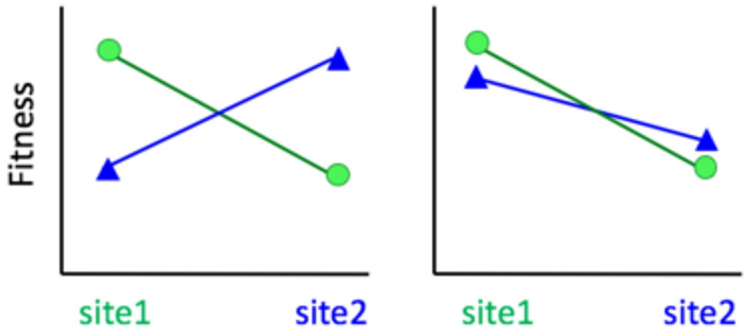
Day 3 - Lecture 1

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

Basic principle: Local adaptation

Geographic heterogeneity in environment

Local populations have been selected by local ecological conditions



Basic principle: Local adaptation

- local adaptation has heritable genetic basis
- phenotypes related to local adaptation

Example of local adaptation

Provenance variation in 8-year-old *Picea sitchensis* from across the species range grown in a common garden in Vancouver, BC, Canada

Aitken et al, 2015 Evol.App

Basic principle: Local adaptation

Can we use genomic data to understand the genetic basis of local adaptation?

Can we find the loci contributing to divergence between populations?

Can we find the loci possibly associated with relevant traits or relevant ecological variables?

Genome scan for local adaptation

Approach 1: Genetic outliers of differentiation between locally adapted populations

- Search for unexpected patterns in allele frequencies across the genome

Approach 2: Genetic associations with environment/phenotype

- Search for correlations between allelic frequencies and other variables

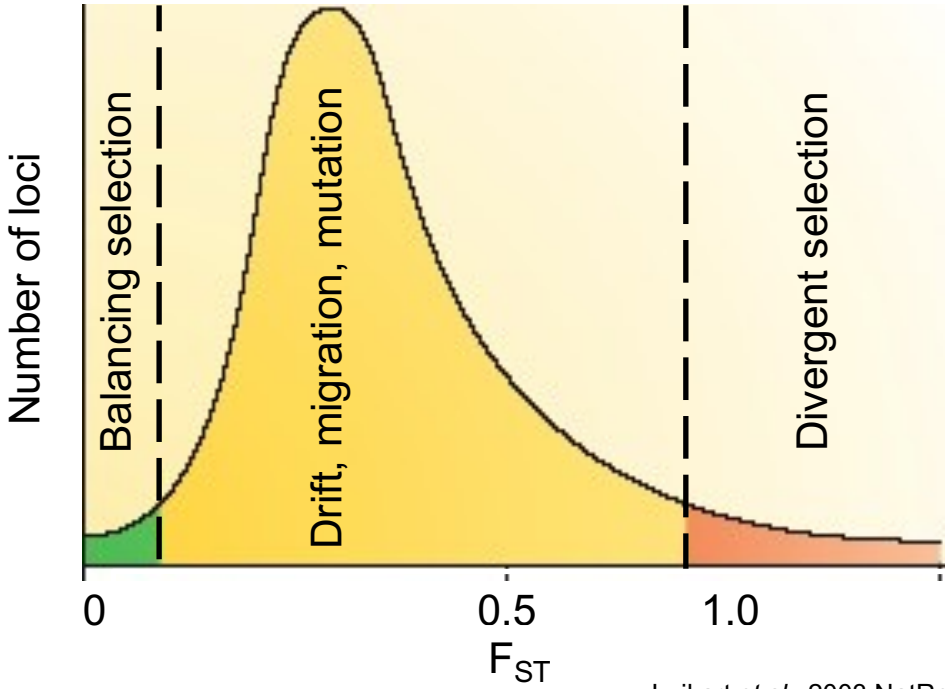
Outliers of divergence – signature of selection

F_{ST} statistics

- A measure of differentiation between populations relatively to intra-population diversity

F_{ST} = 1 : complete fixation of the alleles in each population

F_{ST} = 0 : same allelic frequencies



Luikart *et al.*, 2003 NatRevGen

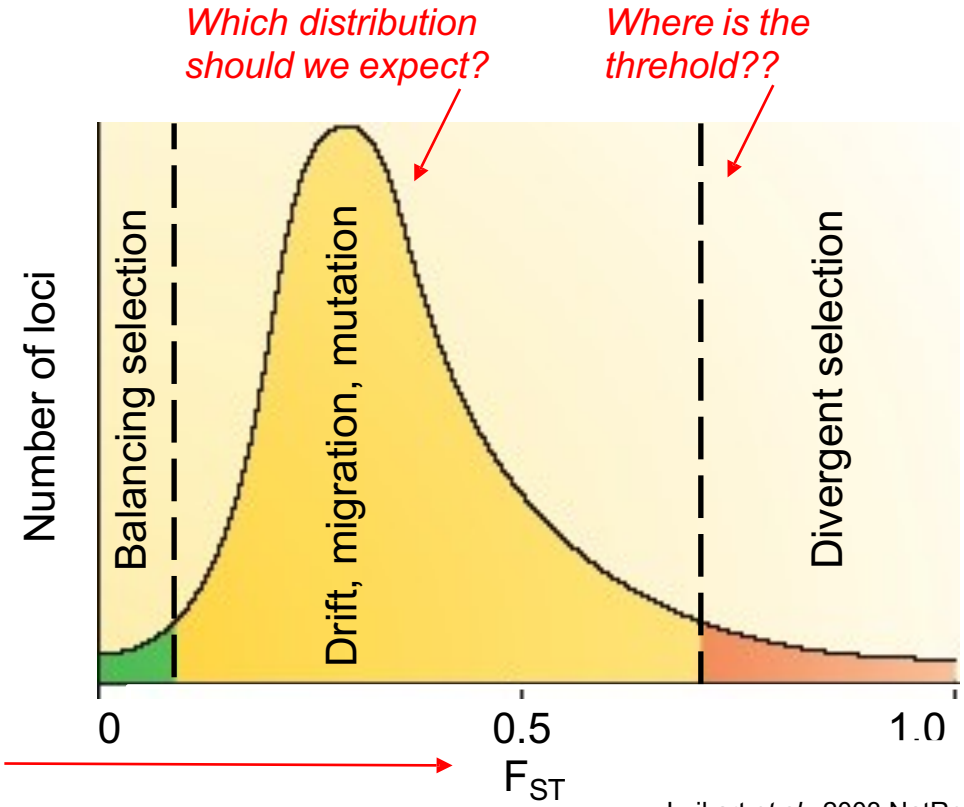
Outliers of divergence – signature of selection

F_{ST} statistics

- A measure of differentiation between populations relatively to intra-population diversity

F_{ST} = 1 : complete fixation of the alleles in each population

F_{ST} = 0 : same allelic frequencies



Luikart *et al.*, 2003 NatRevGen

Outliers of divergence – signature of selection

Factors determining F_{st} distribution:

- Population structure
- Demography
- Sample size
- Background selection
- Recombination

Outliers of divergence – signature of selection

Factors determining F_{st} distribution:

- Population structure
- Demography
- Sample size
- Background selection
- Recombination

Baseline level of differentiation between locally adapted populations is too high

Difficult (impossible?) to detect outliers of differentiation

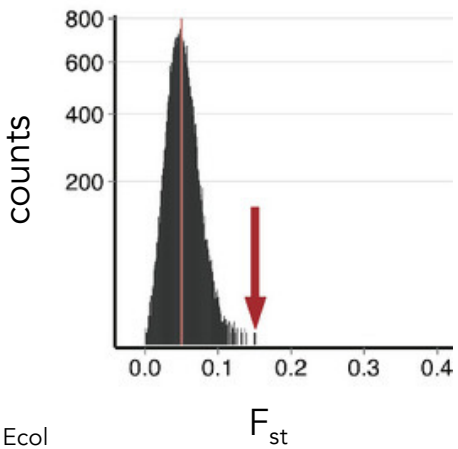
Outliers of divergence – signature of selection

Factors determining F_{st} distribution:

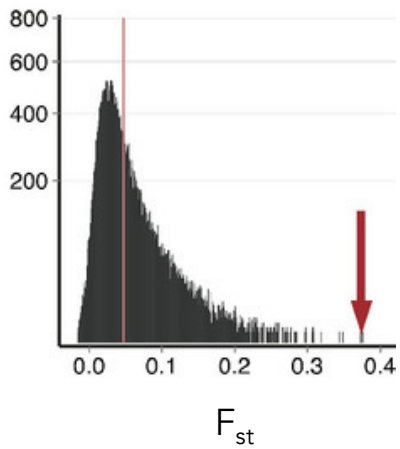
- Population structure
- Demography
- Sample size
- Background selection
- Recombination

History of divergence (demography) between the target populations

Island model



expansion from two refugia with secondary contact



Lotterhos & Whitlock 2014 Mol Ecol

Outliers of divergence – signature of selection

How to account for population structure and demography?

1) Assume a model of dispersion and demography

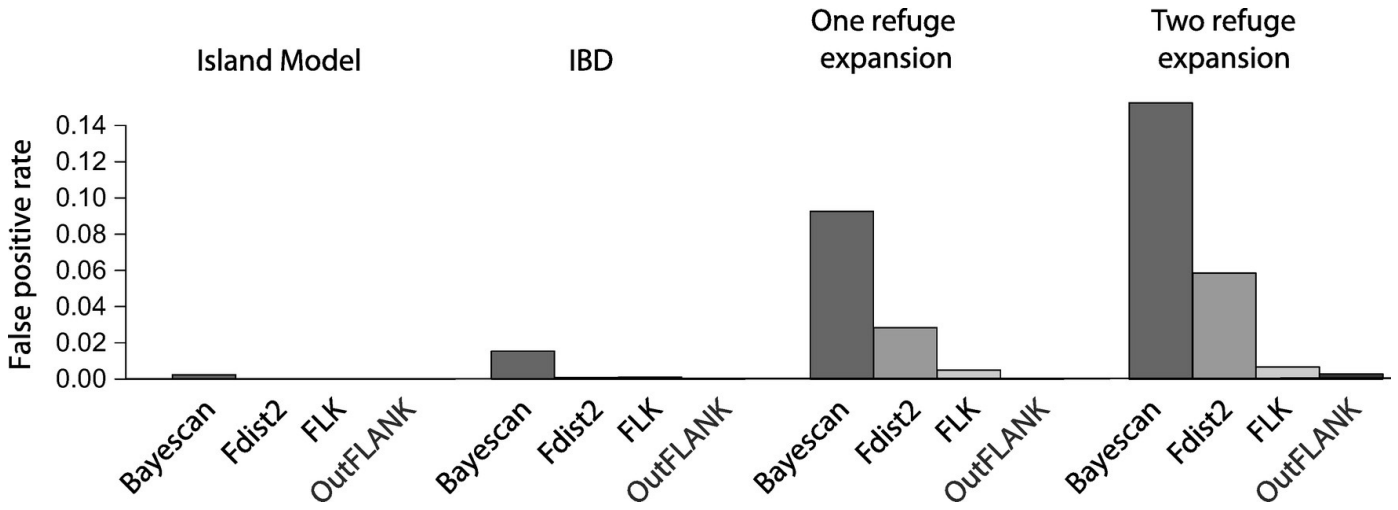
- F_{dist}, Bayescan

2) Estimate a neutral model from data

- covariance matrix between pop (**Baypass**/Bayenv2)
- X^2 distribution of pruned SNPs (**OutFLANK**)

Outliers of differentiation – signature of selection

Many « false positive »…



Or capturing effects unlinked to selection (sampling bias, unaccounted structure, hybridization)…

Lotterhos & Whitlock 2014 Mol Ecol

Outliers of differentiation – signature of selection

Factors determining F_{st} distribution:

- Population structure
- Demography
- Sample size
- Background selection
- Recombination

- Power:
  - sample size within population
  - number of populations

Lotterhos & Whitlock 2014 Mol Ecol

Outliers of differentiation – signature of selection

Factors determining F_{st} distribution:

- Population structure
- Demography
- Sample size
- Background selection
- Recombination

Low heterozygosity

high F_{st} with small differences in allelic frequencies

e.g. Background selection (negative selection)

D_{xy} – genetic divergence allele frequency differences

More on this topic:
Berner 2019, Genes
Matthey-Doret & Whitlock 2019 MolEcol

Outliers of differentiation – signature of selection

Factors determining F_{st} distribution:

- Population structure
- Demography
- Sample size
- Background selection
- Recombination

Variance in F_{st}

higher in low recombination regions
(even without selection)

More on this topic:
Booker, Yeaman & Whitlock 2020
MolEcol

(As we have discussed in day 2)

Outliers of differentiation – signature of selection

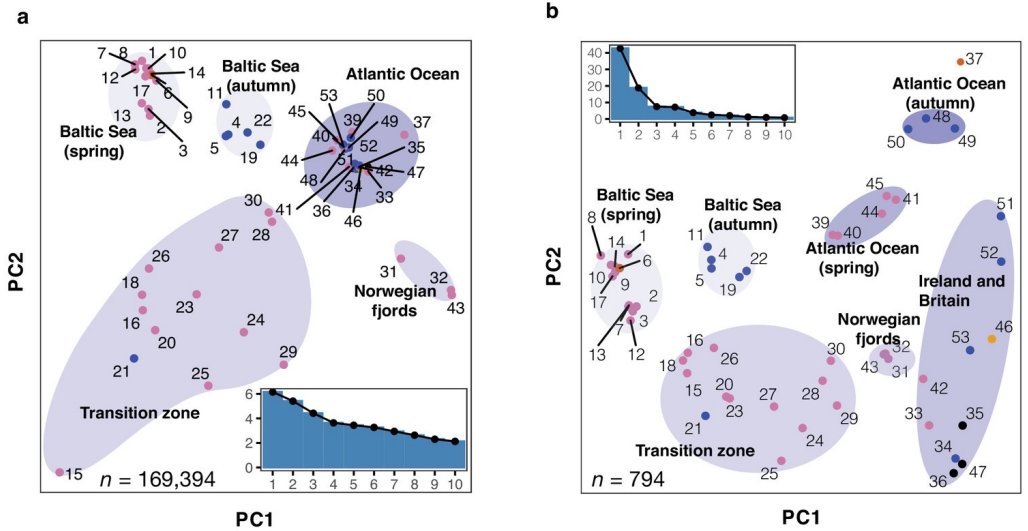
How to interpret outliers of differentiation?

- Demography requires neutral loci
- **But** selected loci may be important for population history
- Definition of "outlier" may depend on study design/ ecological information

Selected loci are important to find differentiation among Atlantic herring populations

Neutral loci

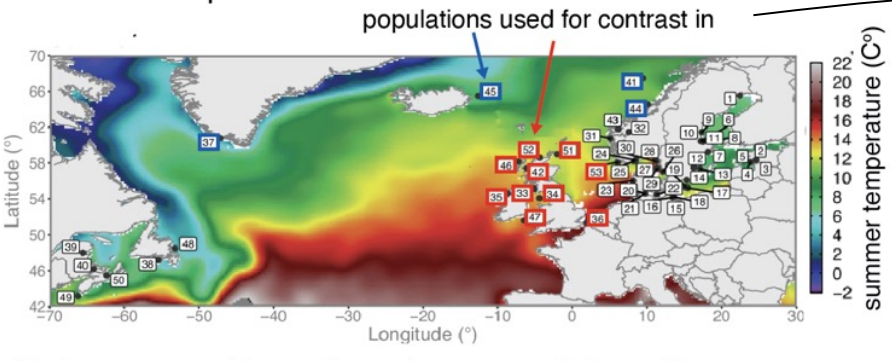
Loci under selection



Han et al 2020, eLife

Outliers of differentiation along geographic and environmental clines

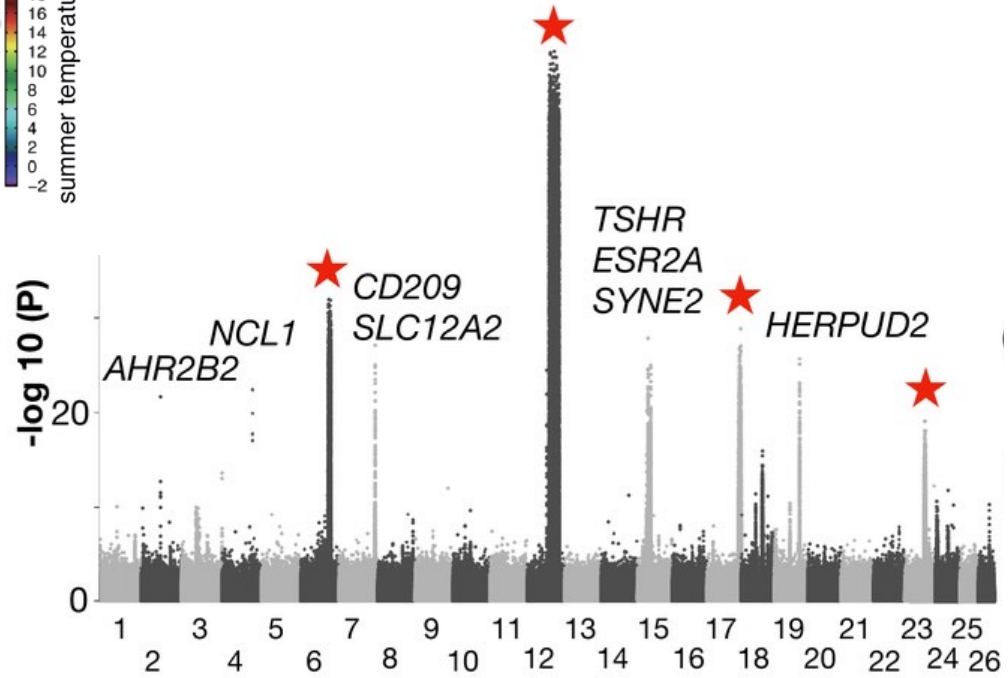
distribution of Atlantic herring populations across a gradient of sea water temperature



Outliers of differentiation:

Atlantic herring populations in warm vs cold seawater

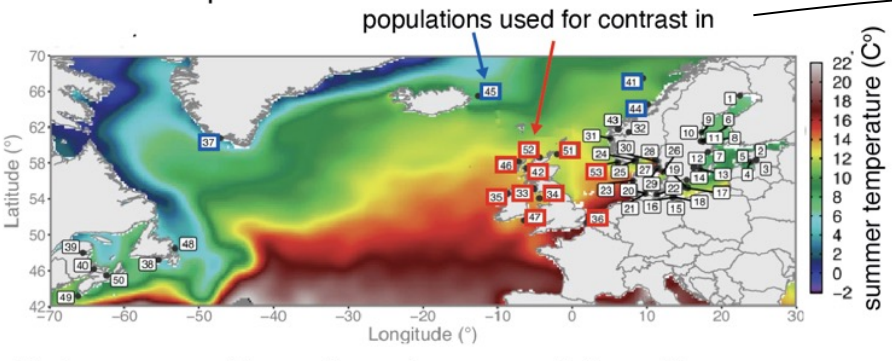
★ Inversions



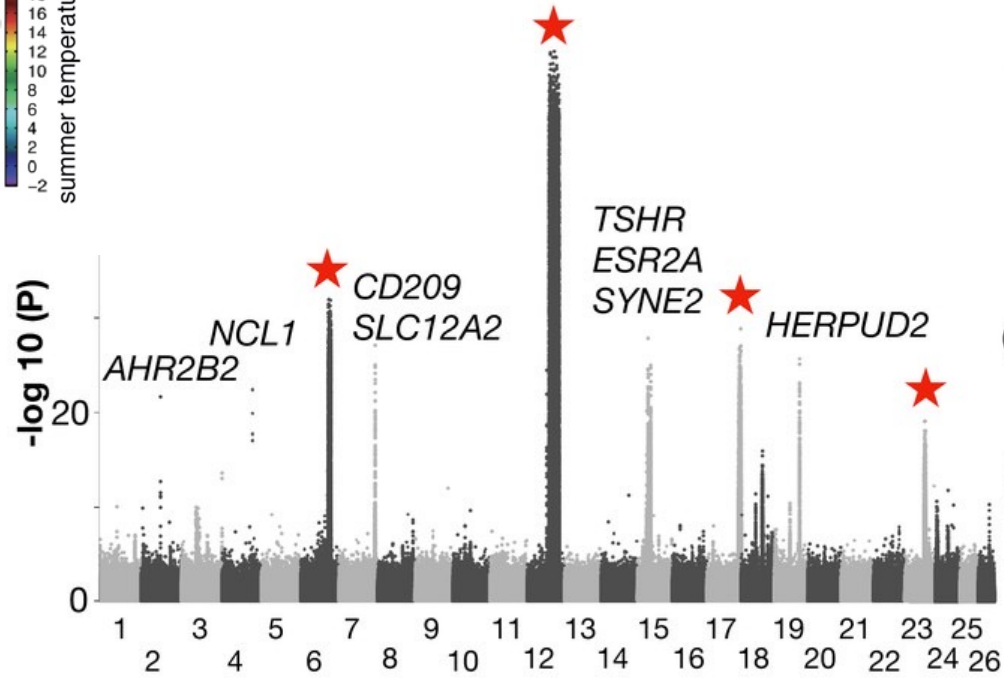
Han et al 2020, eLife

Outliers of differentiation along geographic and environmental clines

distribution of Atlantic herring populations across a gradient of sea water temperature



Populations experiencing the **same environment** will be **less differentiated** than ones in different environments

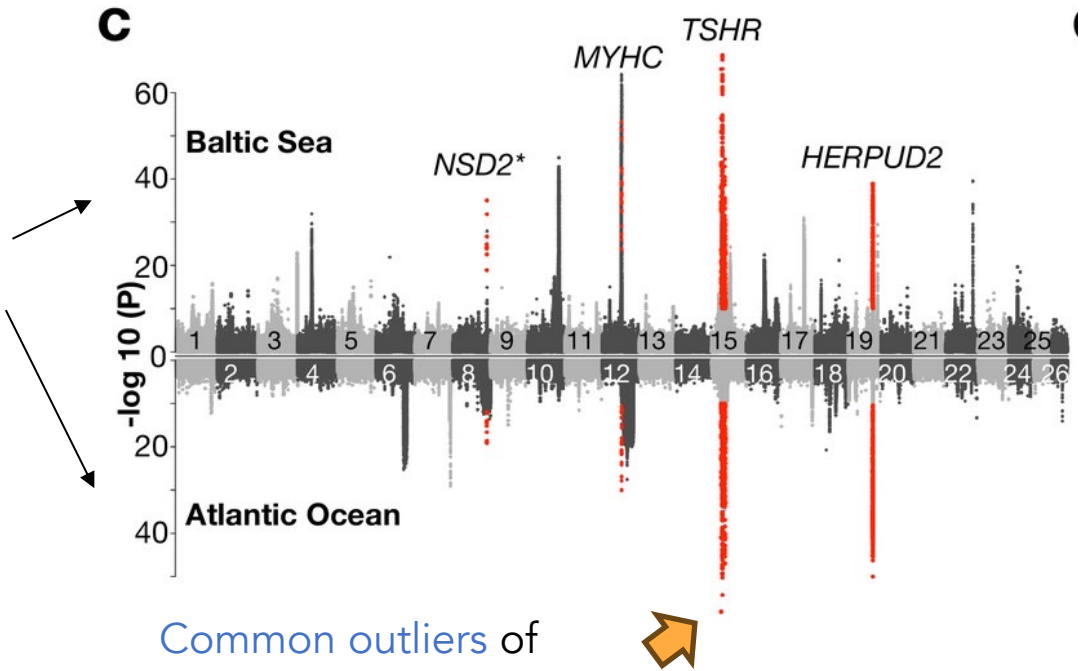


Han et al 2020, eLife

Outliers of differentiation between phenotypes

Spring spawning vs Autumn spawning Atlantic herring

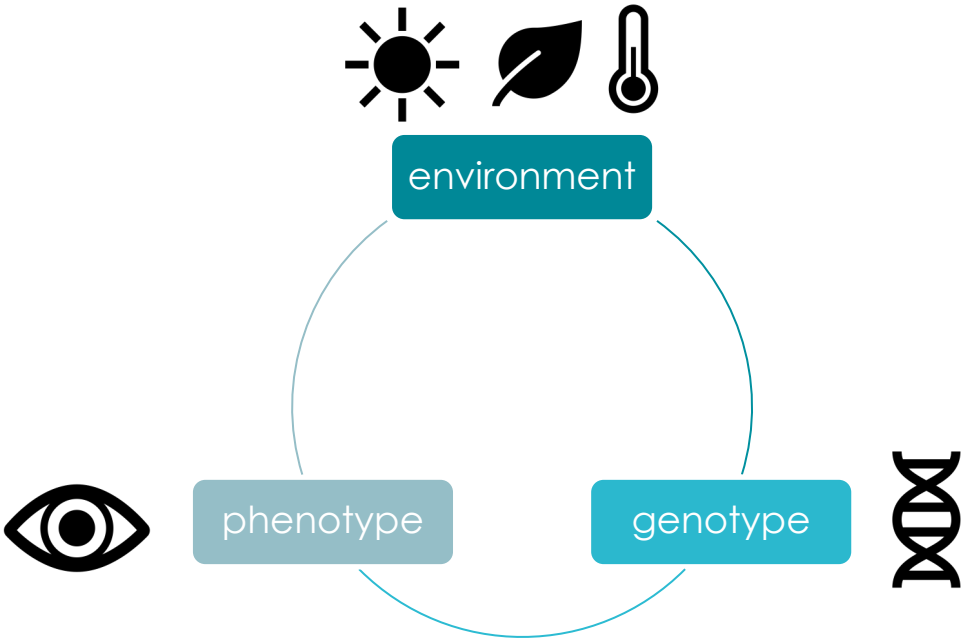
Important to take population structure into account



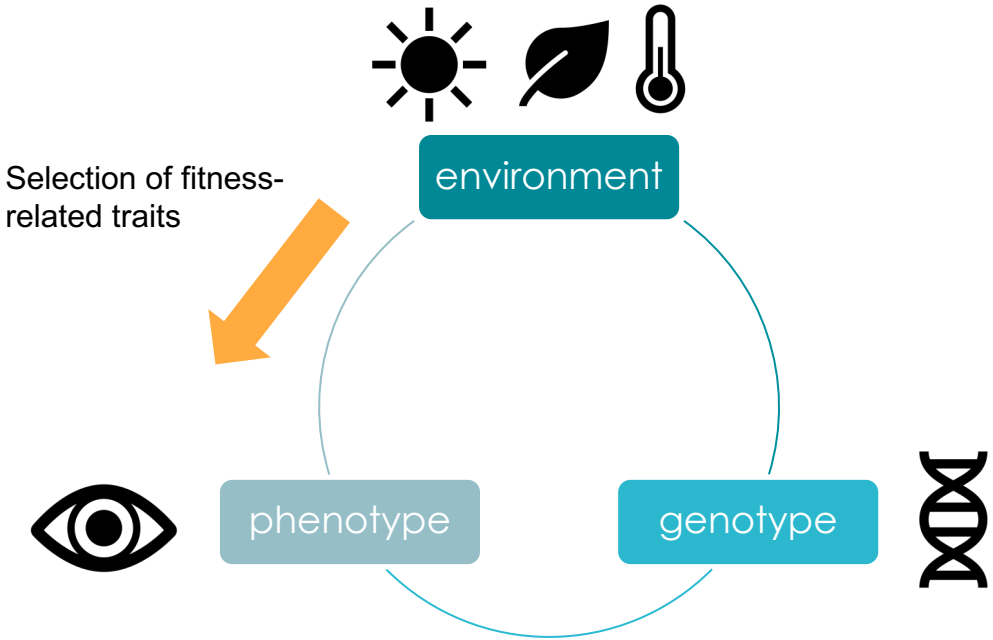
Common outliers of differentiation across populations: Stronger evidence of selection ✓

Han et al 2020, eLife

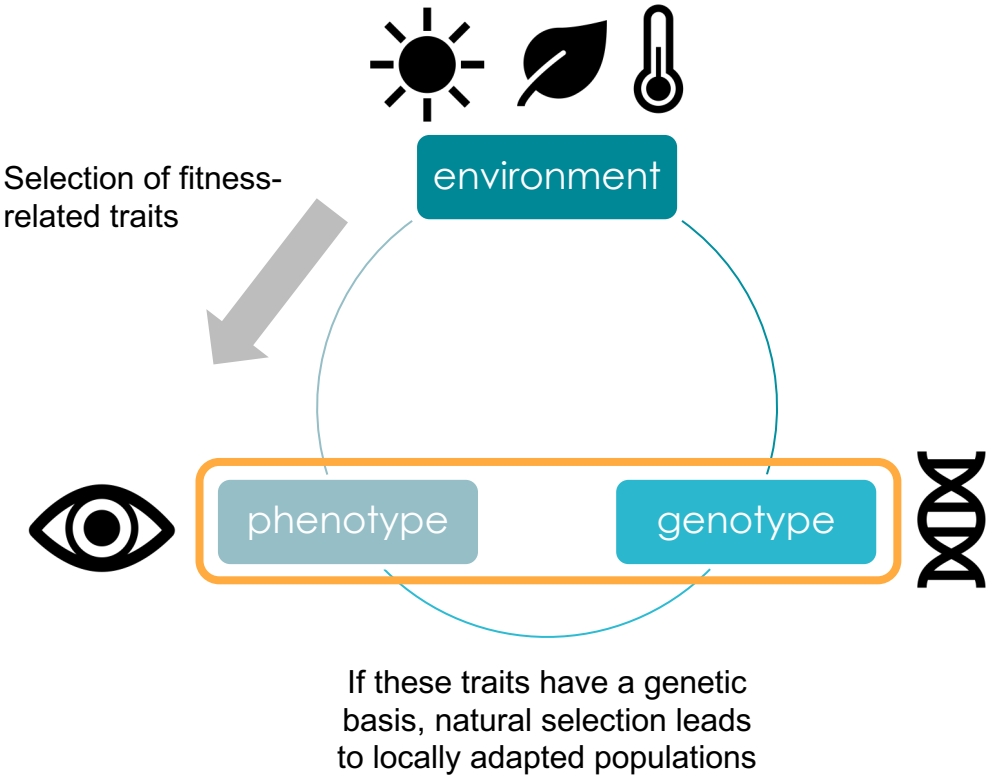
Genome-environment predictive framework



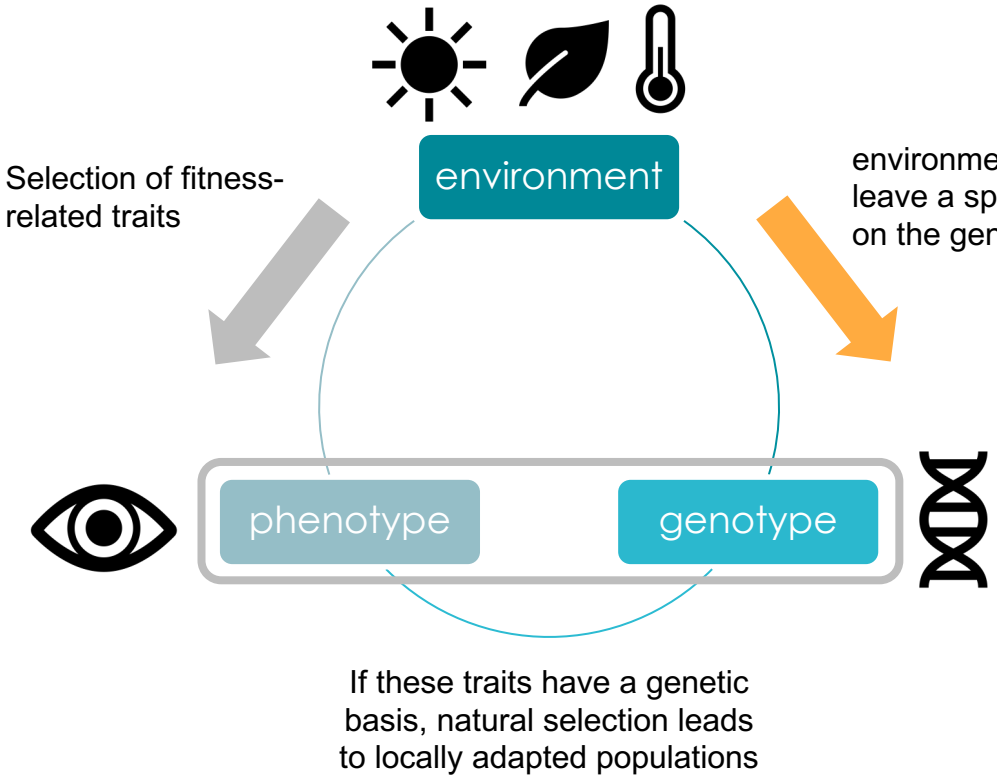
Genome-environment predictive framework



Genome-environment predictive framework



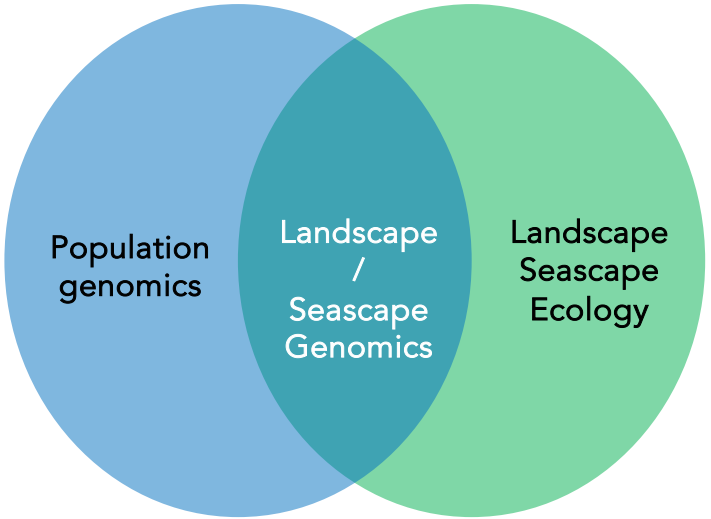
Genome-environment predictive framework



- Possible to infer local adaptation without directly measuring fitness traits
- Focus on determining the effect of environmental factors on allele frequencies using landscape genomic methods such as genotype–environment associations (GEA)

Landscapes genomics and environmental associations

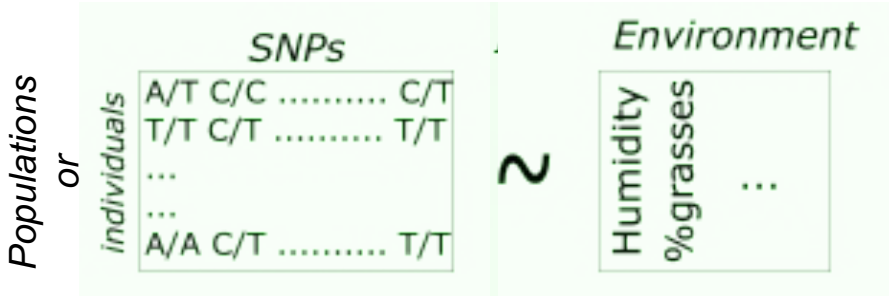
- Investigates how spatial and environmental factors influence geographic patterns of genome-wide genetic variation
- Determines **demographic factors** that affect **spatial distribution** of alleles (gene flow, geographic distance)
- Determines **environmental factors** that affect **spatial distribution** of alleles
- A better sense of **why** populations are differentiated?



Landscapes genomics and environmental associations

GEA methods

Genotype ~ Environment (+ correction?)



Univariate methods (locus by locus)

- Bayenv/Baypass, LFMM

Multivariate methods (all loci at once)

- Redundancy analysis (RDA)

Recommended readings:
Genome scan methods against more complex models: when and how much should we trust them? (Villemereuil et al, 2014 MolEcol)

Comparing methods for detecting multilocus adaptation with multivariate genotype–environment associations (Forester et al 2018, MolEcol)

Landscapes genomics and environmental associations

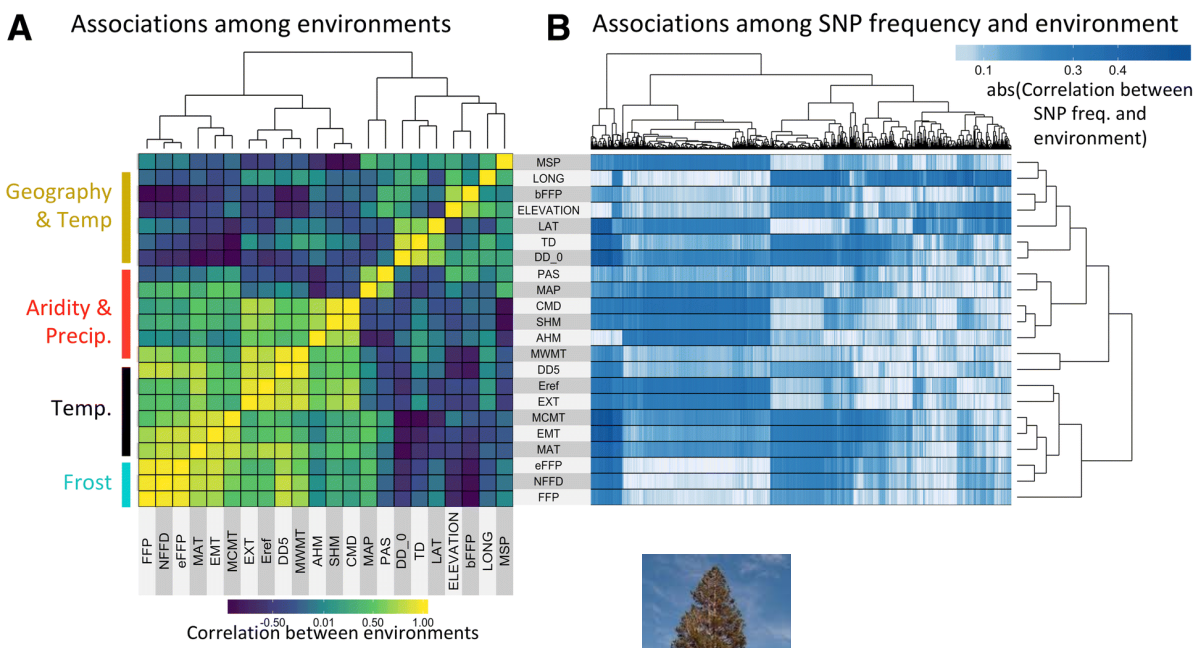
Univariate associations:

Which locus is associated with which environmental variable? (one at a time)

Spearman's correlation

Caution:

- environmental variables may be correlated!
- SNPs may be in physical LD



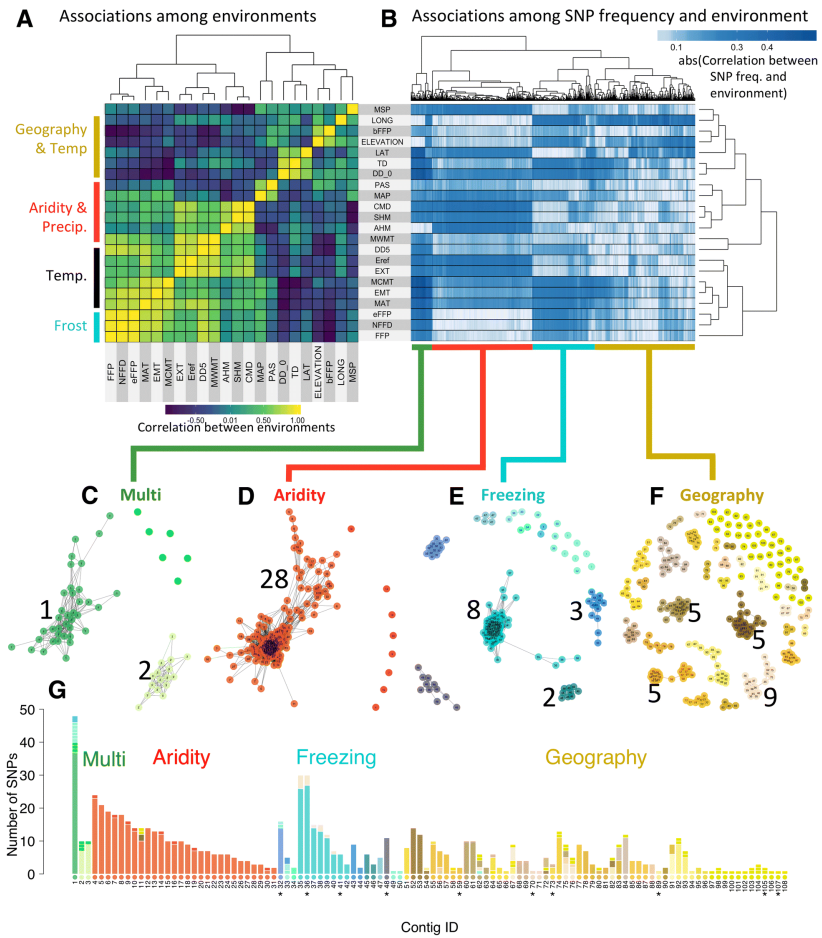
Pinus contorta

Lotterhos et al, 2018
Genome Biology

Landscapes genomics and environmental associations

Univariate associations:

Modular group of adaptive loci to different axis of environmental variation



Lotterhos et al, 2018 Genome Biology

Landscapes genomics and environmental associations

Univariate associations:

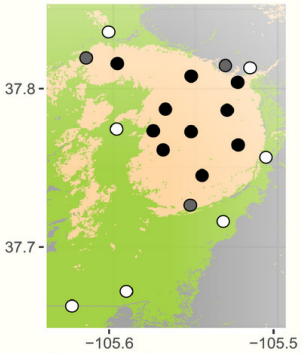
Bayenv2 – Baypass

DATA

Studied species

Helianthus petiolaris
Non-dune/dune ecotypes

Sampling



An environmental gradient

Sequencing

many SNPs

many individuals

A/T C/C ..... C/T
T/T C/T ..... T/T
...
...
A/A C/T ..... T/T

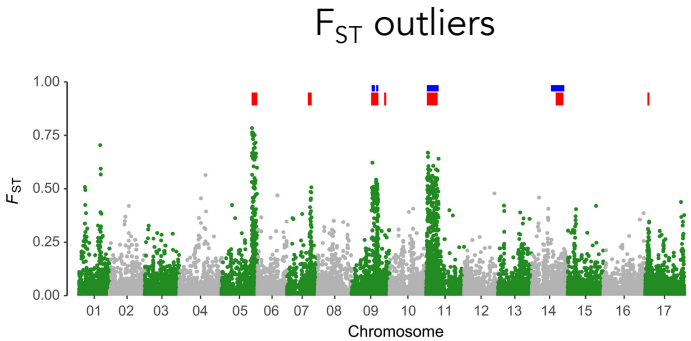
A matrix of SNPs genotypes from reduced-representation sequencing (RAD-seq)

Huang et al, 2020 MolEcol

Landscapes genomics and environmental associations

Univariate associations:

Bayenv2 – Baypass



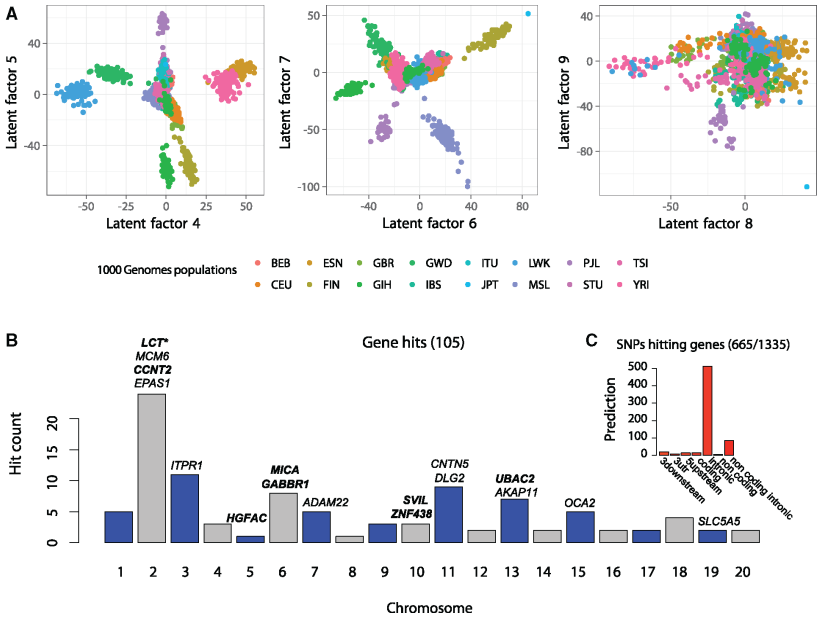
Huang et al, 2020 MolEcol

Landscapes genomics and environmental associations

Univariate associations:

LFMM (latent factor mixed model)

- Detects correlations between environmental and genetic variation while simultaneously inferring background levels of population structure
- Residual population structure is introduced via unobserved K (latent) factors that represent demographic history, IBD, hidden substructure



Human GEA study. Association study based on genomic data from the 1000 Genomes Project database and climatic data from the Worldclim database. (A) Latent factors estimated by LFMM 2.0. (B) Target genes corresponding to top hits of the GEA analysis (expected FDR level of 5%). The highlighted genes correspond to functional variants. (C) Predictions obtained from the VEP program.

Caye et al, 2019, Molecular Biology and Evolution.
https://doi.org/10.1093/molbev/msz008

Landscapes genomics and environmental associations

Univariate associations : Recommendations

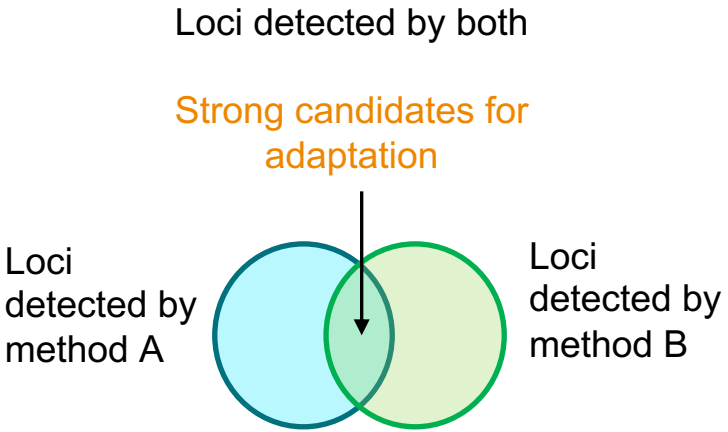
Intersect several methods

- More likely to be strong candidates for adaptation
- Reduce false positive but may also miss variants with less signal

Controlling false discovery rate

Correct for population structure

- An open debate?
- System dependent: high gene flow? IBD? Geography correlated with environmental variation?



Good reading: Controlling false discoveries in genome scans for selection. François et al, 2015, Molecular Ecology, https://doi.org/10.1111/mec.13513

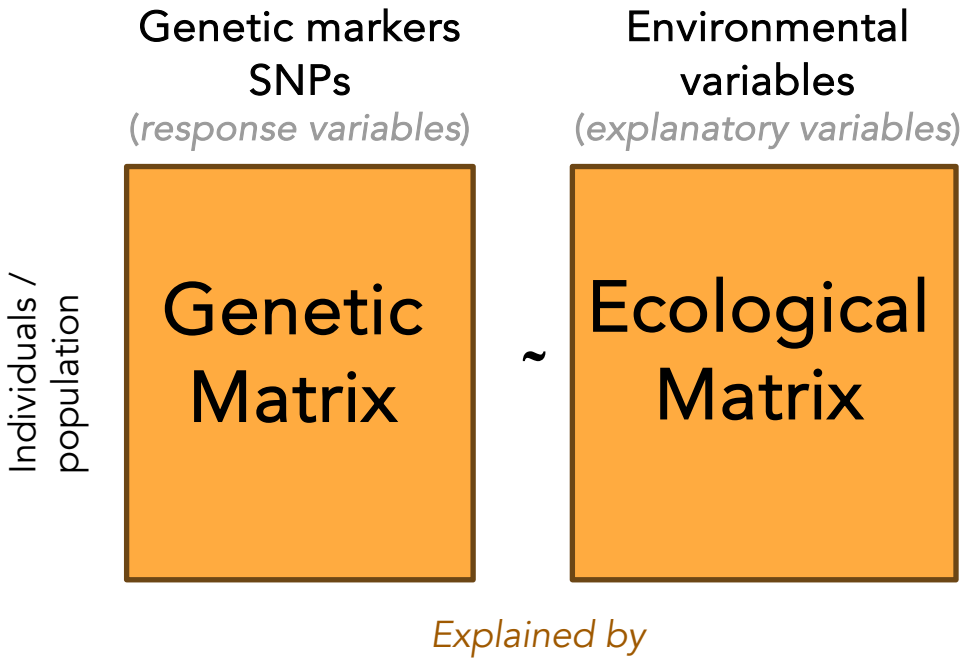
Landscapes genomics and environmental associations

Univariate associations : Limitations

- Test one genetic marker at a time, and may also test only one environmental predictor at a time
- Do not account for covariation among environmental variables and/or genetic markers
- For large genomic datasets including hundreds of thousands or even millions of genetic markers and multiple environmental predictors, millions of univariate tests would be required

Landscapes genomics and environmental associations

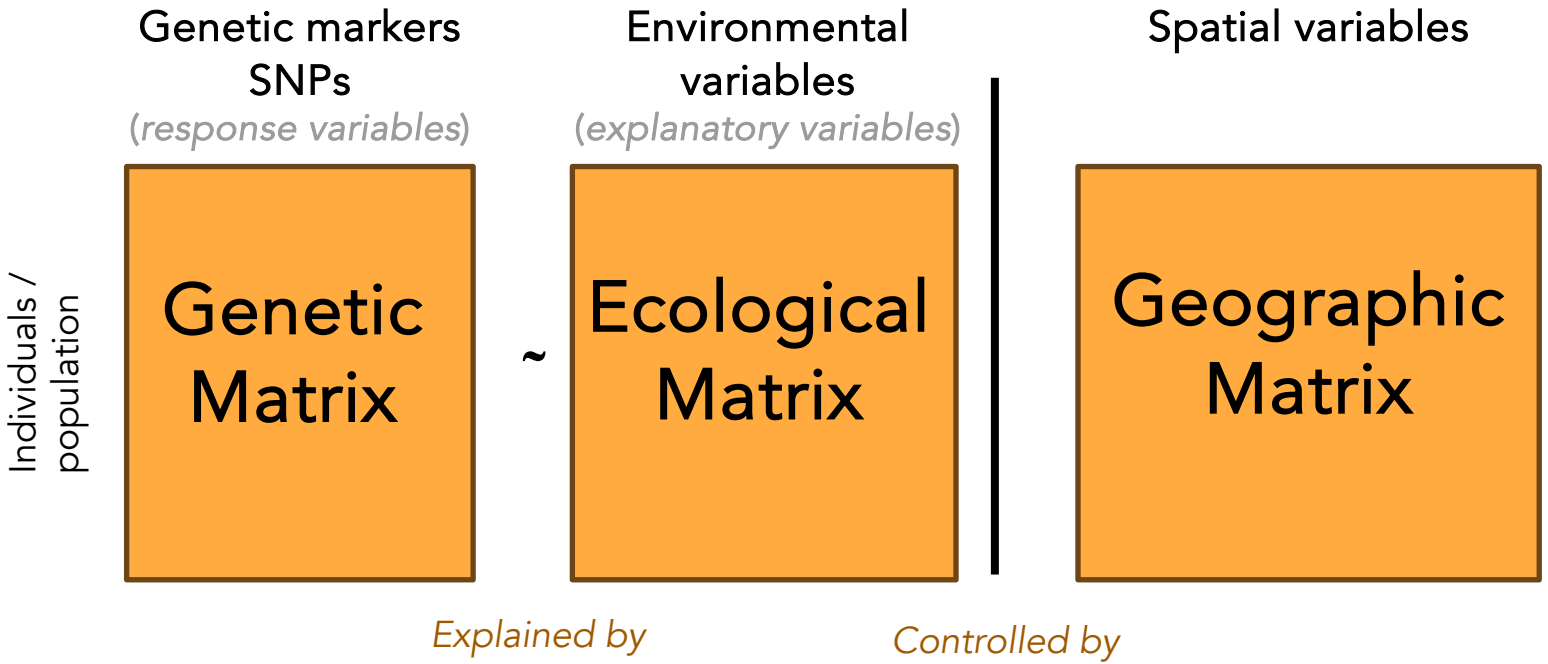
Multivariate associations : RDA



https://doi.org/10.1016/B978-0-444-53868-0.50014-9

Landscapes genomics and environmental associations

Multivariate associations : RDA



https://doi.org/10.1016/B978-0-444-53868-0.50014-9

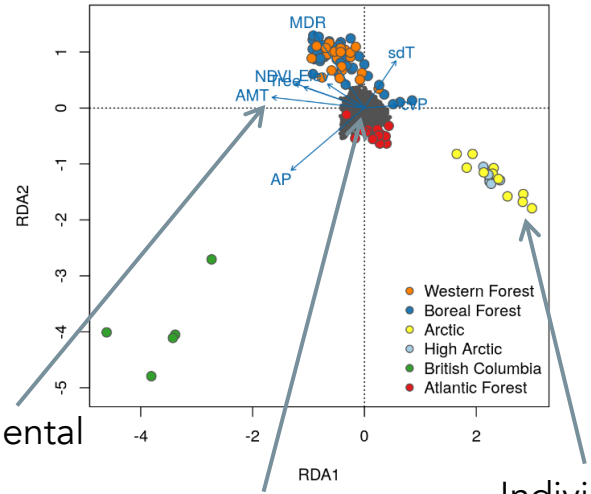
Landscapes genomics and environmental associations

Multivariate associations : RDA

- Use the contribution of genetic markers along the different axis to detect putatively-selected loci

```
points(X.rda, display="sites")
```

Triplot RDA (individuals centered)



Environmental variable

Genetic markers

Individual

AP: high annual precipitation
sdT: and low temperature seasonality
MDR: small mean diurnal temperature range
ATM: low annual mean temperature (AMT)
Tree: levels of tree cover
NDVI: a measure of vegetation greenness

Forester et al 2018 Mol Ecol

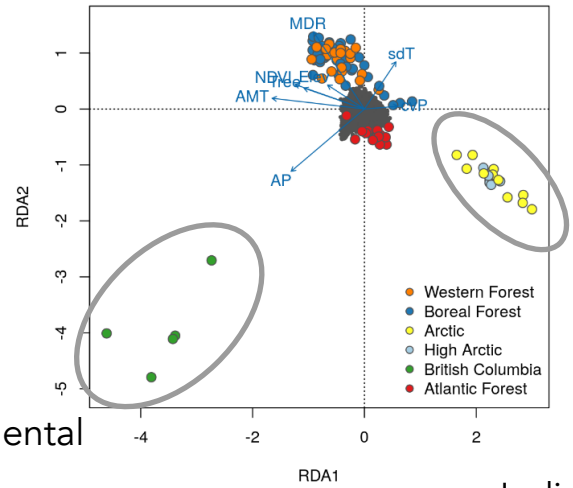
Landscapes genomics and environmental associations

Multivariate associations : RDA

- Use the contribution of genetic markers along the different axis to detect putatively-selected loci

```
points(X.rda, display="sites")
```

Triplot RDA (individuals centered)



AP: high annual precipitation
sdT: and low temperature seasonality
MDR: small mean diurnal temperature range
ATM: low annual mean temperature (AMT)
Tree: levels of tree cover
NDVI: a measure of vegetation greenness

Environmental variable

Genetic markers

Individual

Forester et al 2018 Mol Ecol

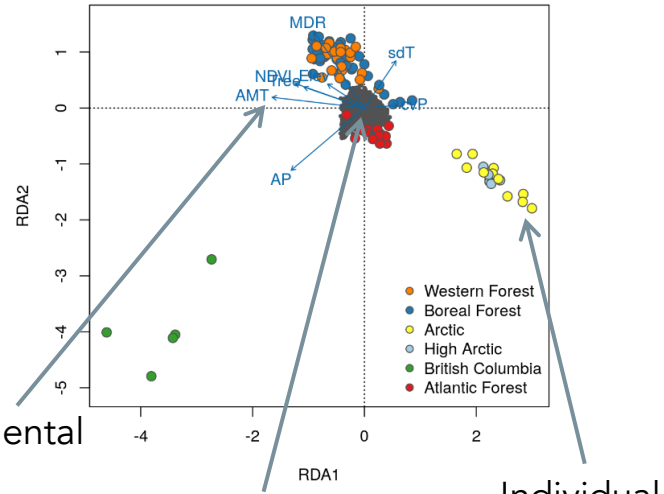
Landscapes genomics and environmental associations

Multivariate associations : RDA

- Use the contribution of genetic markers along the different axis to detect putatively-selected loci

```
points(X.rda, display="sites")
```

Triplot RDA (individuals centered)



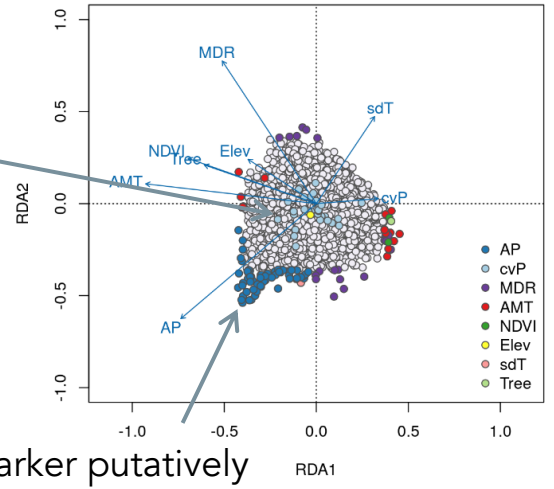
Environmental variable

Genetic markers

Individual

```
points(X.rda, display="species")
```

Triplot RDA (SNPs centered)



Neutral marker

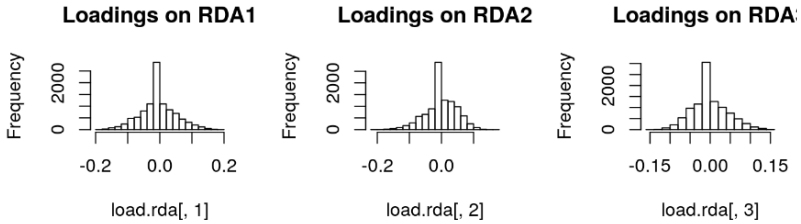
Outlier marker putatively associated to adaptive variation

Forester et al 2018 Mol Ecol

Landscapes genomics and environmental associations

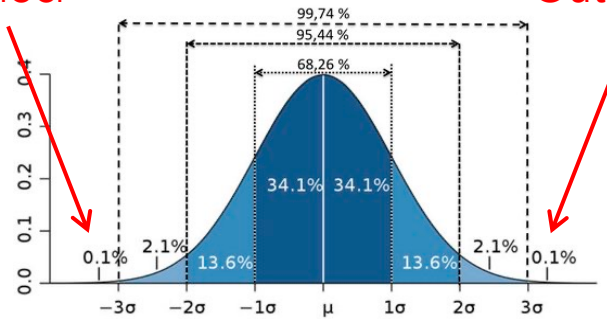
Multivariate associations : RDA

- Use the contribution of genetic markers along the different axis to detect putatively-selected loci



Outlier loci

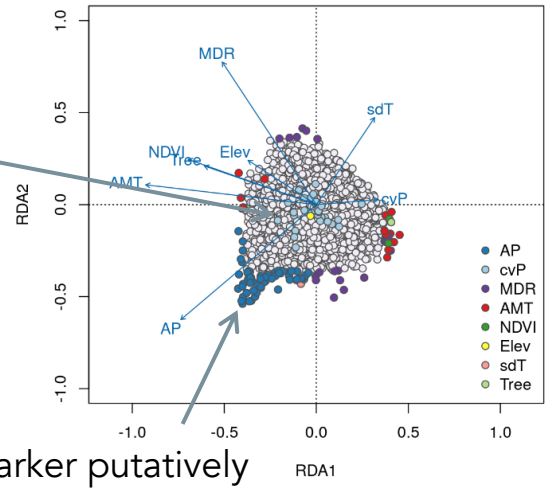
Outlier loci




```
points(X.rda, display="species")
```

Triplot RDA (SNPs centered)

Neutral marker



Outlier marker putatively associated to adaptive variation

Forester et al 2018 Mol Ecol

Landscapes genomics and environmental associations

Multivariate associations : RDA

- Use the contribution of genetic markers along the different axis to detect putatively-selected loci

Recommended tutorials:

- https://popgen.nescent.org/2018-03-27_RDA_GEA.html
- https://github.com/Capblancq/RDA-landscape-genomics

Landscapes genomics and environmental associations

Multivariate associations : RDA

Advantages

- Multi-locus: Polygenic adaptation possible to detect?
- Multi-variable analysis: More realistic environmental characterization
- Very fast + global information:
  - Which variables explain genetic variance?
  - Correction possible by population/geographic structure

Landscapes genomics and environmental associations

Multivariate associations : RDA

Limitations

- Assumes a linear dependence between the response variables (genotypes) and the explanatory variables (environmental predictors)
  - Meaning that nonlinear relationships will not be detected
  - Fortunately, there are nonlinear statistical methods that have been adapted to landscape genomic analyses (logistic regression and gradient forest)
  - Variations on classic RDA have been developed to accommodate nonlinear relationships, though these have yet to see development or use in landscape genomics

Landscapes genomics and environmental associations

Recommended readings:

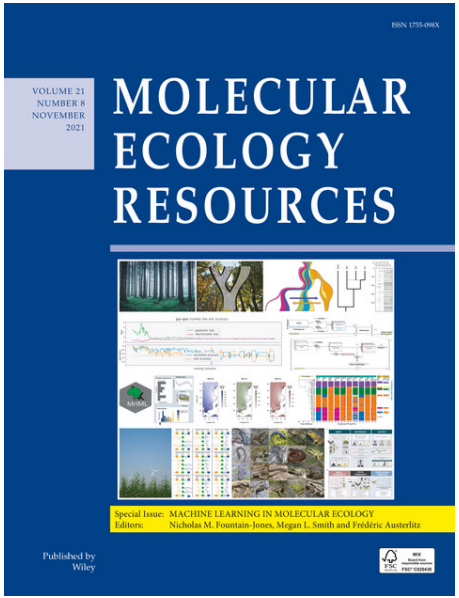
Finding the Genomic Basis of Local Adaptation: Pitfalls, Practical Solutions, and Future Directions. Hoban et al 2016 Am Nat. https://www.journals.uchicago.edu/doi/full/10.1086/688018

Capblancq, T., & Forester, B. R. (2021). Redundancy analysis: A Swiss Army Knife for landscape genomics. Methods in Ecology and Evolution, 12, 2298–2309. https://doi.org/10.1111/2041-210X.13722

The future (or the present?)

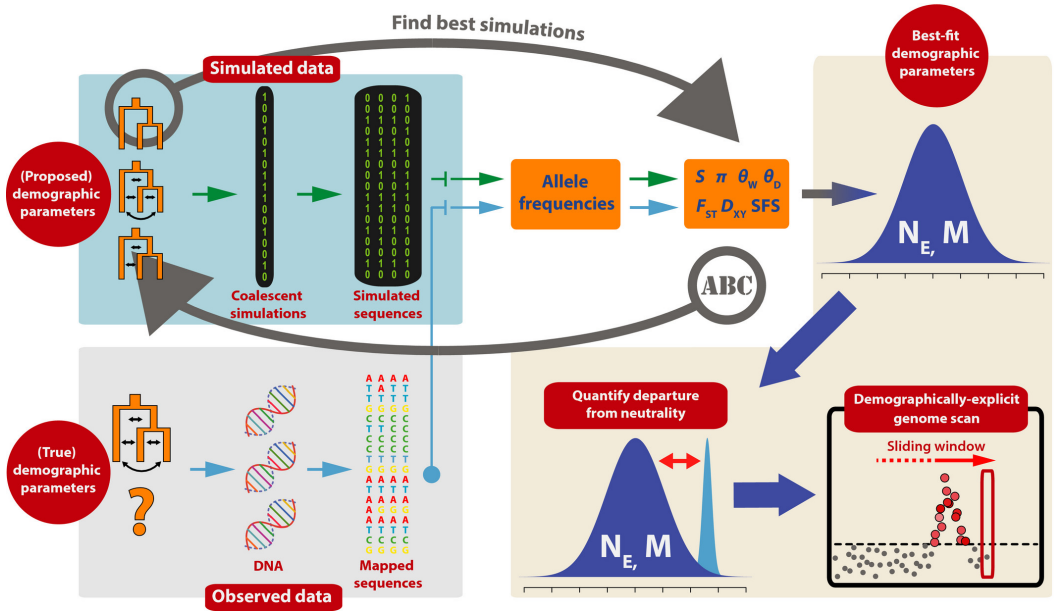
Machine learning methods

Special Issue
Machine learning in molecular ecology
Nicholas M. Fountain-Jones, Megan L. Smith, Frédéric Austerlitz



Identifying putative adaptive loci via explicit demographic models

Luqman, H., Widmer, A., Fior, S. and Wegmann, D. (2021), Identifying loci under selection via explicit demographic models. Mol Ecol Resour, 21: 2719-2737. https://doi.org/10.1111/1755-0998.13415



Landscape genetics with machine learning

Fountain-Jones, N. M., Kozakiewicz, C. P., Forester, B. R., Landguth, E. L., Carver, S., Charleston, M., Gagne, R. B., Greenwell, B., Kraberger, S., Trumbo, D. R., Mayer, M., Clark, N. J., & Machado, G. (2021). MrIML: Multi-response interpretable machine learning to model genomic landscapes. *Molecular Ecology Resources*, 21, 2766–2781. https://doi.org/10.1111/1755-0998.13495