

Genome-scans

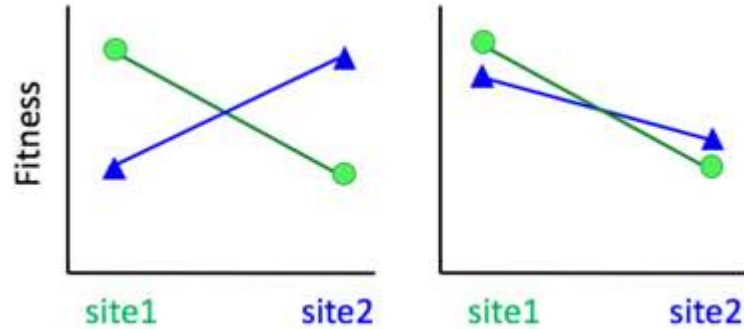
Landscape genomics

Claire Mérot & Anna Tigano
Physalia Courses
September 2020

Basic principle: Local adaptation

Geographic heterogeneity in environment

Local populations have been selected by local ecological conditions

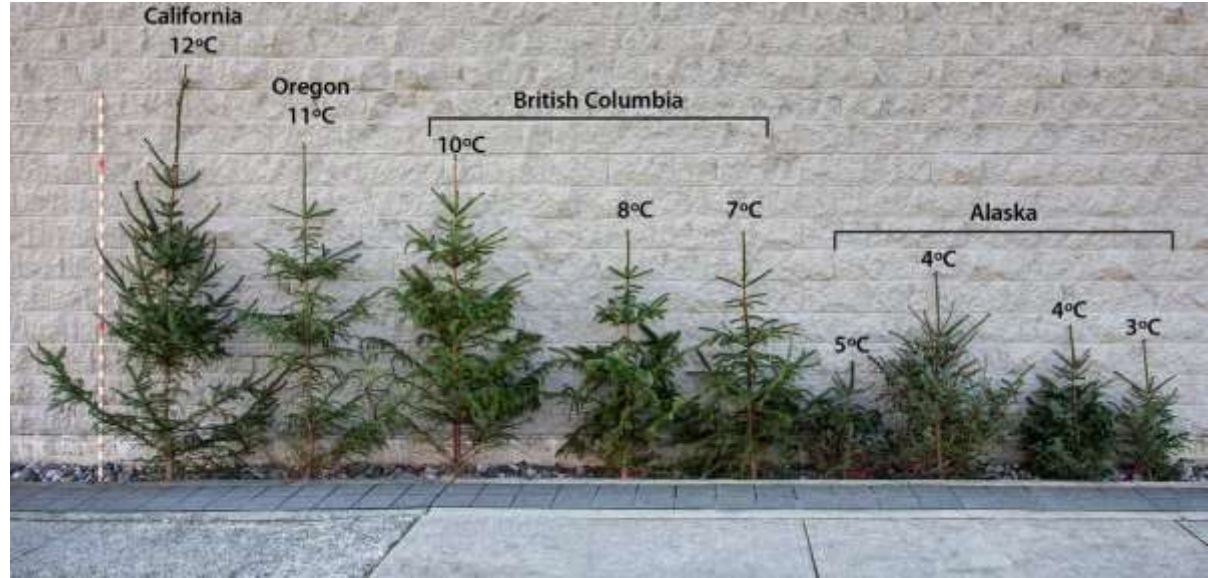


Basic principle: Local adaptation

Common garden experiment

⇒ local adaptation has heritable genetic basis

⇒ phenotypes related to 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* <https://www.onlinelibrary.wiley.com/doi/10.1111/eva.12293>

⇒ *This may be visible as « signature of selection »*

⇒ *We expect differences of allelic frequency between sites*

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 : detect outliers of divergence

-> Search for unexpected patterns in allele frequencies accross the genome

Approach 2: detect associations with environment/phenotype

-> Search for correlations between allelic frequencies and variables

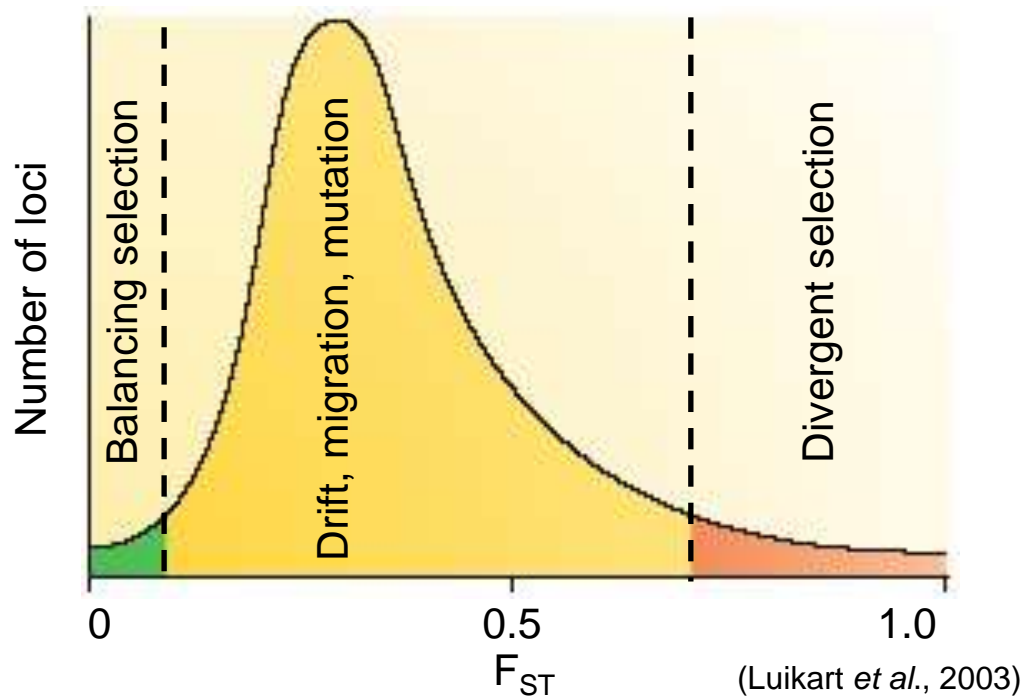
Outliers of divergence – signature of selection

Fst 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



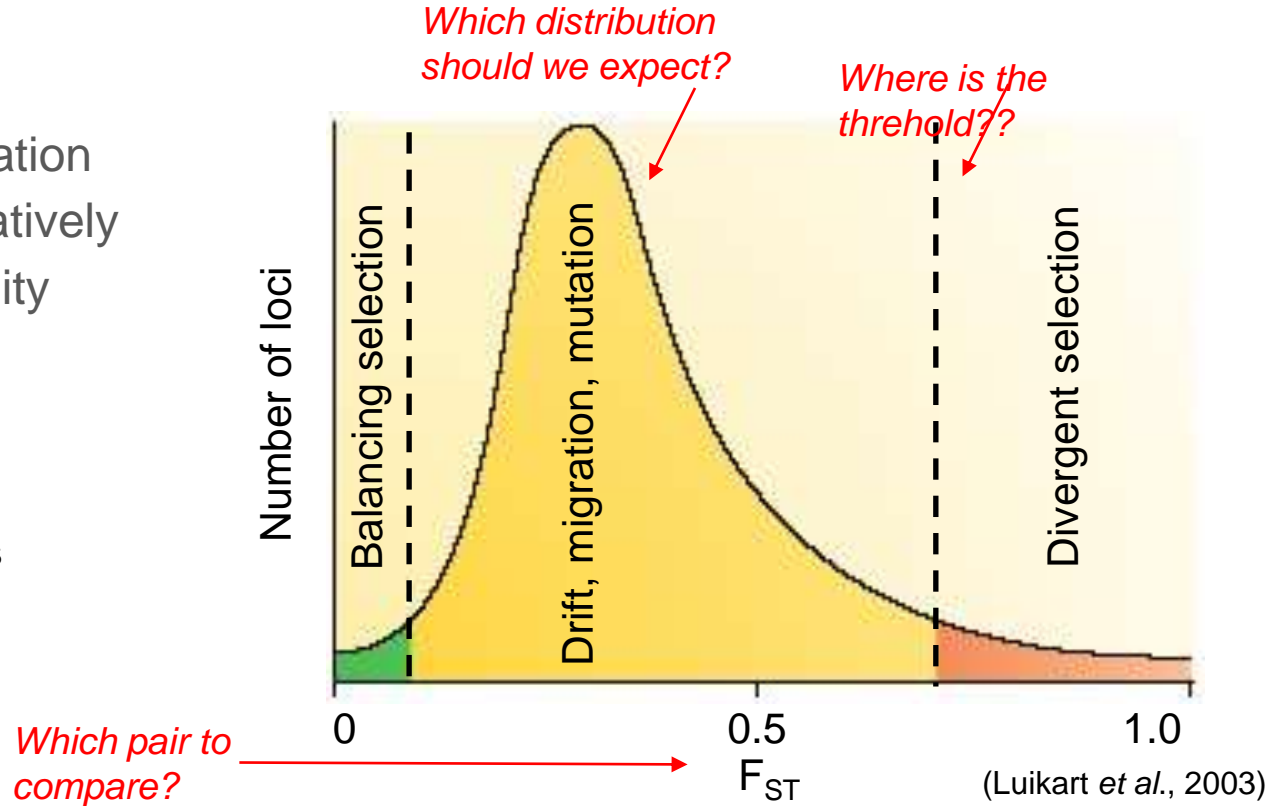
Outliers of divergence – signature of selection

Fst statistics

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

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

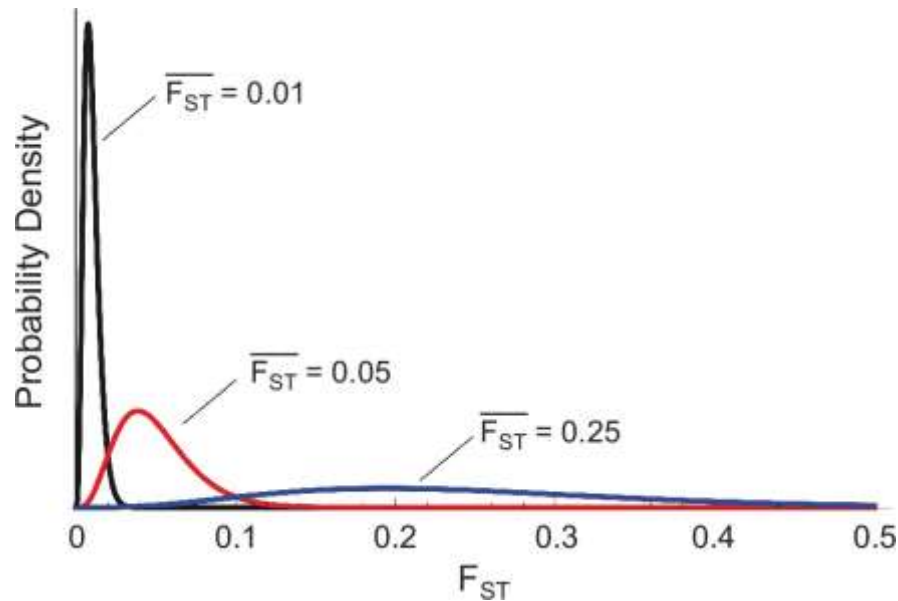
$F_{ST}=0$: same allelic frequencies



Outliers of divergence – signature of selection

Problem 1: Population structure

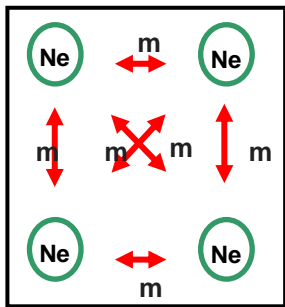
When the average level of differentiation is high, the variance in F_{ST} values among loci increases with average F_{ST} , which makes detection of outlier loci difficult for highly differentiated populations



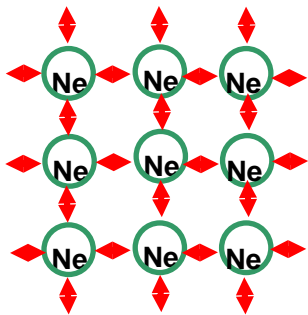
Outliers of divergence – signature of selection

Problem 2: Demography

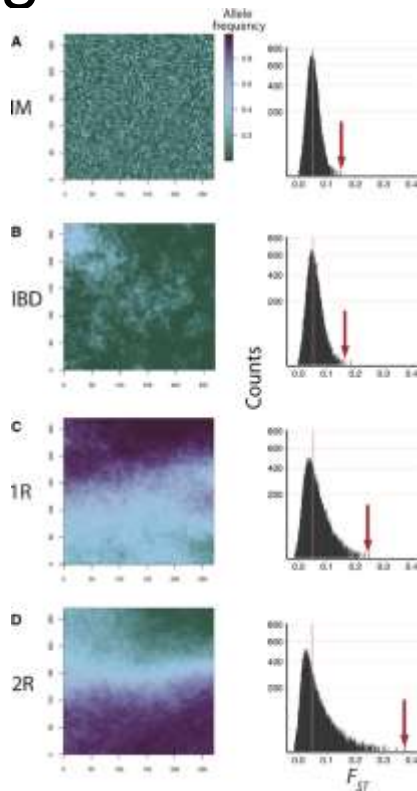
The distribution of F_{ST} will depend on the real demography of the populations



island model



Stepping-stone



island model

isolation by distance

expansion from one refugium

expansion from two refugia with secondary contact

Landscapes of frequency and F_{ST} for an outlier neutral locus at the end of the simulation (75 random samples)

Outliers of divergence – signature of selection

1st solution: assume a model of dispersion and demography
(Fdist, Bayescan)

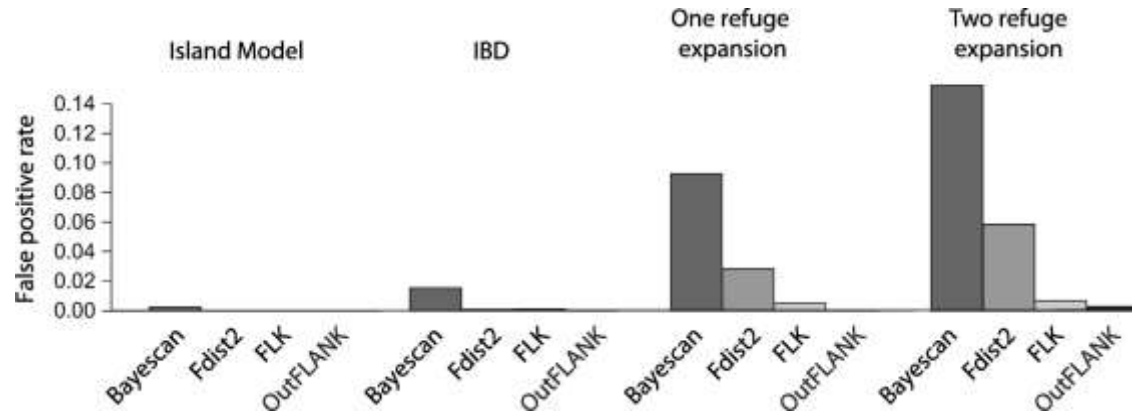
2nd solution: try to estimate a neutral model from data

-> covariance matrix between pop (Baypass/Bayenv2, FLK)

-> χ^2 distribution of pruned SNPs (OutFLANK)

Outliers of divergence – signature of selection

Many « false positive »...



Or capturing effects unlinked to selection (sampling bias, unaccounted structure, hybridisation)...

Outliers of divergence – signature of selection

Problem 3: Sample size

The power to detect meaningful differences based on genome-scan will depend on the number of populations, and the number of samples within a population

Table 2: Power for the one-refuge case analyzed with OutFLANK as a function of sample sizes, based on loci with $s_L = 0.01$, using >1,700 tests per case with a q value threshold of 5%

No. individuals per population	5 populations	10 populations	20 populations	40 populations
5	0	.09	.52	.84
10	.10	.56	.82	.94
20	.37	.75	.90	.95
40	.55	.81	.94	.97

Few samples per pop
-> stochasticity in
allelic frequency

Note: Parameters otherwise similar to those in figure 2.

Few populations
-> low statistical power

Outliers of divergence – signature of selection

Problem 4: Background selection

Fst is a relative measure of variation among populations. Low heterozygosity can inflate Fst values even for small differences in allelic frequencies...

For e.g. background selection (negative selection in regions of low-recombination)

-> Should we use other measures: dxy ? (but lower power for early stages of divergence)

AFD : Allelic frequency differences?

(Berner 2019, Genes [10.3390/genes10040308](https://doi.org/10.3390/genes10040308))

-> but the effect of background selection on Fst may not be that bad for populations connected by high gene flow.

See Matthey-Doret & Whitlock 2019 MolEcol <https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15197>

Outliers of divergence – signature of selection

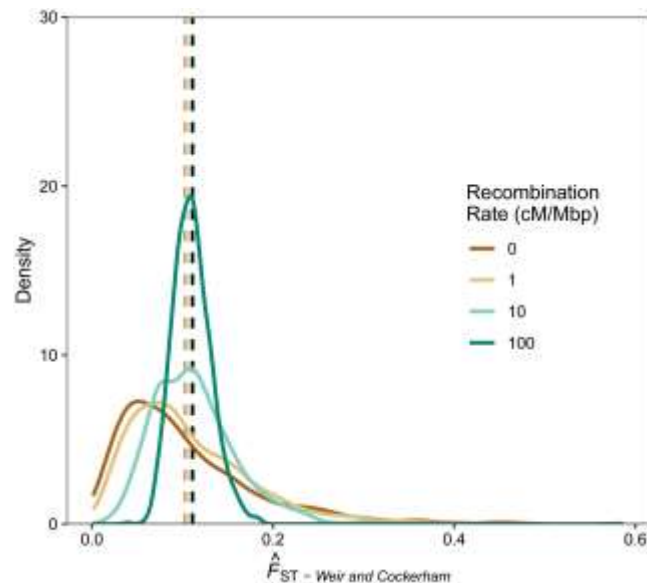
Problem 5: Recombination

Even without selection, F_{ST} variance is expected to be higher in low-recombination regions...

-> Should we use a different threshold depending on recombination??

See Booker, Yeaman & Whitlock 2020 MolEcol
<https://onlinelibrary.wiley.com/doi/abs/10.1111/mec.15501>

A field in development with upcoming whole-genome data...



Outliers of divergence – signature of selection

But the main problem is: how can we interpret the results?

Outliers tests can be useful to remove “putatively-adaptive” loci from analysis of neutral structure/demography...

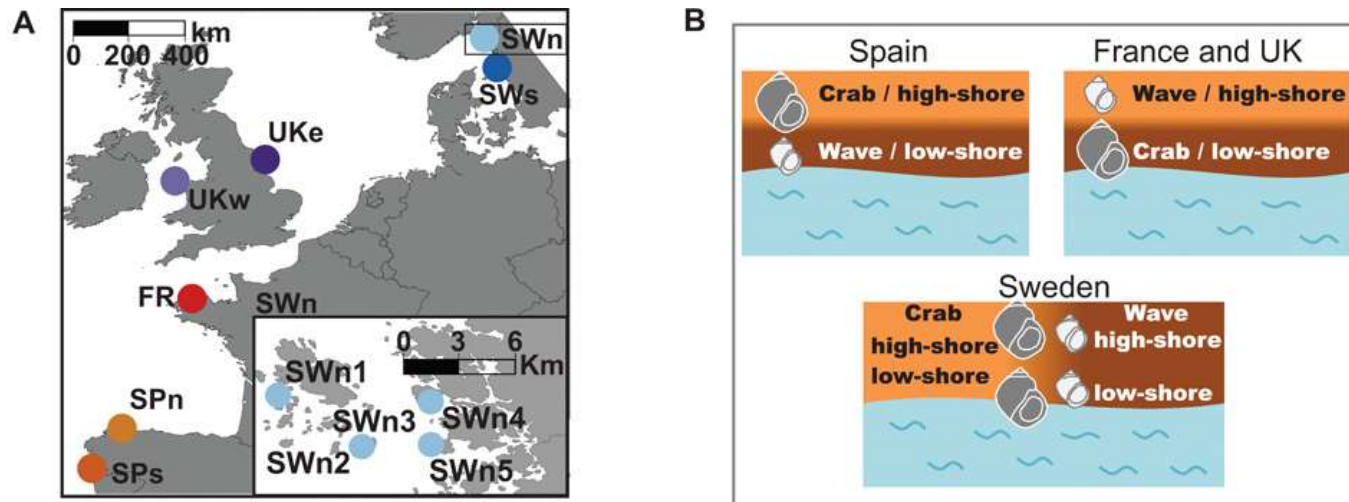
-> But do we want to avoid taking into account adaptation or specific loci showing traces of differentiation?!

What does outliers of differentiation means?

-> depend on the study design/ecological information...

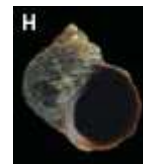
Outliers of divergence – signature of selection

Pairs of populations: contrast F_{st}



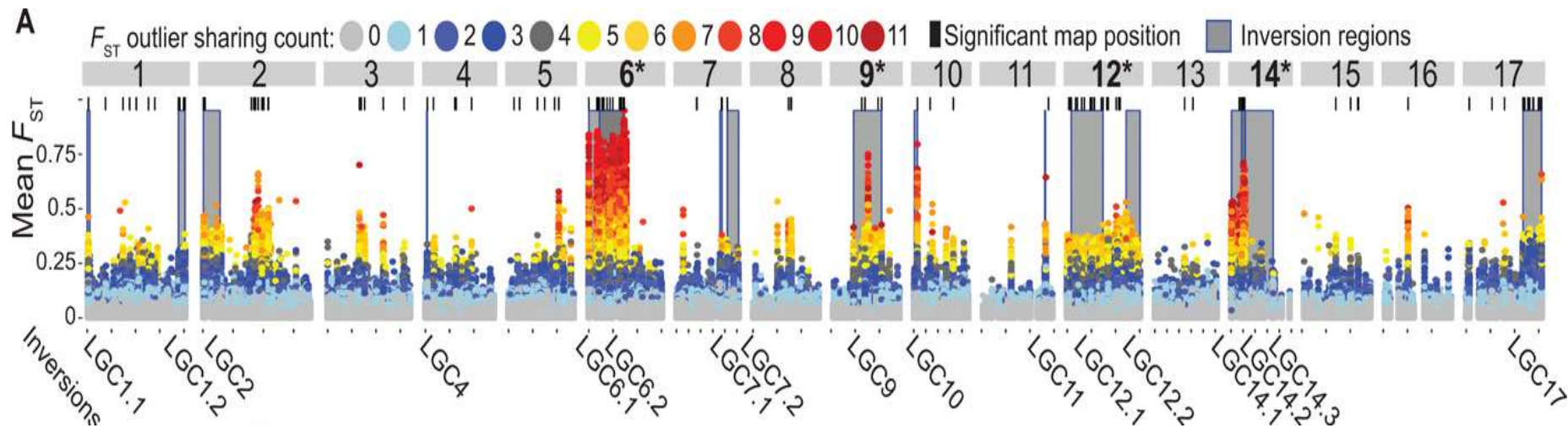
Outliers in pairs of *Littorina* ecotypes Crab vs Wave at different localities accross Europe...

Morales et al, 2019 Science advances
<https://doi.org/10.1126/sciadv.aav9963>



Outliers of divergence – signature of selection

Pairs of populations: contrast F_{ST}



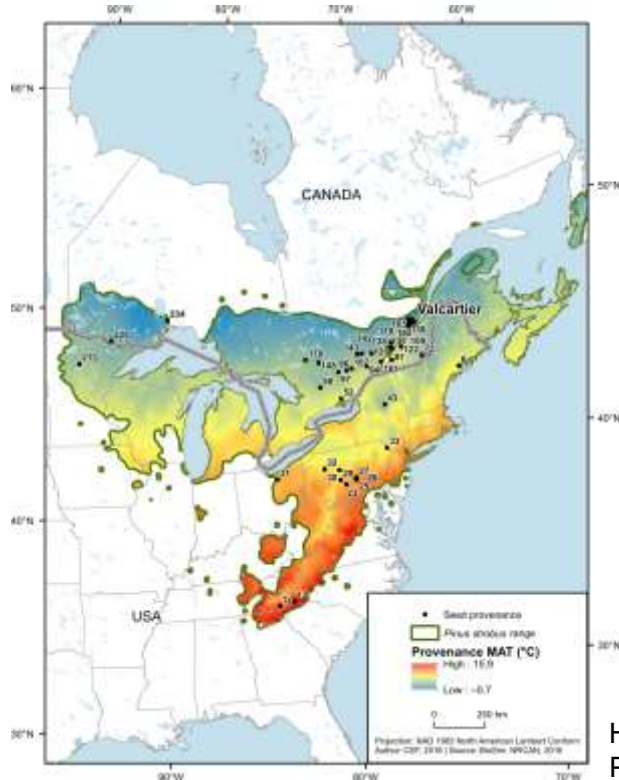
F_{ST} Crab vs Wave at different localities across Europe: outlier sharing

Morales et al, 2019 Science advances <https://doi.org/10.1126/sciadv.aav9963>



Outliers of divergence – signature of selection

Along a cline, accross a geographic gradient



Ecological information
+
Genetic information

⇒ hypothesis: populations experiencing the same selective pressures (same environment) will be less differentiated than populations experiencing different ecological conditions at adaptive loci

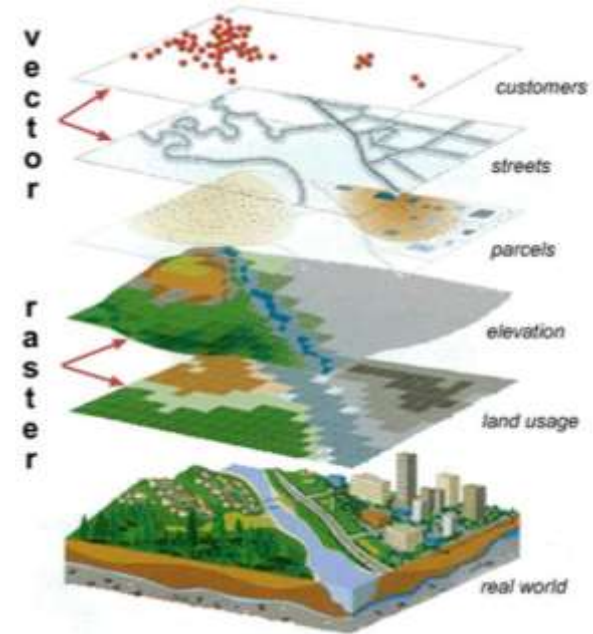
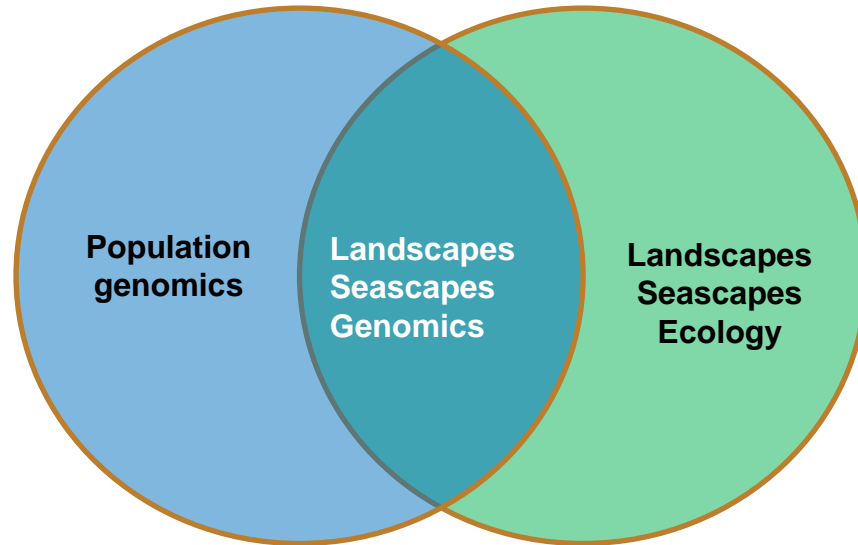
Housset, 2018, New
Phytologist

Landscapes genomics and environmental associations

Lanscapes or seascapes genomics contribute to understand both neutral and adaptive genetic variation

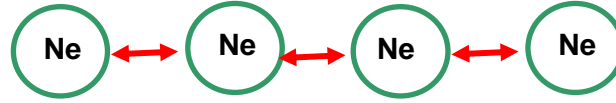
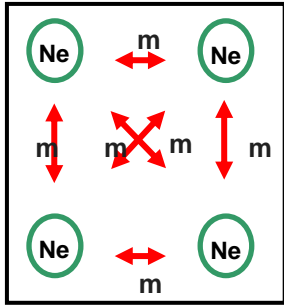
Genetic diversity

- Many samples
- Many populations



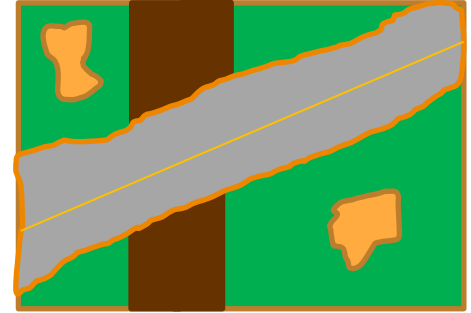
Landscapes genomics and environmental associations

- a better understanding of gene flow & connectivity

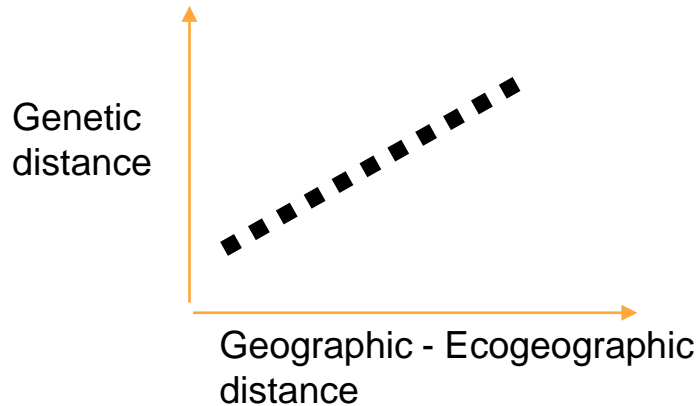


Isolation-by-distance?

Isolation-by-distance
along least-cost-path?

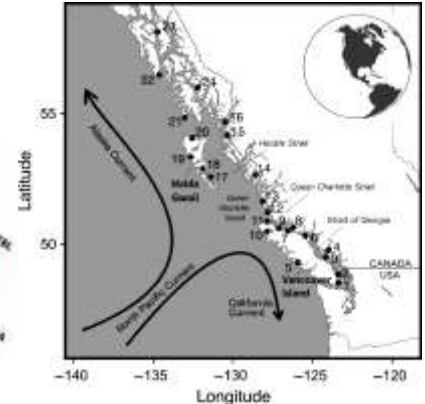
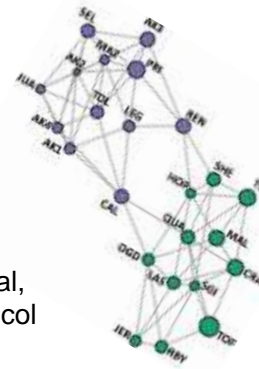


Isolation-by-resistance?



Complex
connectivity

Xuereb et al,
2018 MolEcol

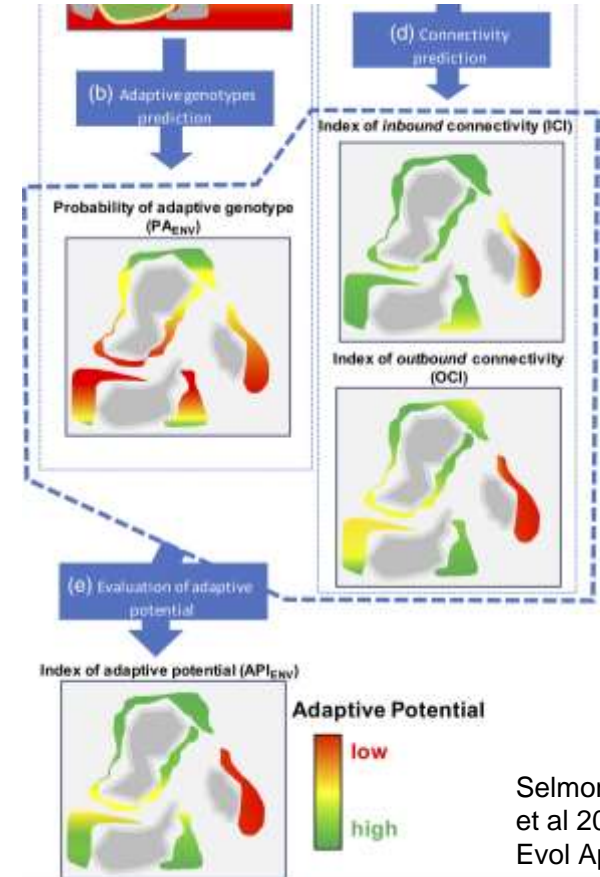
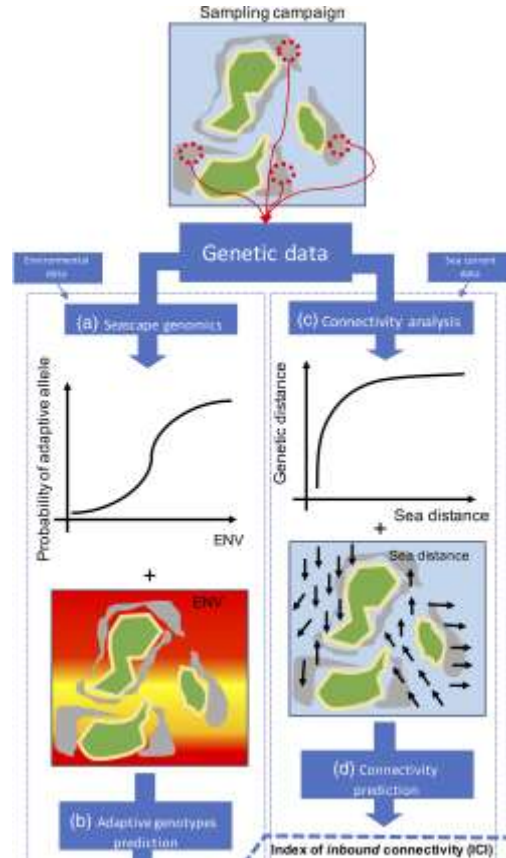


Landscapes genomics and environmental associations

- Analyzing adaptation to known ecological variables

⇒ A better sense of why populations are differentiated?

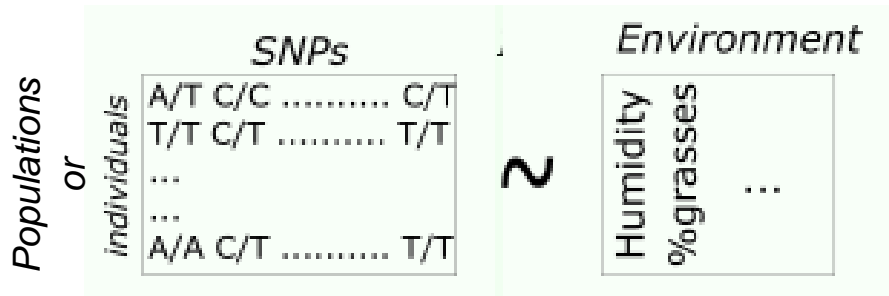
⇒ Predictions for models of adaptation, response to climate change, etc



Landscapes genomics and environmental associations

Environmental associations methods

$G \sim E$ (+ correction?)



- Univariate methods : locus by locus

(correlation freq/env, LFMM, Bayenv/Baypass, etc)

Genome scan methods against more complex models: when and how much should we trust them?

(Villemereuil et al, 2014 MolEcol)

- Multivariate methods

(redundancy analysis RDA)

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

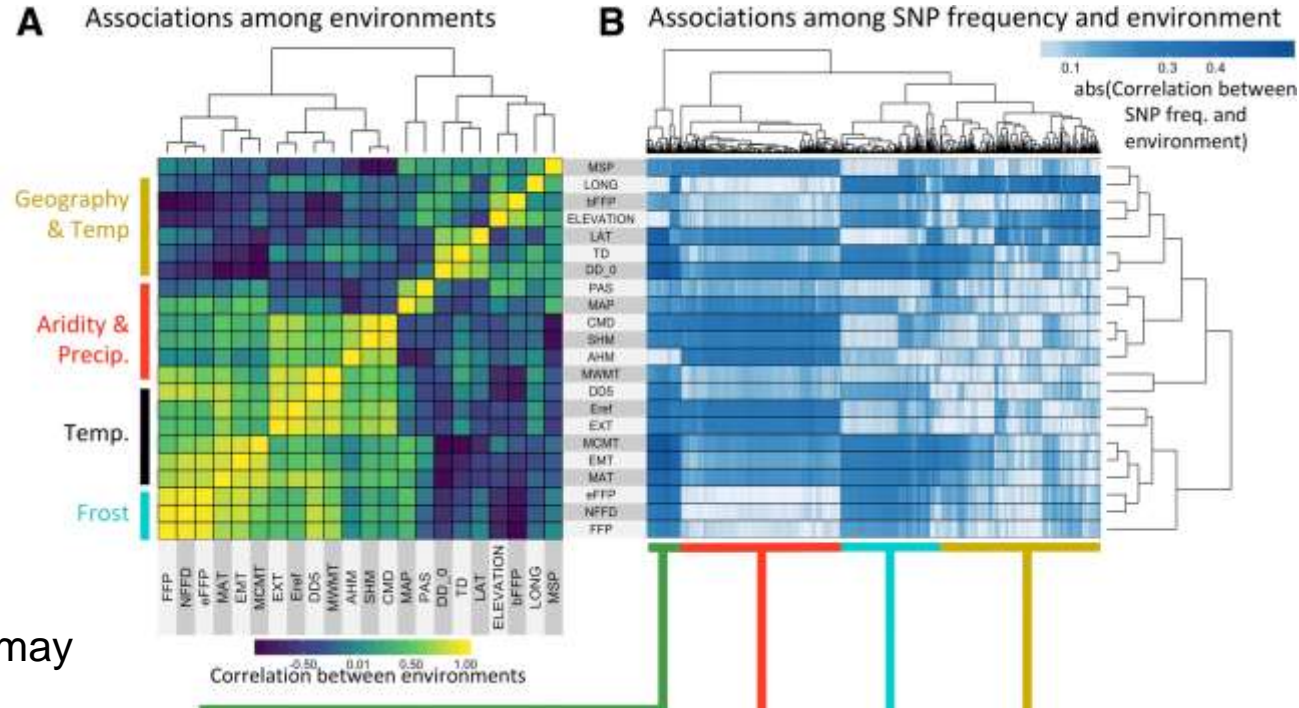
Landscapes genomics and environmental associations

Univariate associations:

Which locus is associated with which environmental variable?
-> Spearman's correlation

Caution:

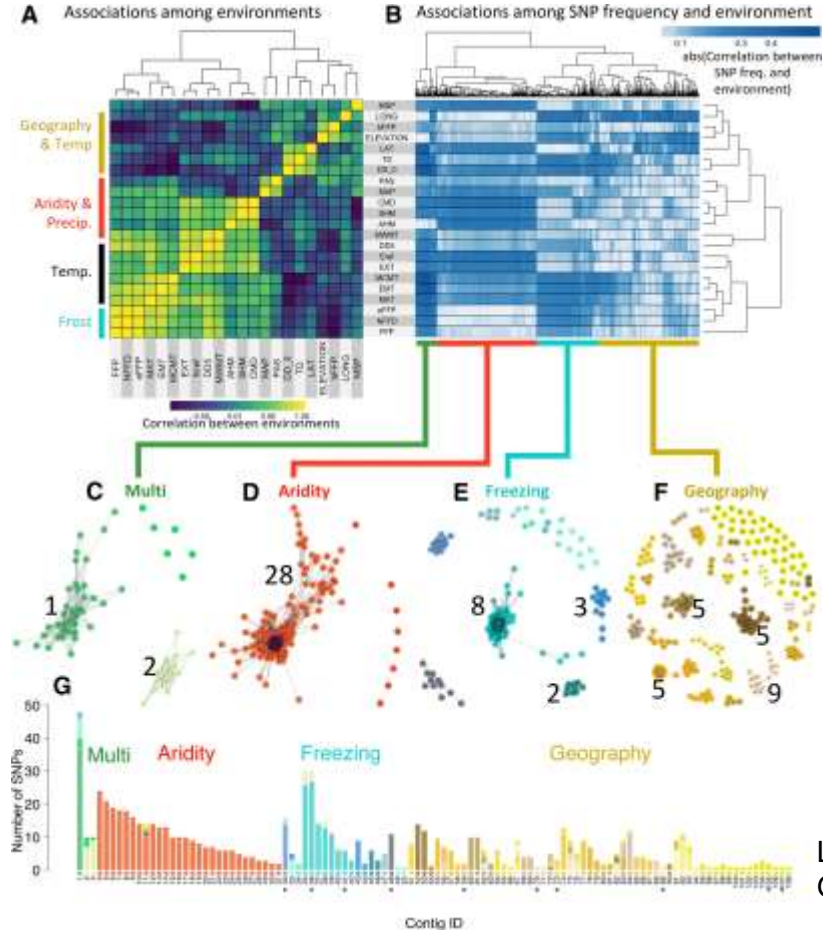
- environmental variables may be correlated!
- SNPs may be in physical LD



Landscapes genomics and environmental associations

Univariate associations:

Modular group of adaptive loci to different axis of environmental variation

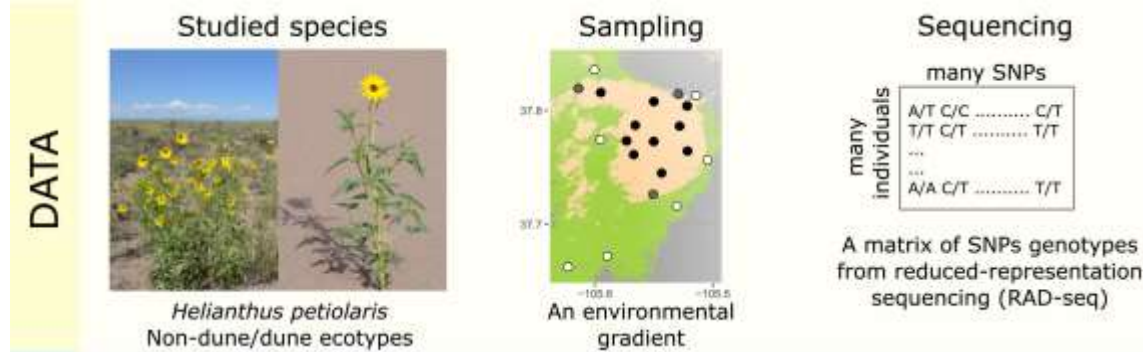


Lotterhos et al, 2018
Genome Biology

Landscapes genomics and environmental associations

Univariate associations:

Bayenv2 – Baypass

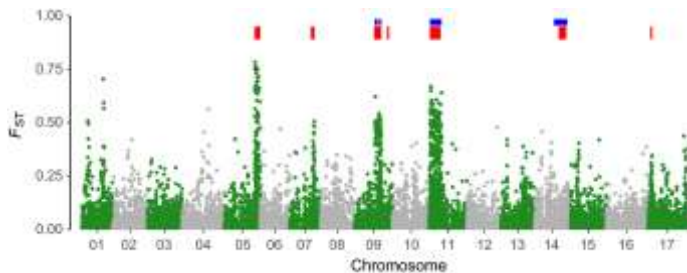


Huang et al,
2020
MolEcol

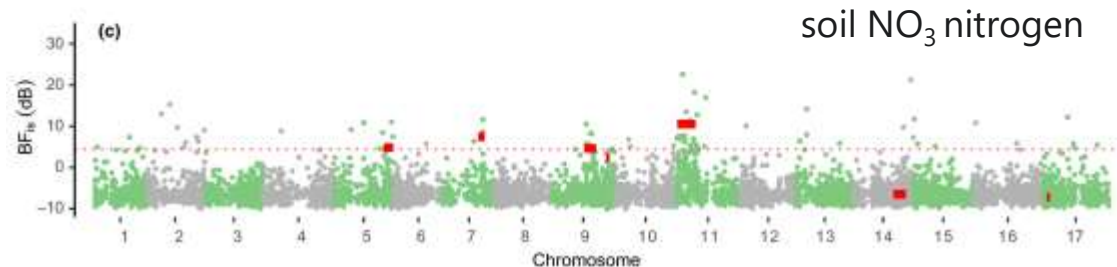
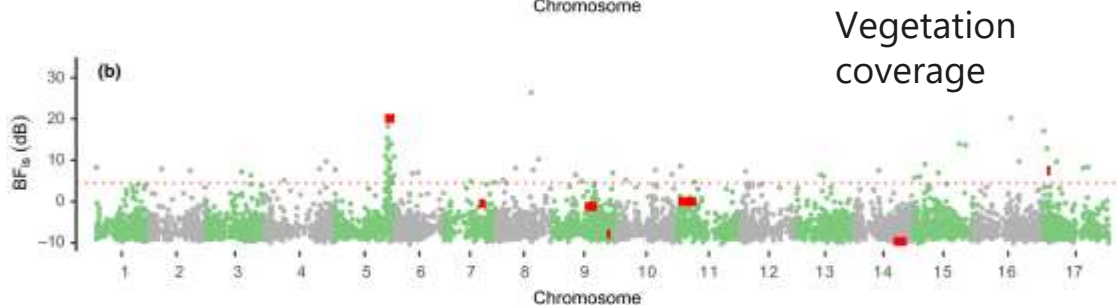
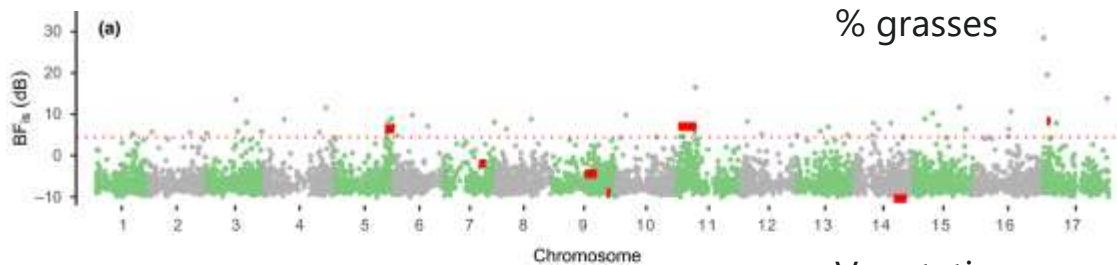
Landscapes genomics and environmental associations

Univariate associations:
Bayenv2 – Baypass

FST outliers



Env. Assoc.

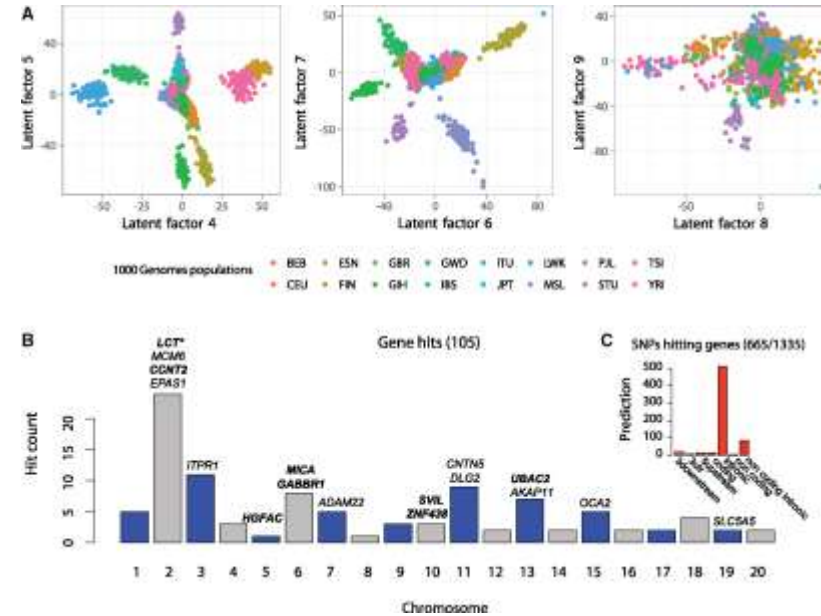


Huang et al,
2020
MolEcol

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
 - Latent factors represent demographic history, IBD, hidden substructure
 - Lfmm2: New & faster version which can work on SNP or methylation matrix

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



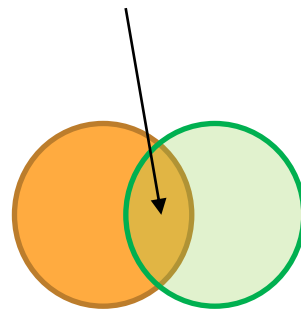
Human GEA study. Association study based on genomic data from the 1000 Genomes Project database and climatic ...

Landscapes genomics and environmental associations

Univariate associations:

Recommendation:

- intersect of 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?
 - ⇒ Likely depends on the system: high gene flow? IBD? Geography correlated with environmental variation?



Controlling false discoveries in genome scans for selection

François et al, 2015, Molecular Ecology

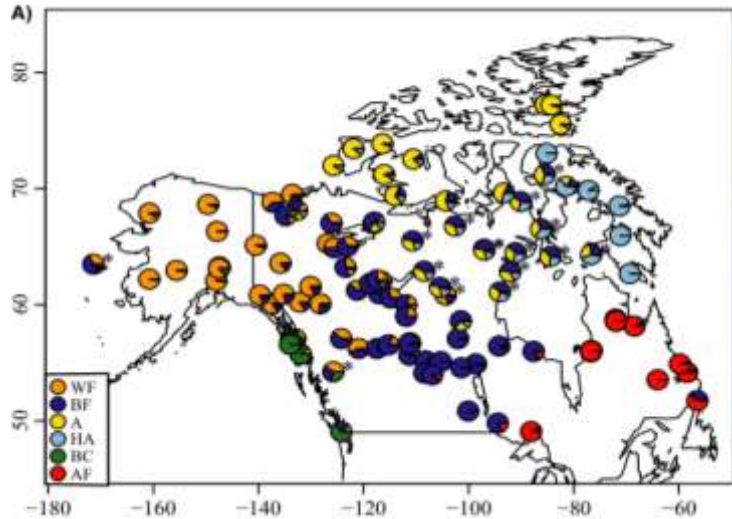
<https://doi.org/10.1111/mec.13513>

Landscapes genomics and environmental associations

Multivariate associations:

94 wolves

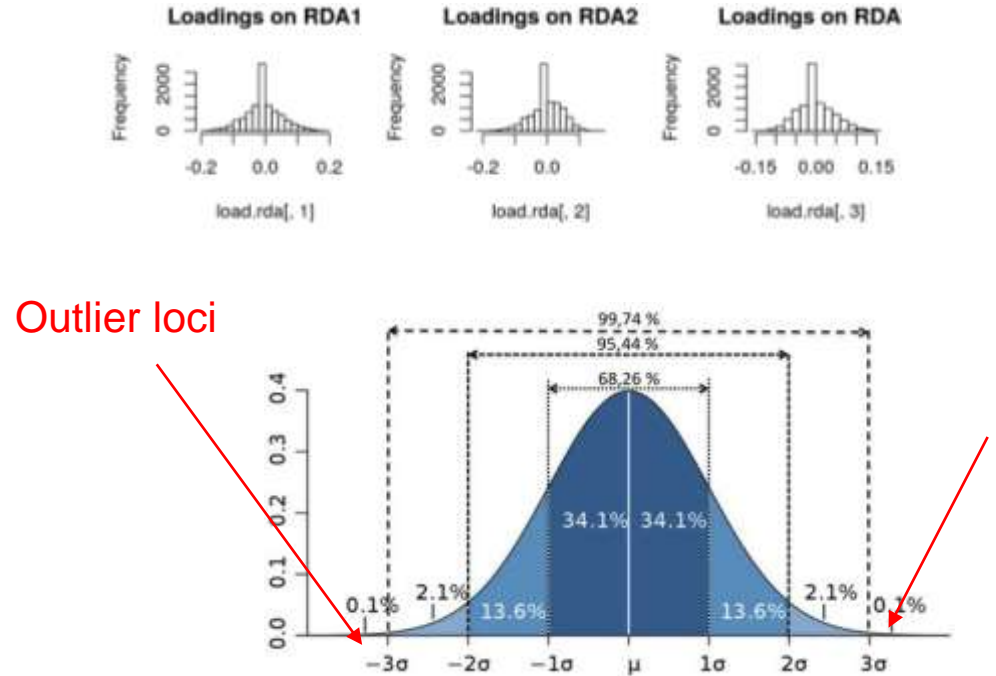
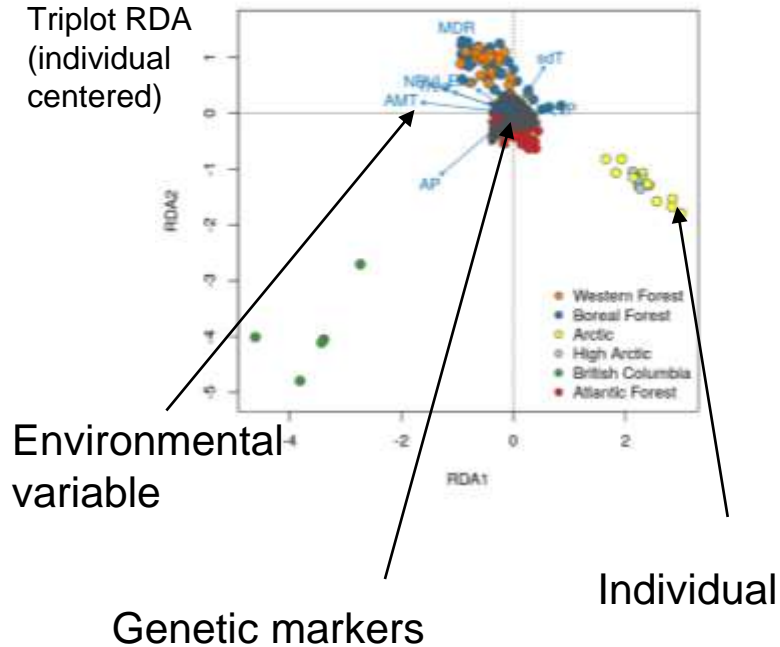
42 597 SNPs



Forester et al 2018
Mol Ecol

Landscapes genomics and environmental associations

Multivariate associations:



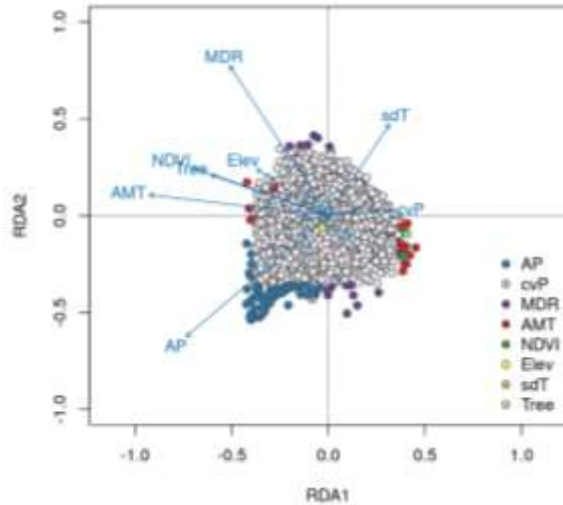
Forester et al 2018
Mol Ecol

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

Landscapes genomics and environmental associations

Multivariate associations:

Triplot RDA
(marker
centered)



Forester et al
2018 Mol Ecol

The color represent the
environmental variable the most
correlated to the marker

Multi-locus

-> Polygenic adaptation possible to
detect?

Multi-variable analysis

-> realistic environmental
characterisation

Very fast + Global information:

- Which variables explain genetic
variance?
- Correction possible by
population/geographic structure

Landscapes genomics and environmental associations

A must-read:

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>