

Landscape genomics

30/05/2025

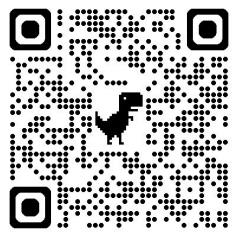
Physalia course

Yann Bourgeois, Thibault Leroy

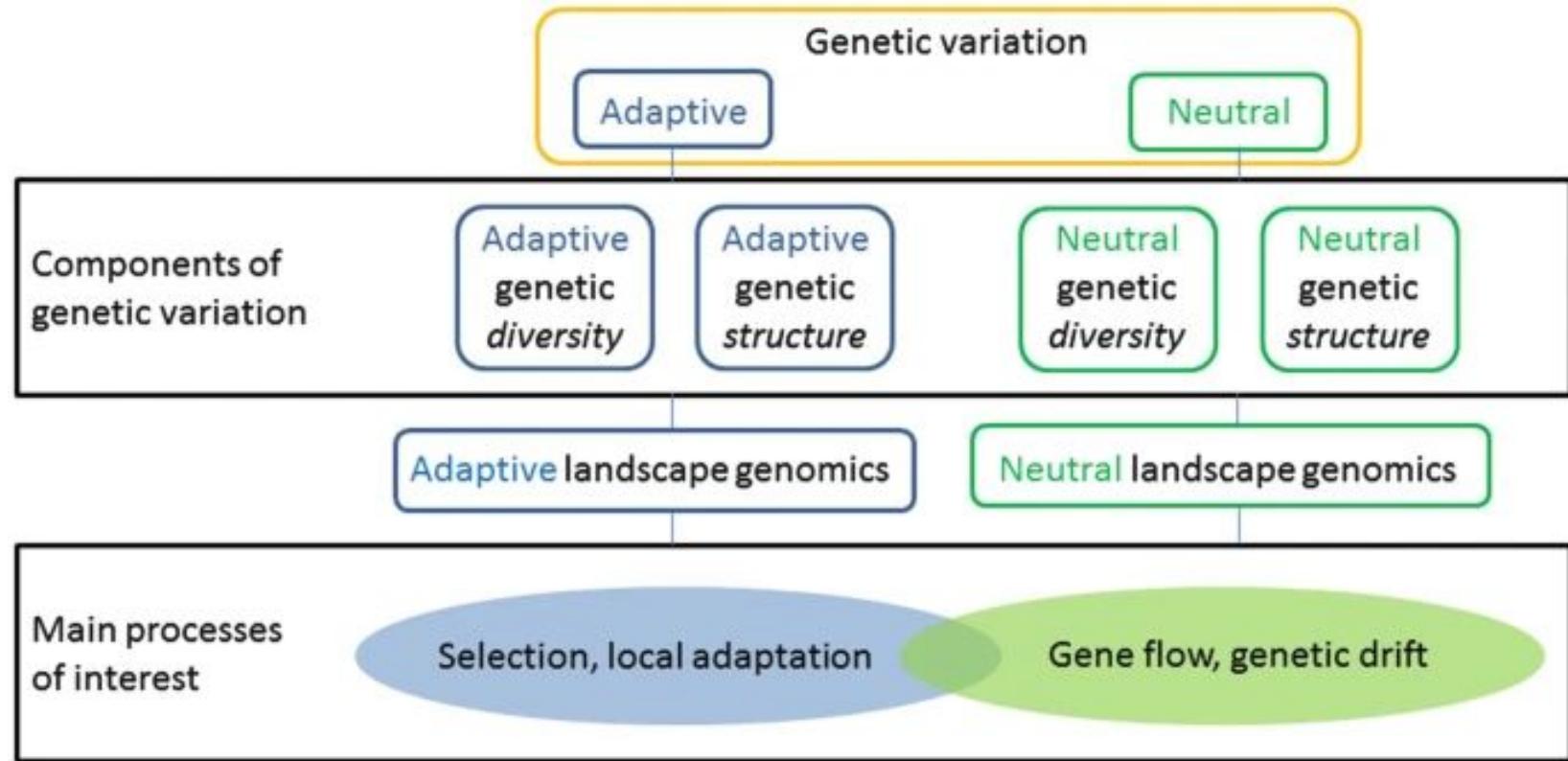


Goals for today's lecture

- How can we incorporate spatial information in popgen?
- How can we detect barriers to dispersal?
- How can we test hypotheses about the role of landscape features?
- How can we identify signatures of local adaptation using genotype-environment associations?
- Can GEA provide insights into (mal)adaptation in the context of climate change?
- And throughout, as usual: a few limitations to keep in mind.

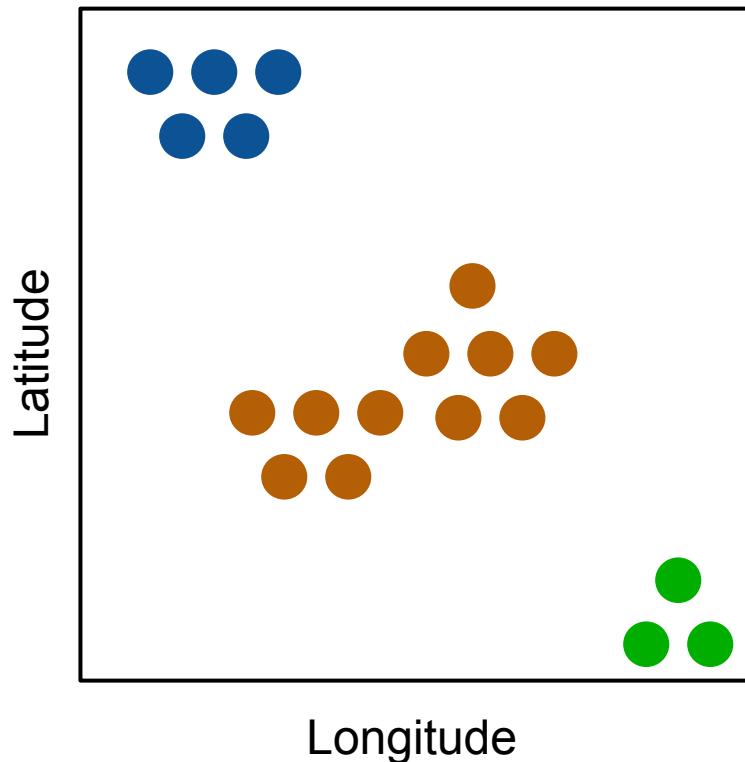


General framework

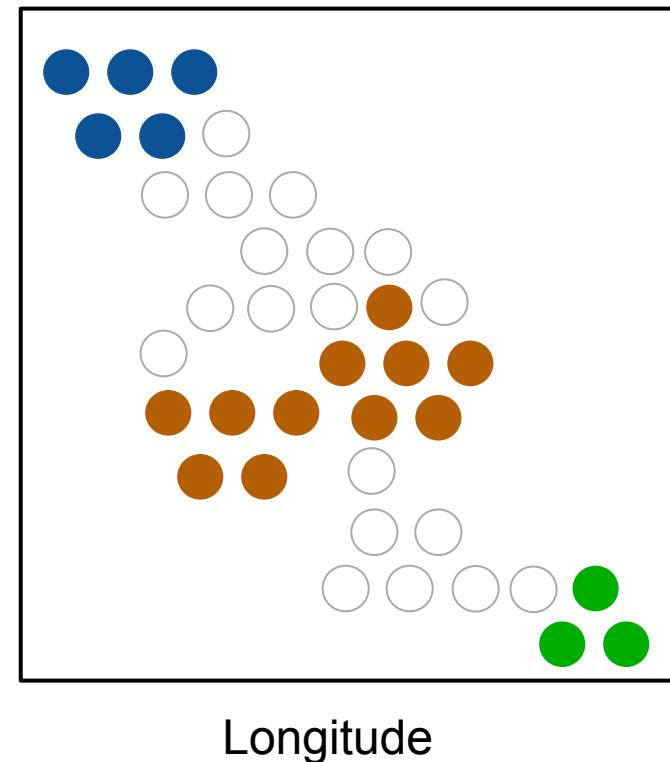


Non-spatial vs. spatial (neutral) population structure: introducing the issue

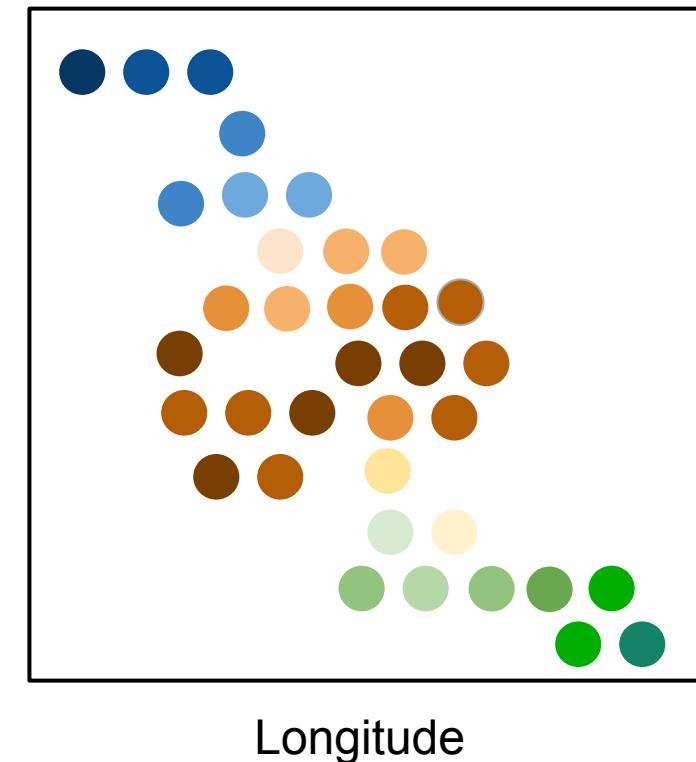
Non-spatial clustering (e.g. Structure)



Reality: sampling discontinuities

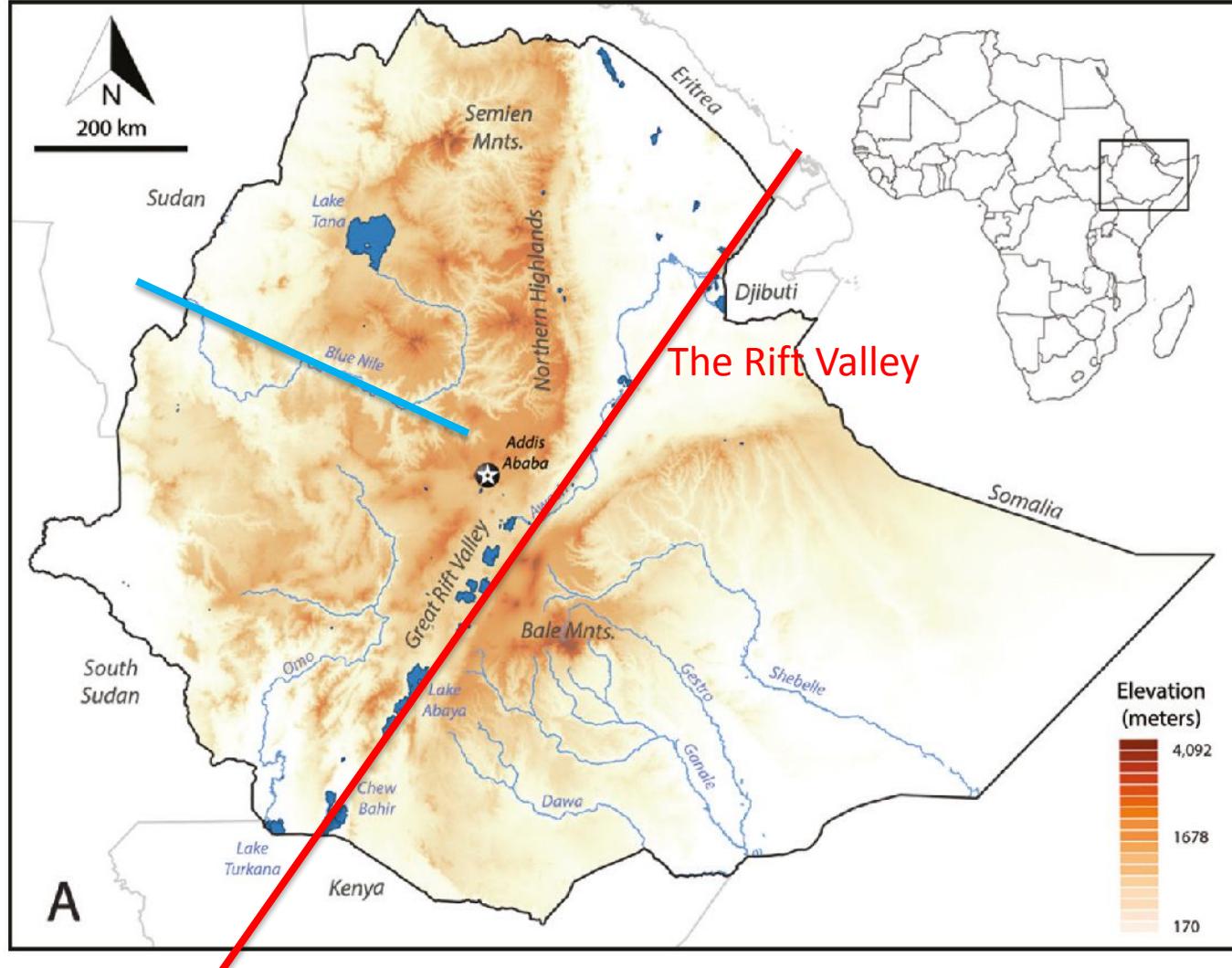


Reality: spatial structure



Spatializing structure analyses (Tuesday lecture, in space!)

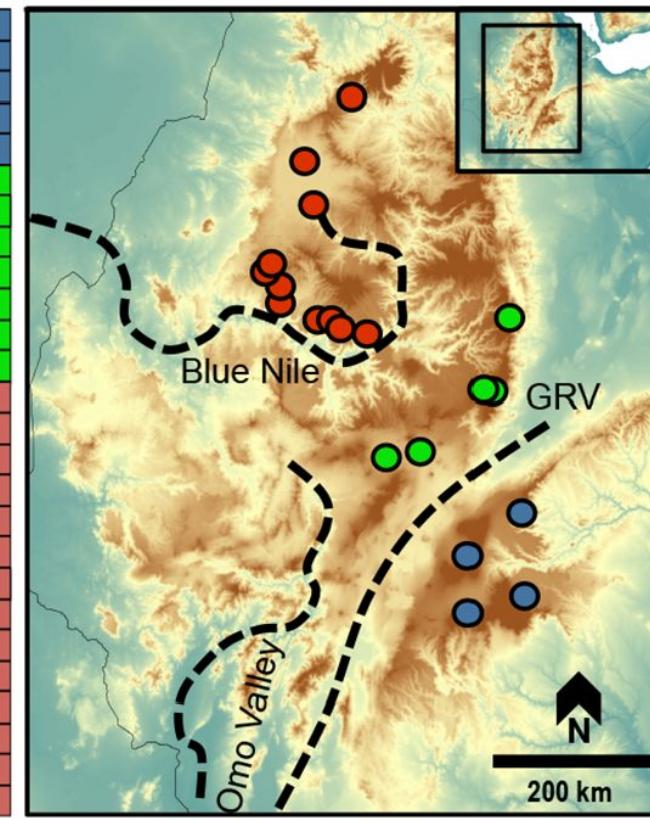
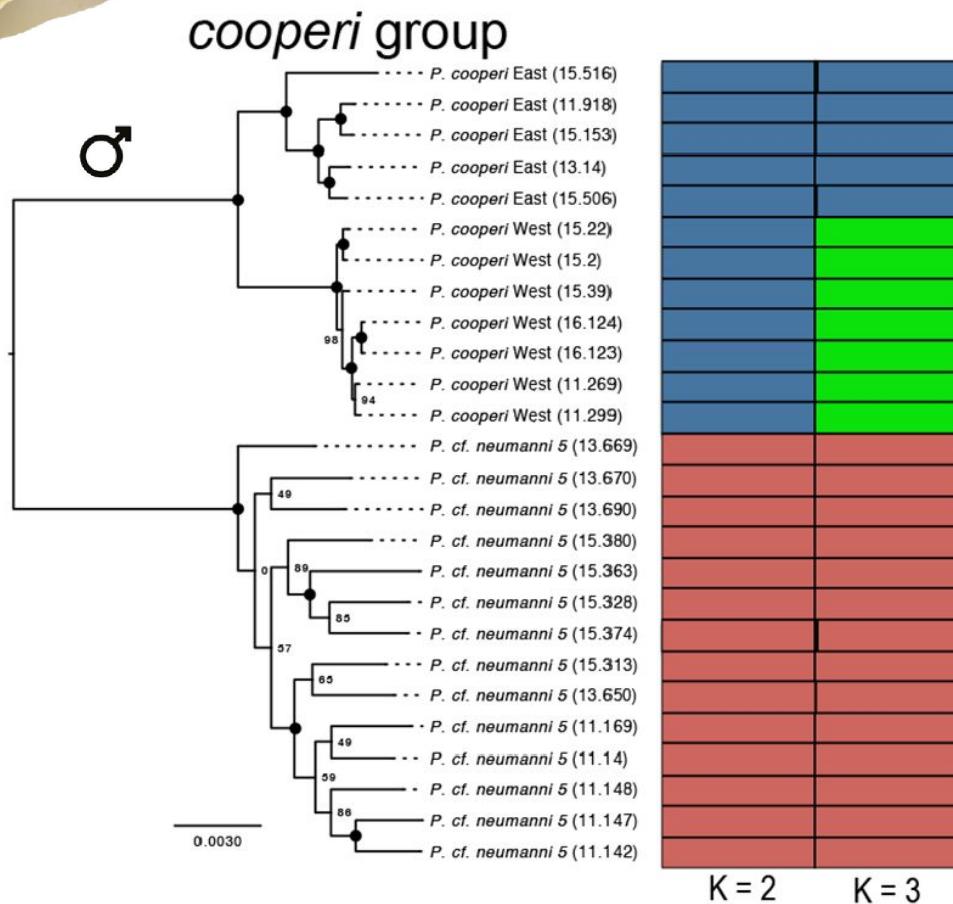
The Blue Nile

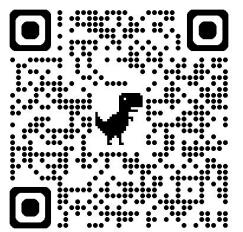


Spatializing structure analyses

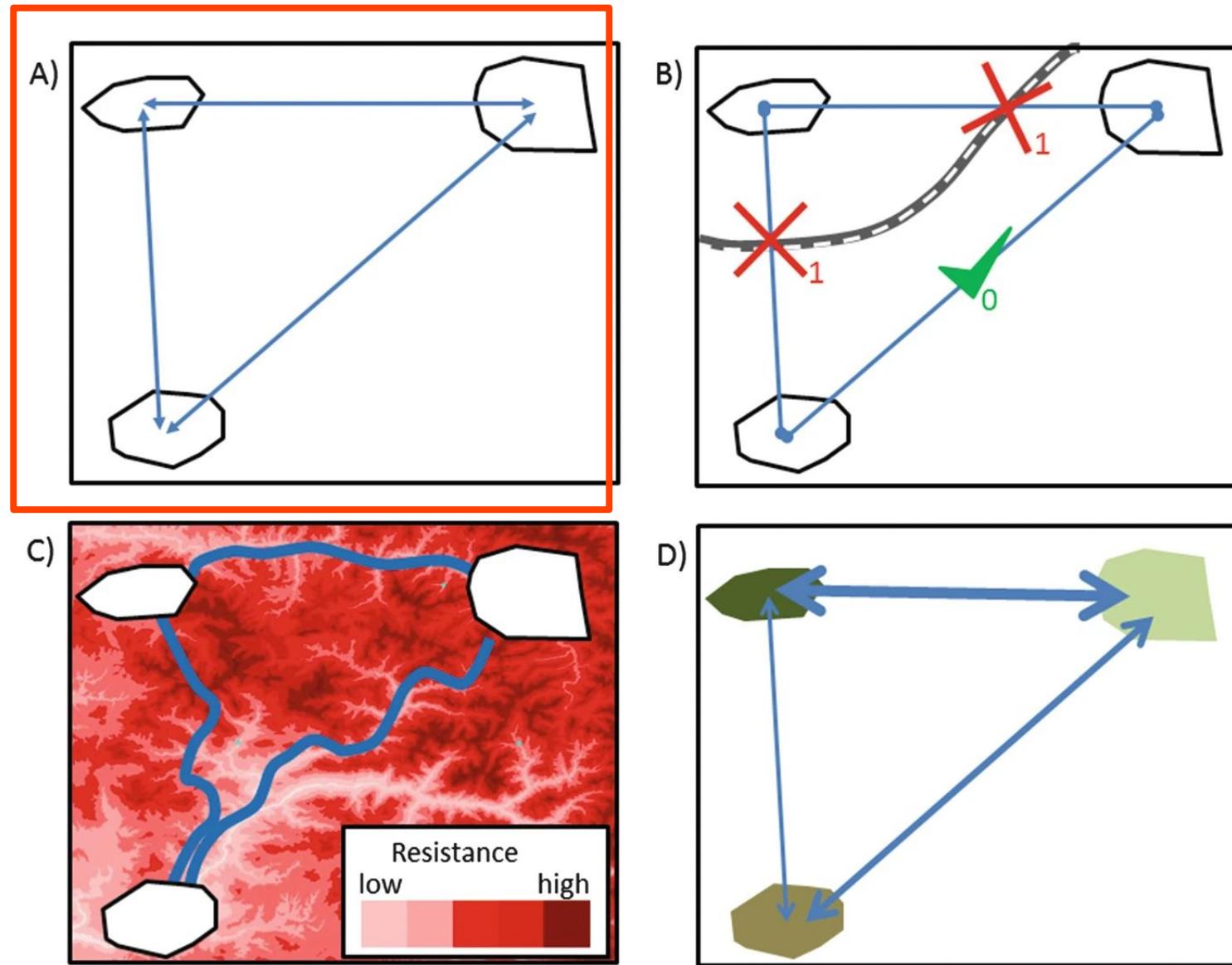


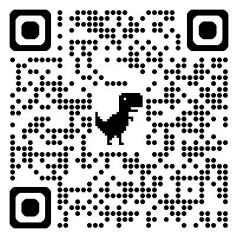
*Ptycha
dena
cooperi*



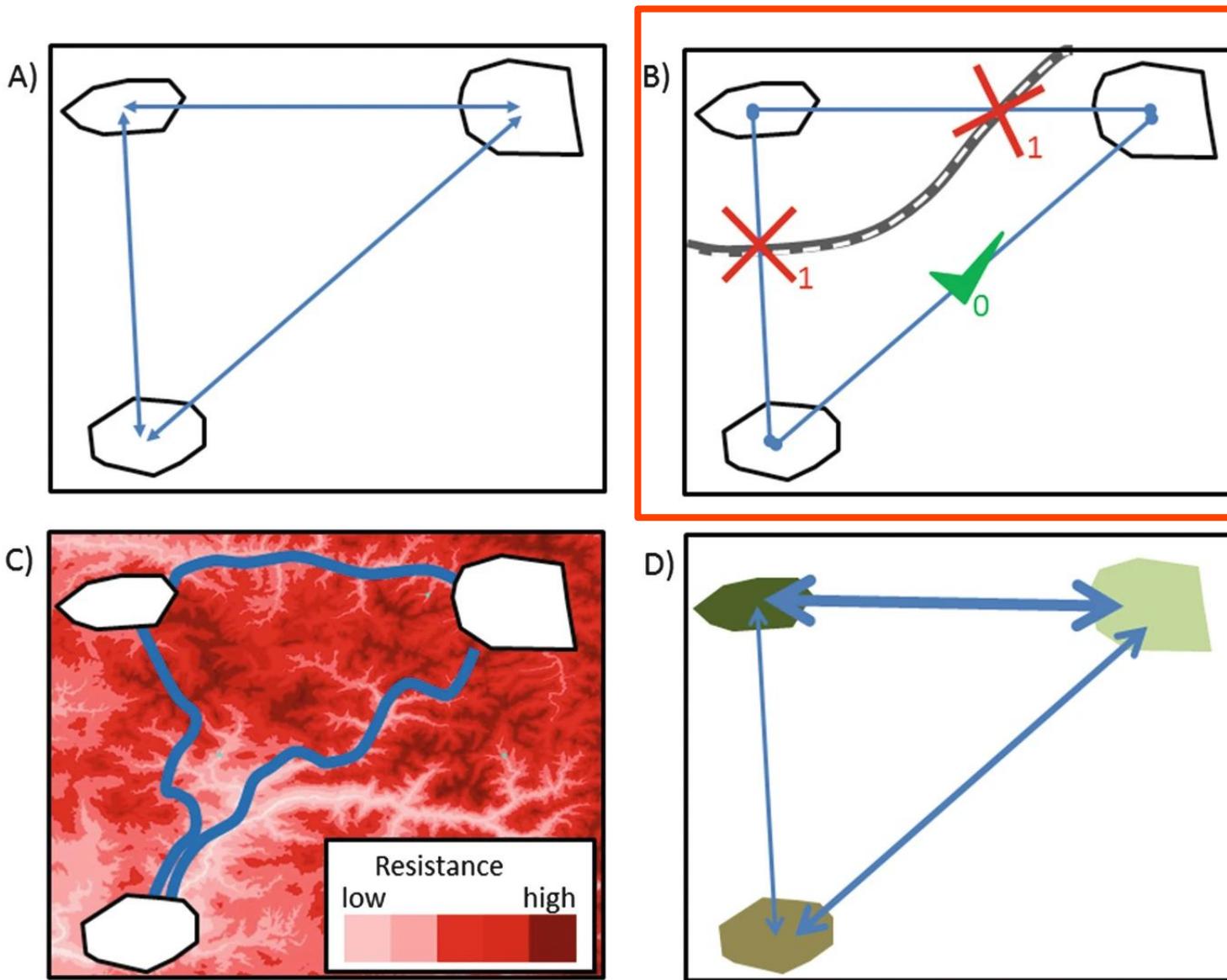


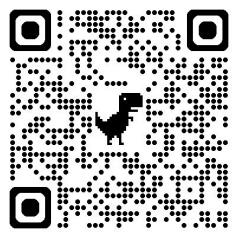
Going a bit further: Isolation-by-Distance



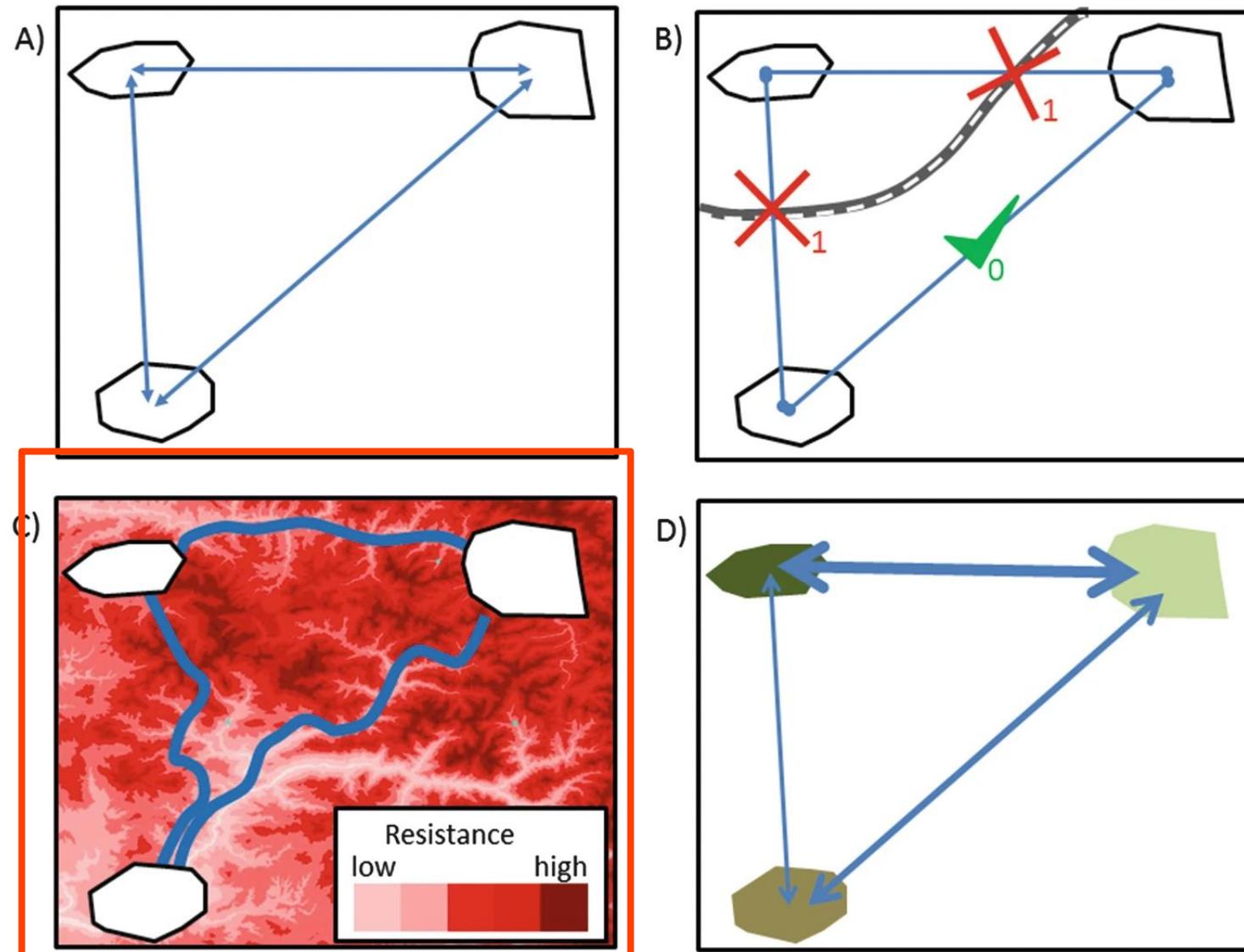


Going a bit further: Isolation-by-Barrier



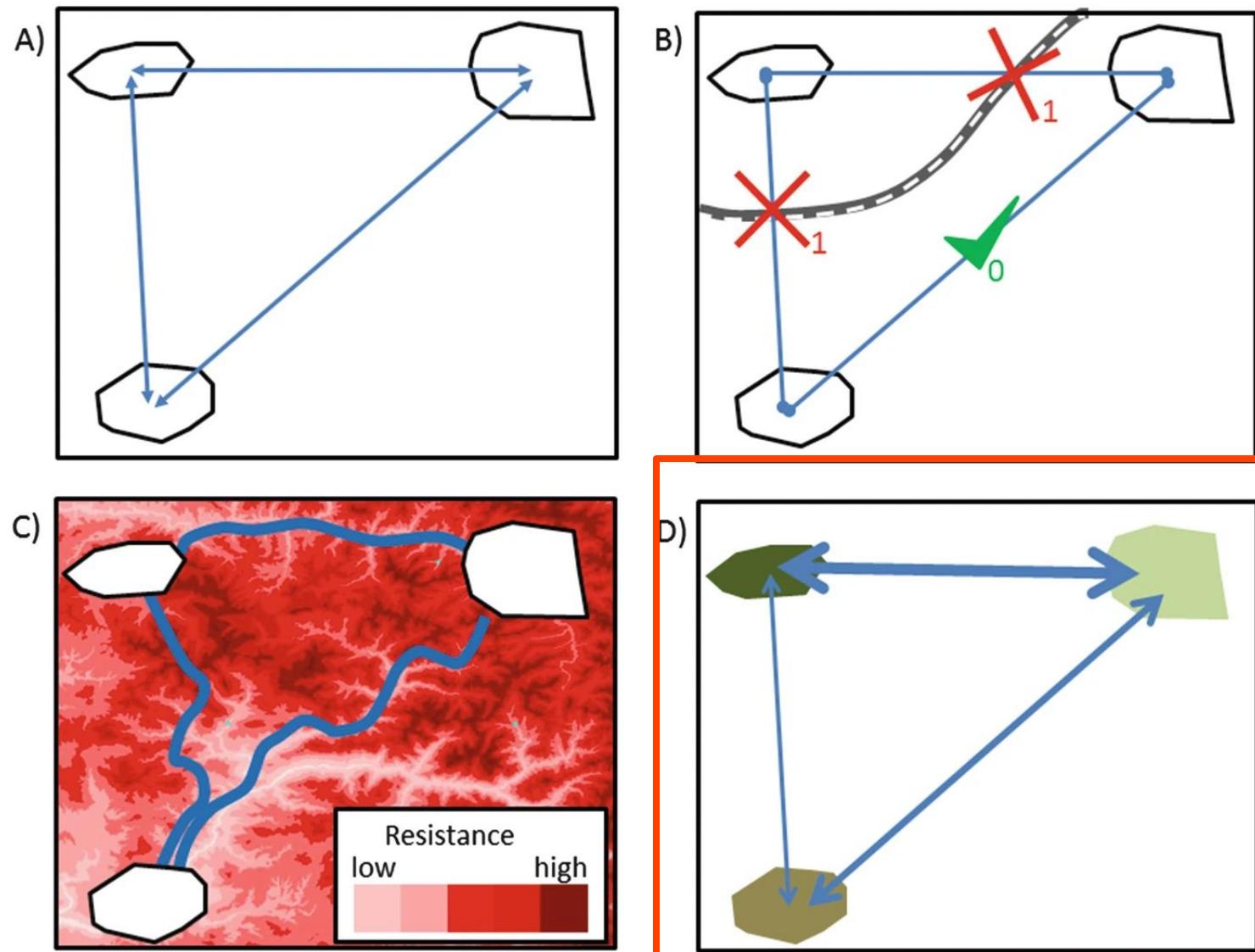


Going a bit further: Isolation-by-resistance



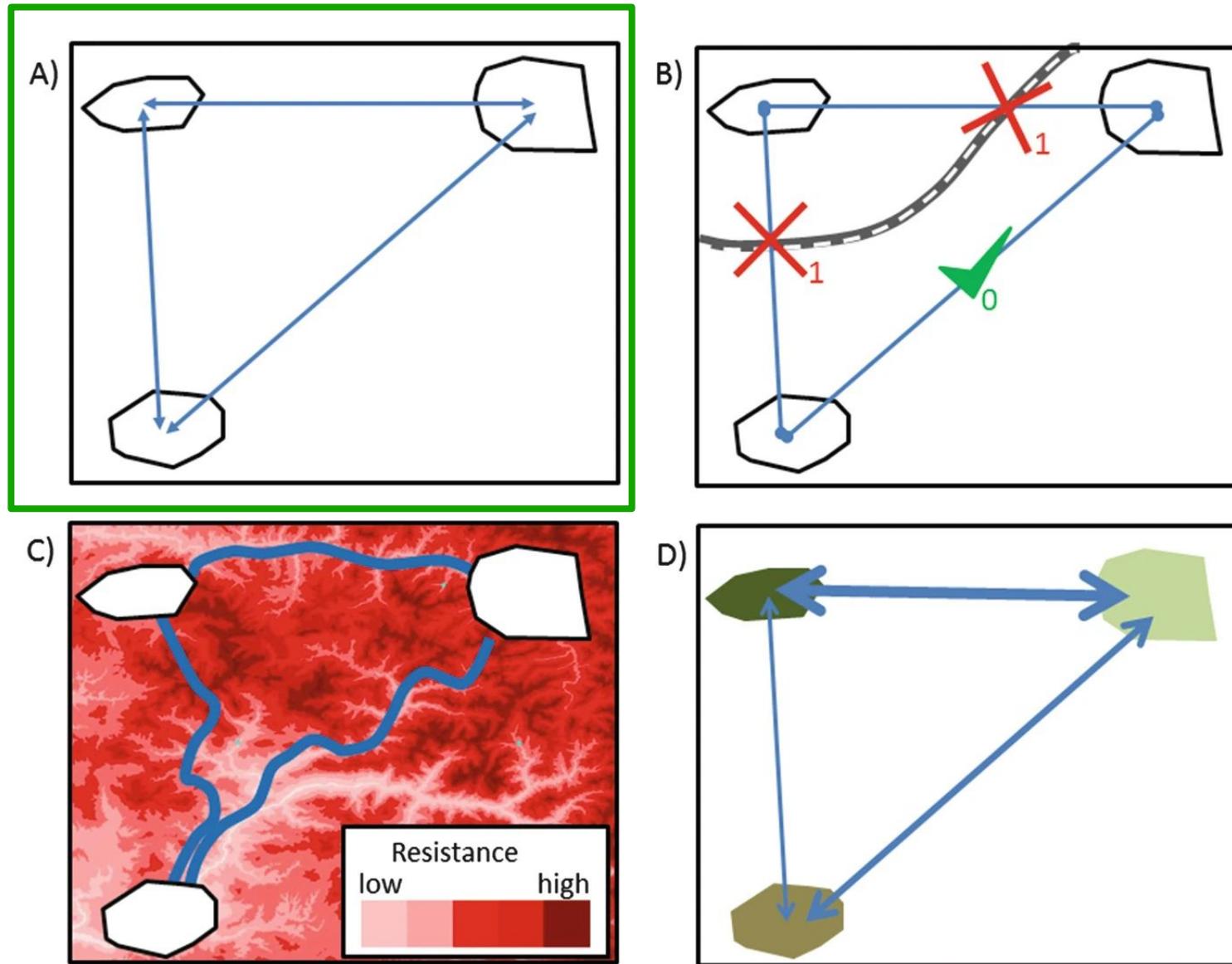


Going a bit further: Isolation-by-Environment

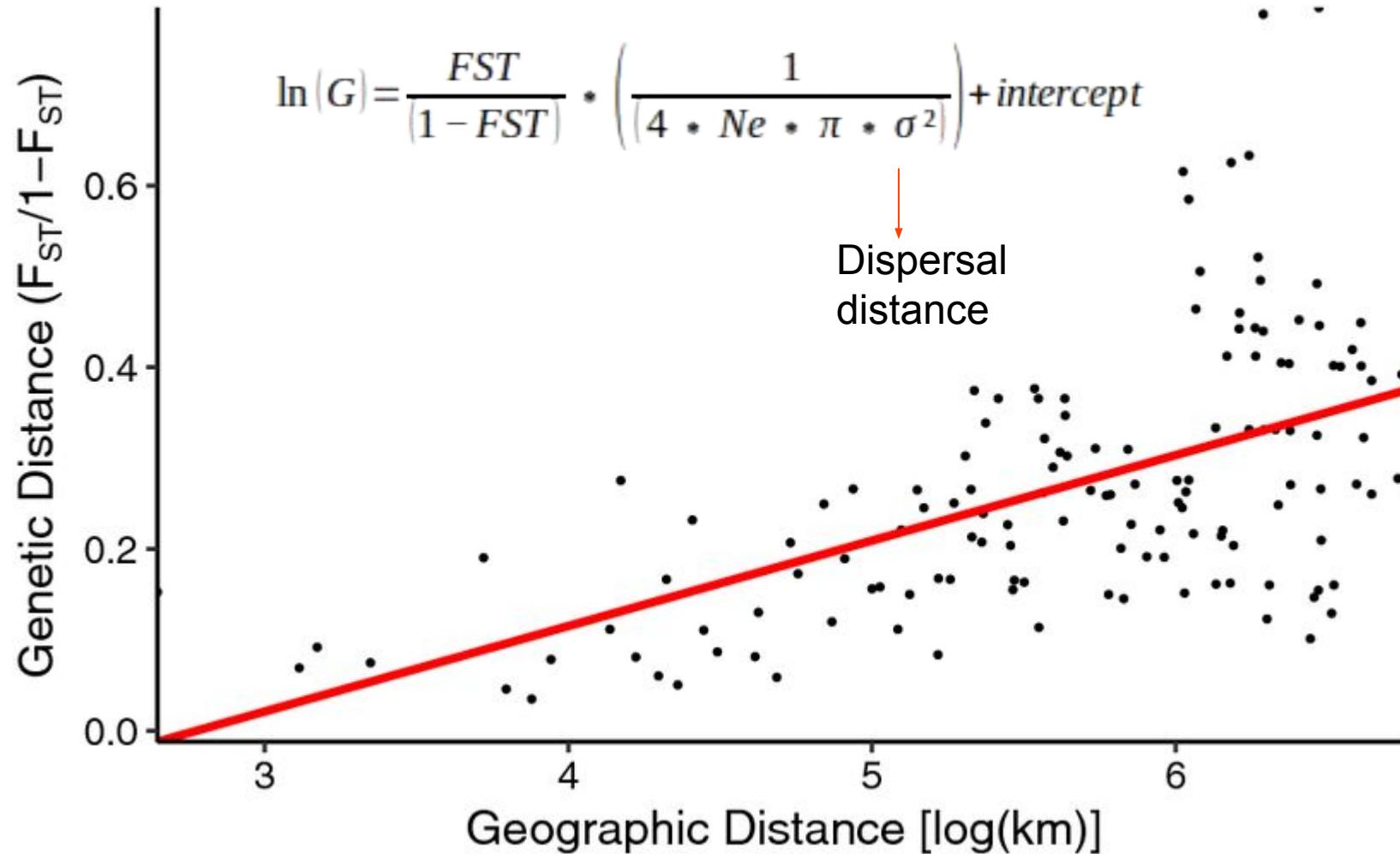




Let's start with Isolation-by-Distance



Isolation by distance



Mantel tests.

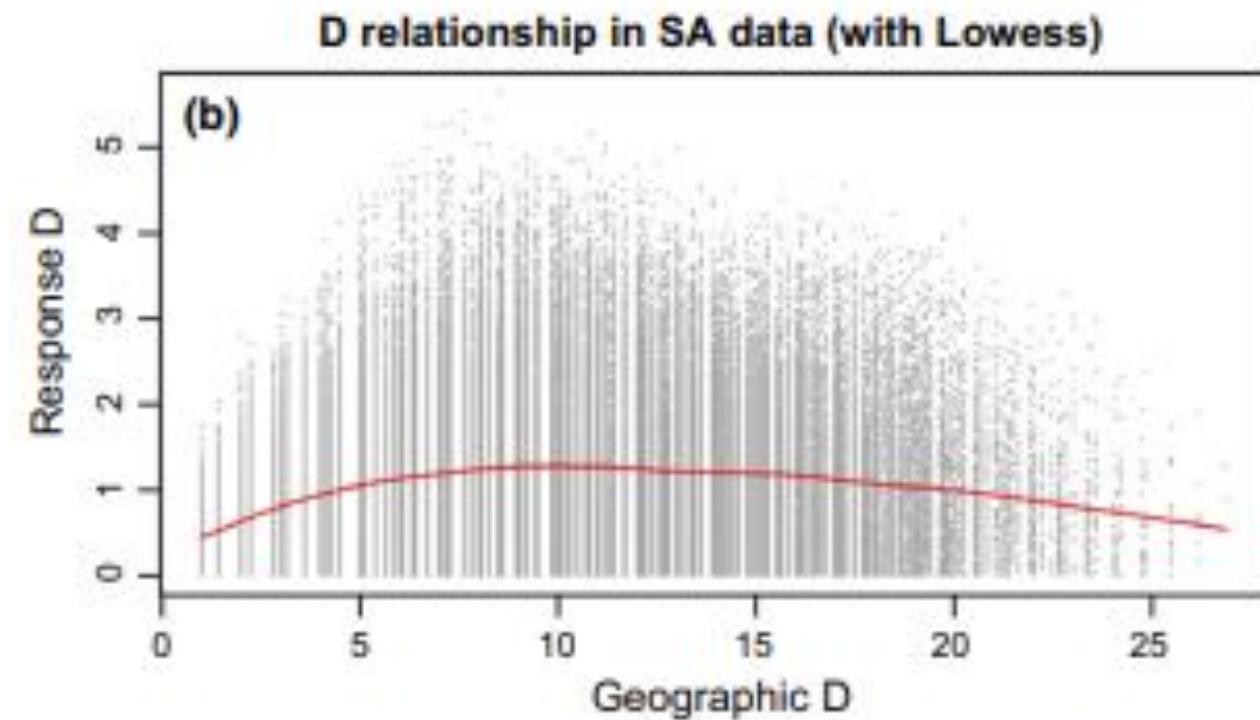
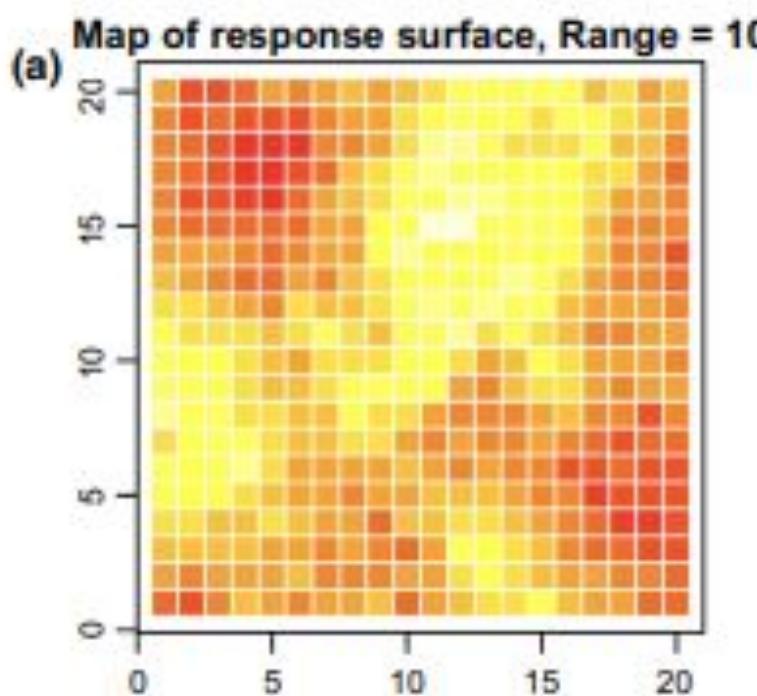
- Compares dissimilarity matrices (genetic distance G v. geographical distance S)
- Partial Mantel tests can be used to test for both environment E and geography S : $G \sim S + E$

Mantel tests: Limitations

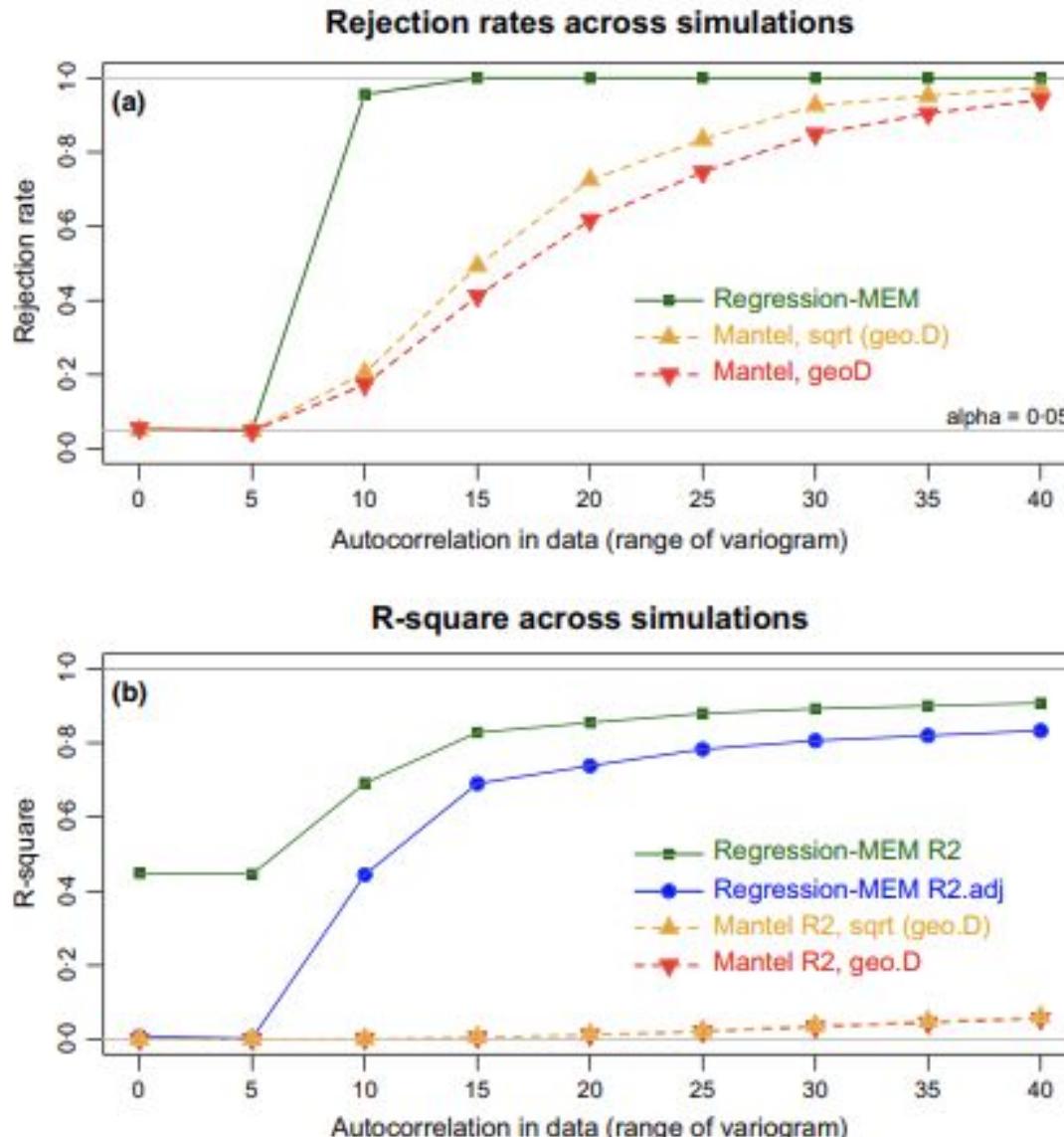
- Does not provide information about where in space autocorrelation varies (~ a bit crude)
- Not very sensitive.
- Mantel's R^2 is NOT equivalent to R^2 in regressions
- The null hypothesis H_0 : No linear or monotonic relation of distances among objects provided in the two/three matrices.



Mantel tests: Limitations.

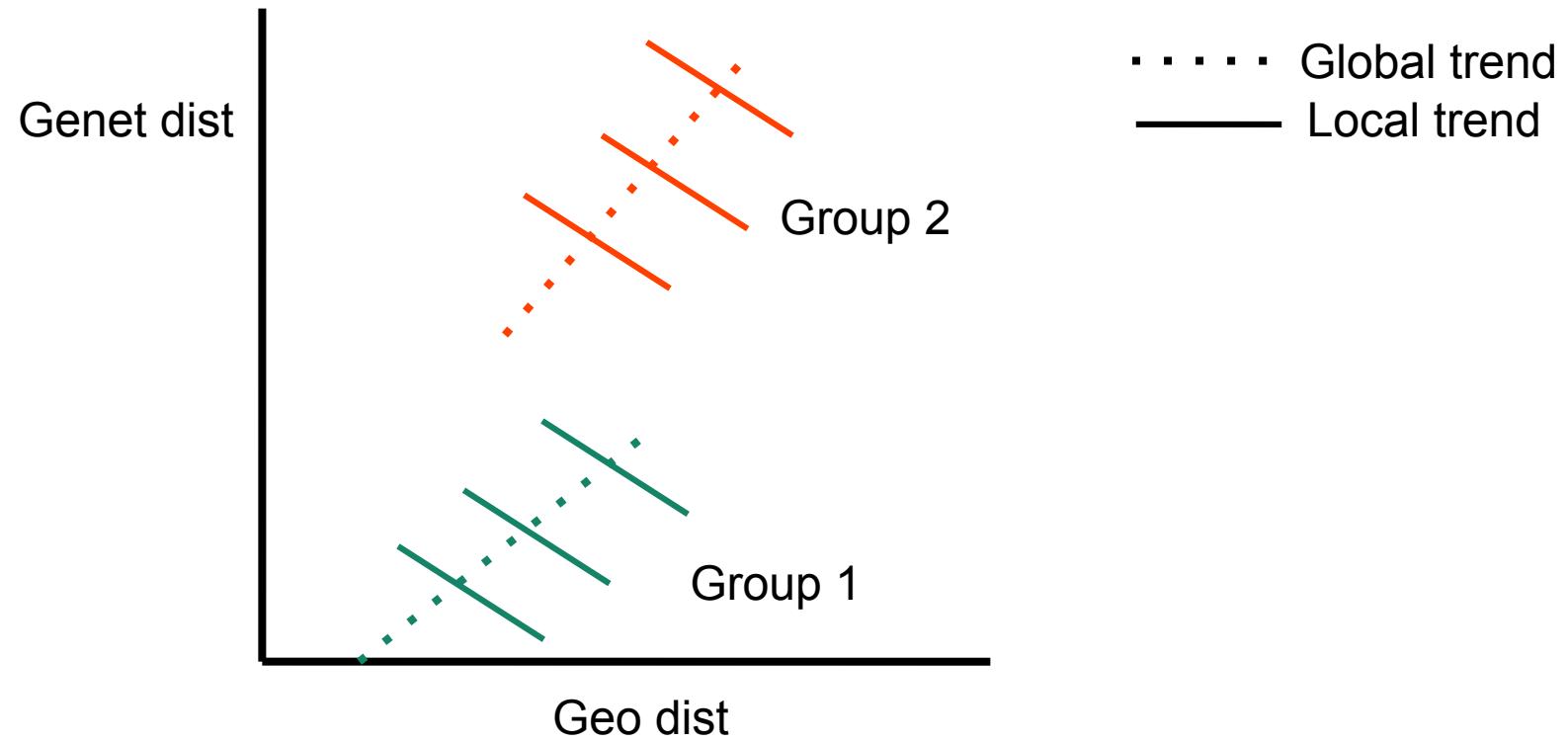


Mantel tests: Limitations.

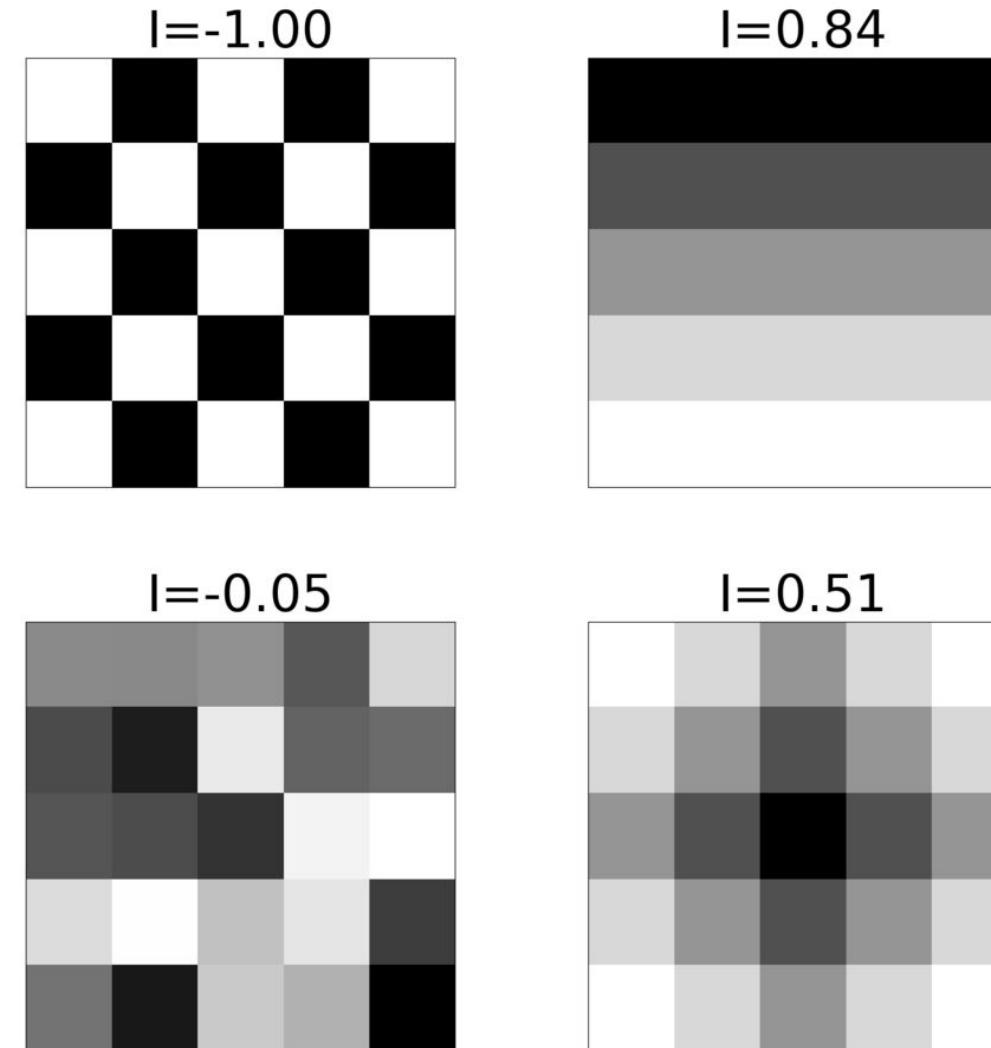


Limitations to Mantel test

- What to do when structure and autocorrelation varies across scales?

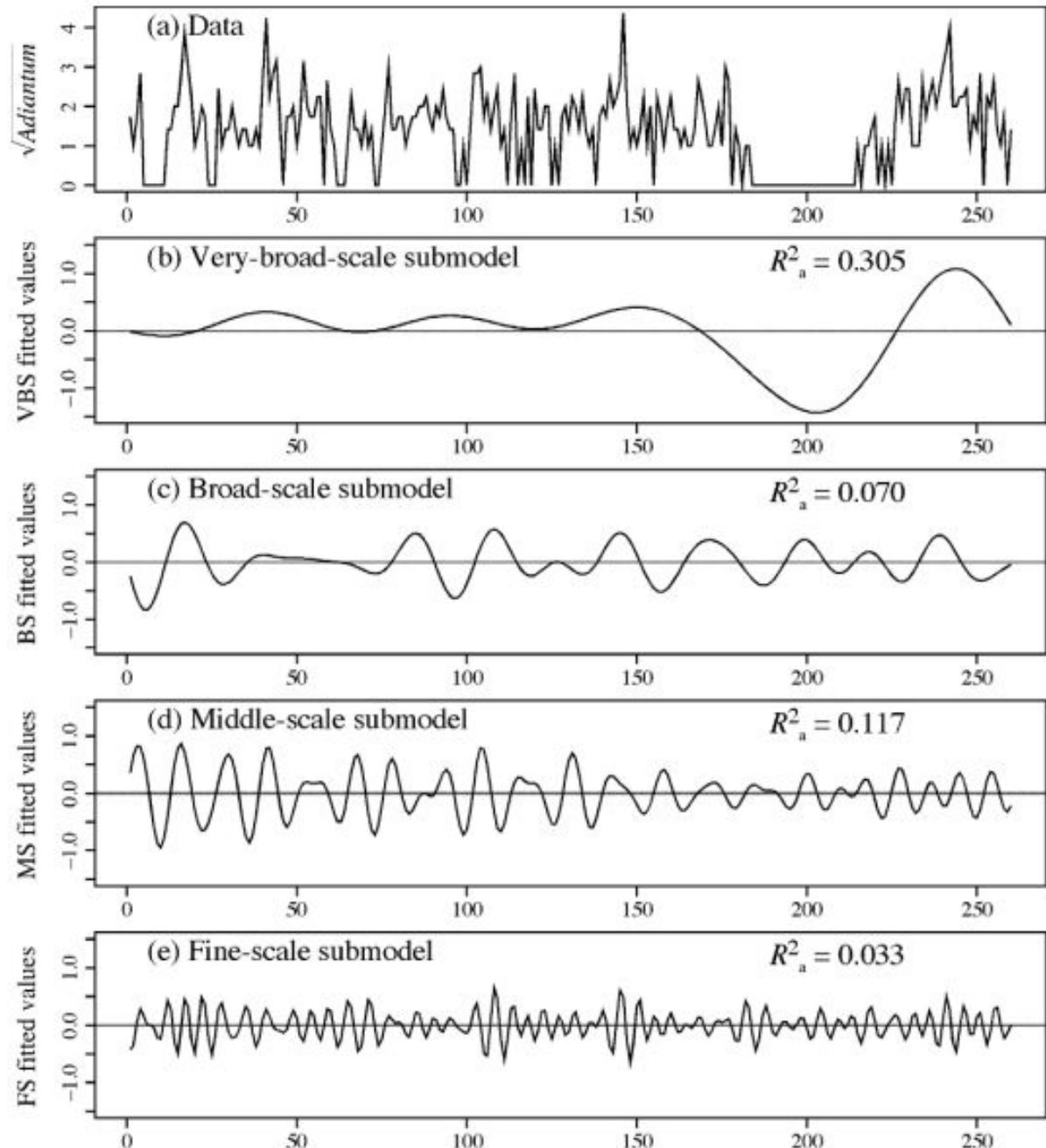


Describe spatial autocorrelation. Moran's I.

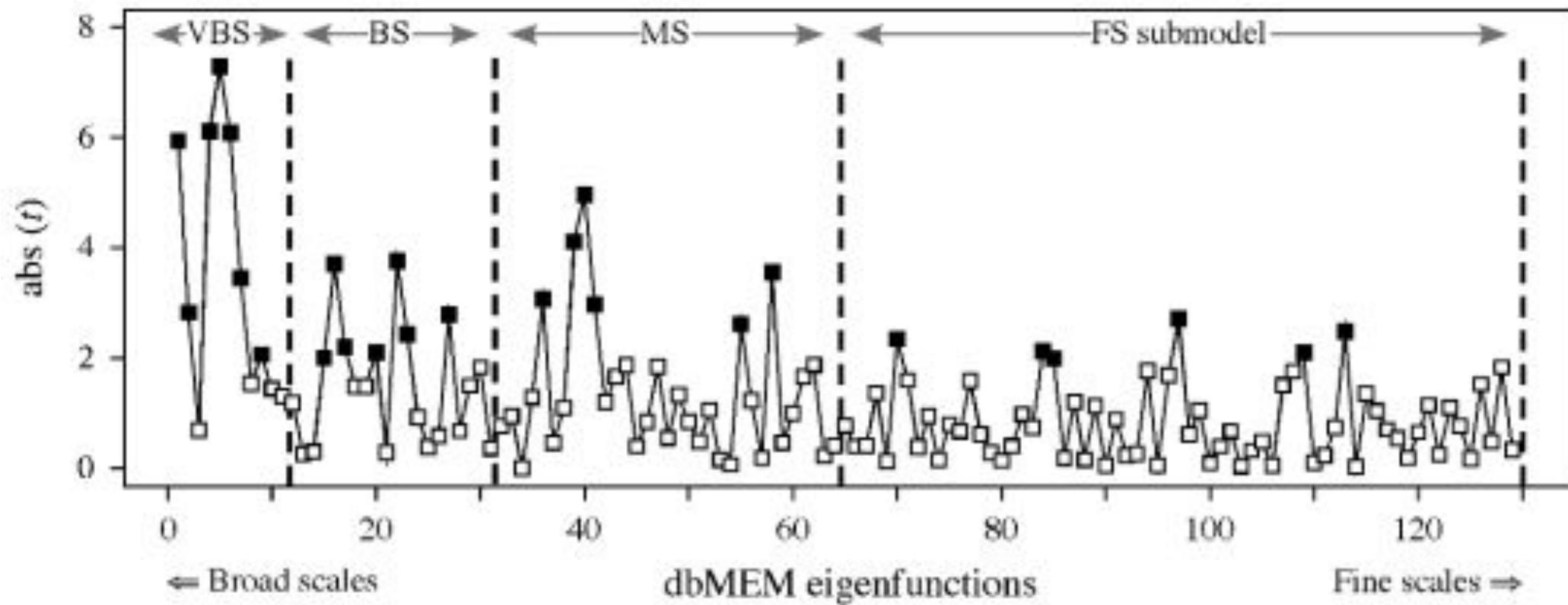


dbMEM.

- We can decompose spatial variation along a gradient (below, 1D) into a combination of independent functions (eigenvectors)
- Some will reflect positive correlations between distances (positive Moran's I), some are negative.
- Approaches based on so-called 'distance-based Moran's eigenvector MEM

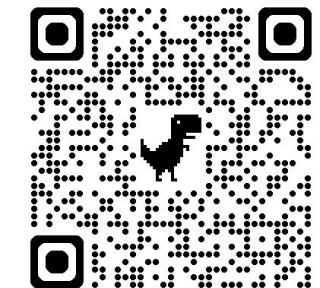
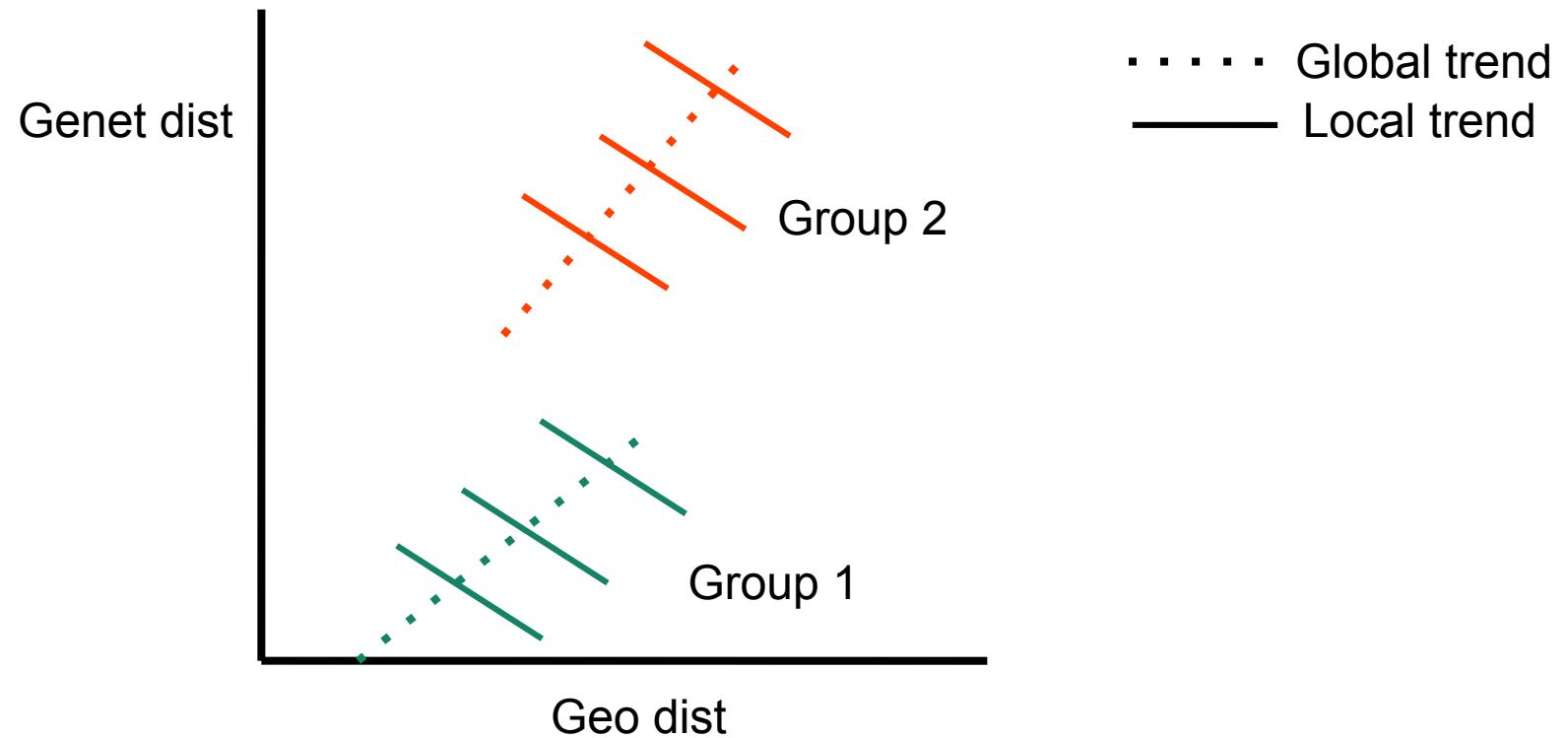


dbMEM.



Spatial PCA (sPCA).

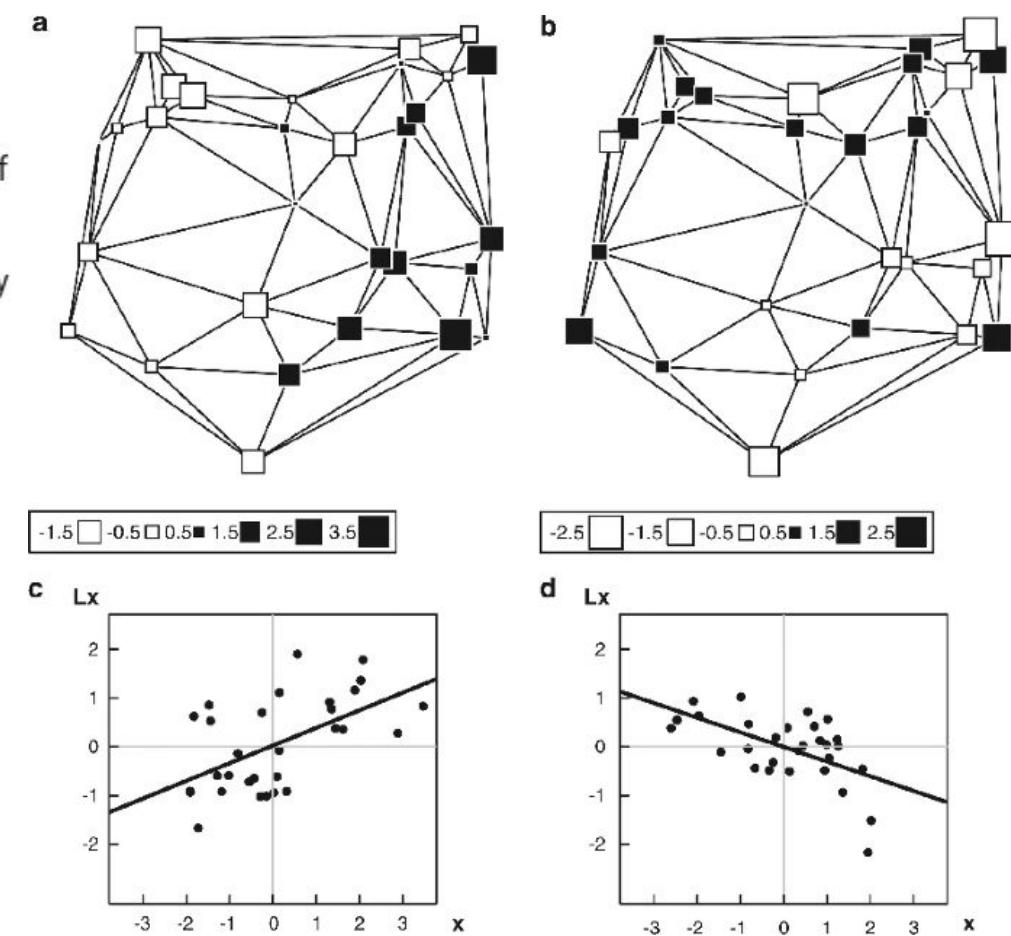
- A modification of PCA that takes spatial autocorrelation between individuals/populations into account.
- Outputs maps of scores to assess relevant spatial structure
- Two types of patterns can be investigated: global (population differentiation), and local (can reflect repulsion between neighbours)



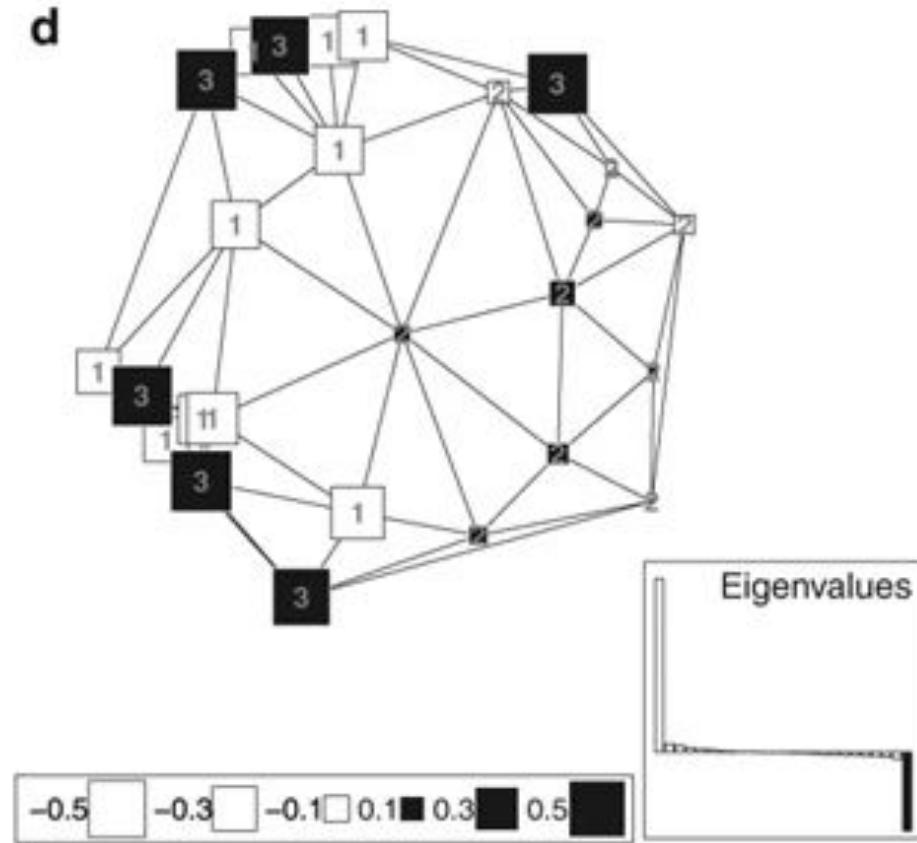
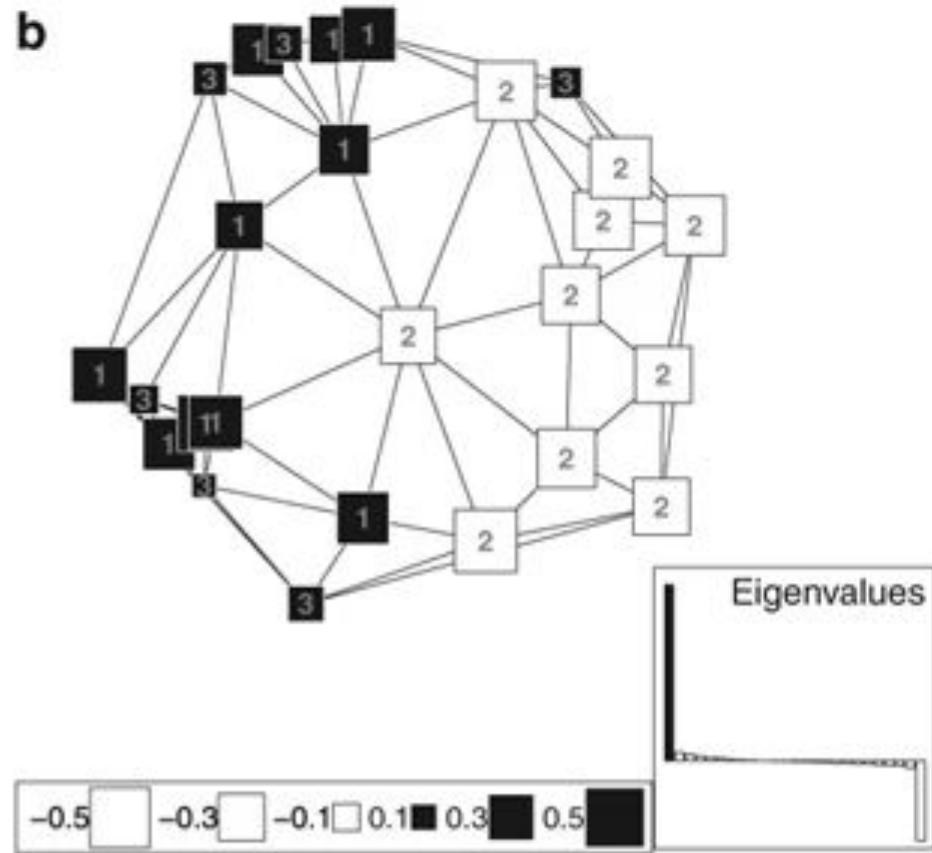
Spatial PCA (sPCA).

Illustration of global and local patterns of an allelic frequency for 20 fictitious populations overlying their sampling area. Each square represents the frequency of a population. Edges correspond to the connection network (Gabriel's graph). (a) Example of global structure, corresponding to $I(x) > I_0$. (b) Example of local structure, corresponding to $I(x) < I_0$. (c) Moran's scatterplot showing that in the global structure (a), the allelic frequency x of a population is positively correlated with the mean frequency of its neighbors, Lx . The line corresponds to the linear regression of Lx on x . (d) Conversely, the Moran's scatterplot associated with the local structure (b) shows that frequency x of a population is negatively correlated with the mean value of its neighbors, Lx .

- First a graph connecting neighbours is fitted.
- This graph can be adjusted but is usually chosen to maximize the number of neighbours.
- This is used to derive a spatial matrix of distances
- Two important aspects: **autocorrelation** (Moran's I) and **% of genetic variance explained**.



Spatial PCA (sPCA).

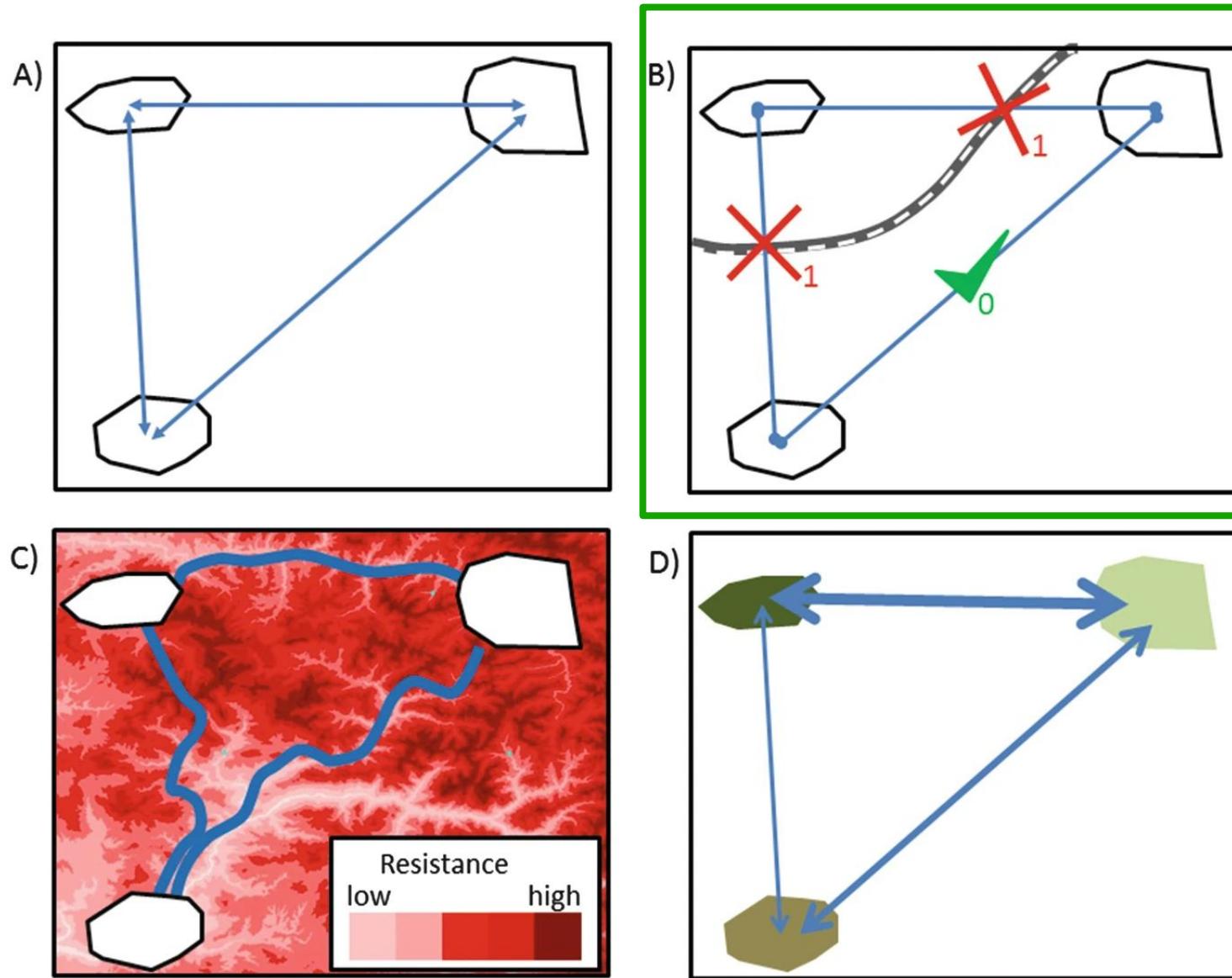


Limitations

- Mostly descriptive.
- Patterns are sometimes difficult to interpret.
- But more informative about spatial structure than Mantel.



Isolation-by-Barrier



Barriers and corridors to gene flow

EEMS: isolation by resistance to naively identify barriers

Uses matrix of genetic dissimilarities: fast, but loss of information

Expects a broad pattern of IBD

Does not reveal past dynamics

Migrati
on rate
↓

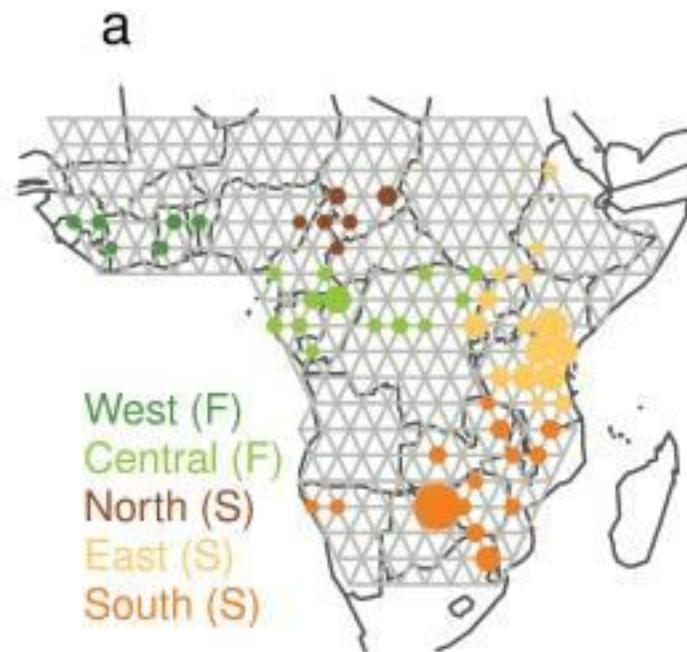
Forest



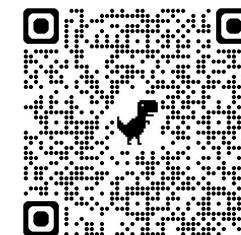
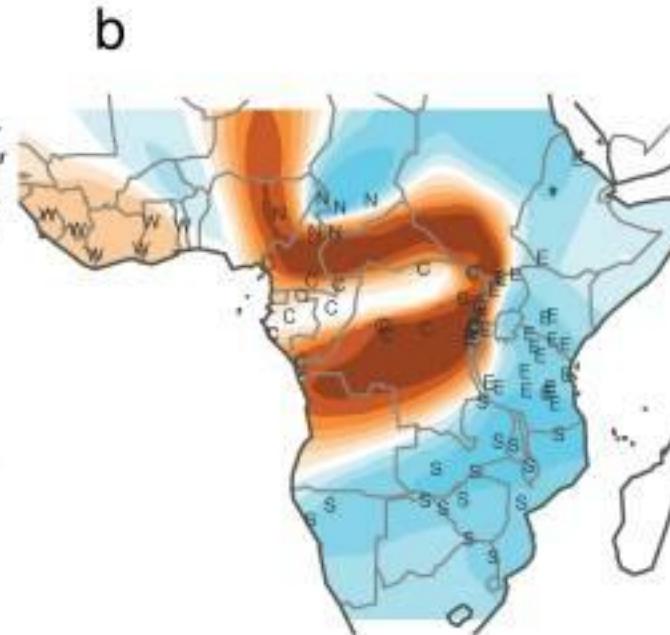
Savann
a



a



b



Petkova et al. 2016, Nat. Genet.

Barriers and corridors to gene flow

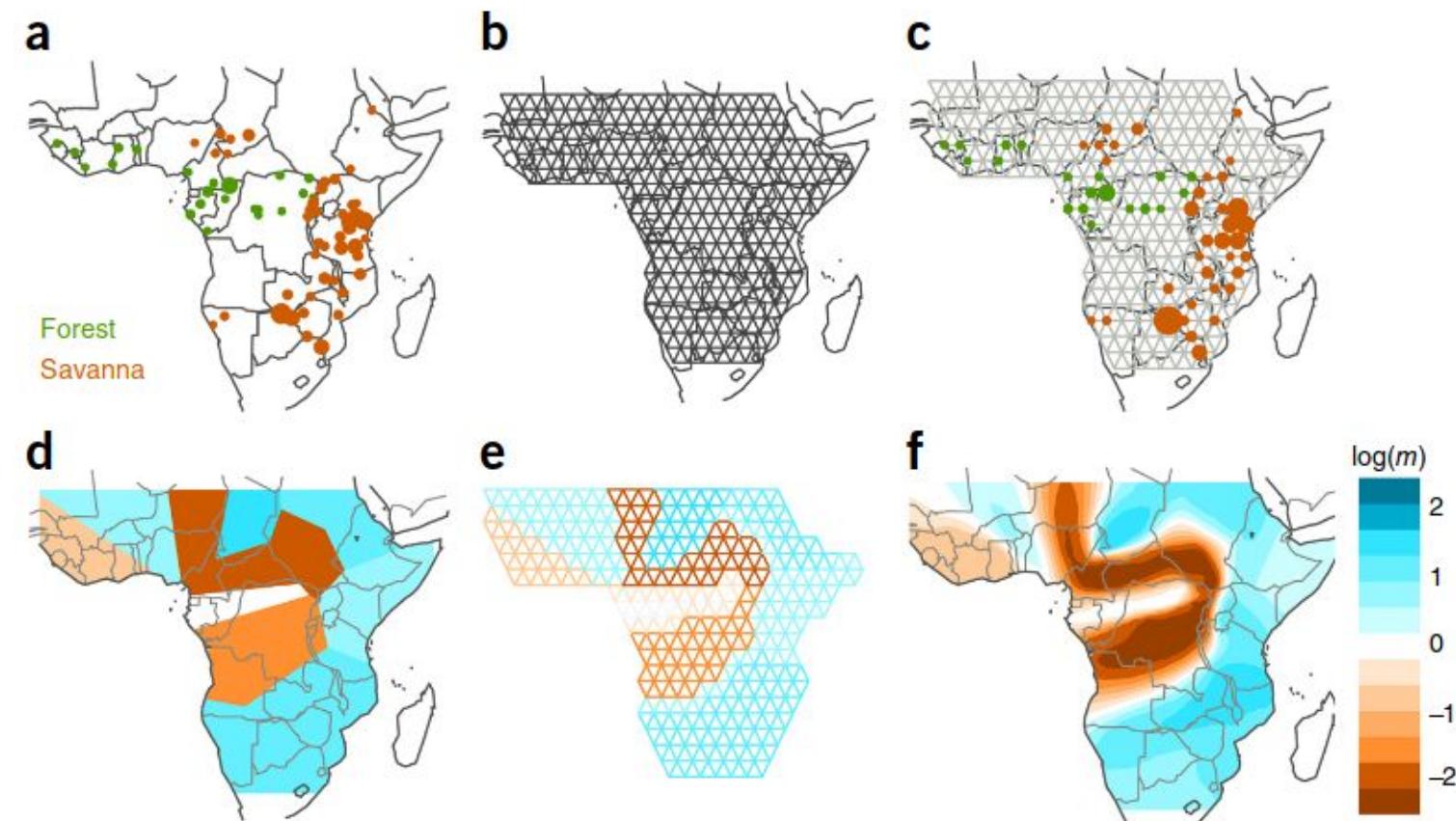
Figure 1 A schematic overview of EEMS (Estimated Effective Migration Surfaces), using African elephant data for illustration.

(a–c) Setting up the population grid.

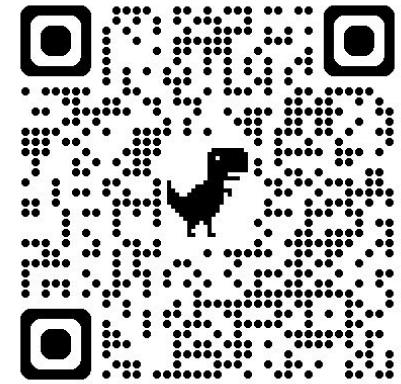
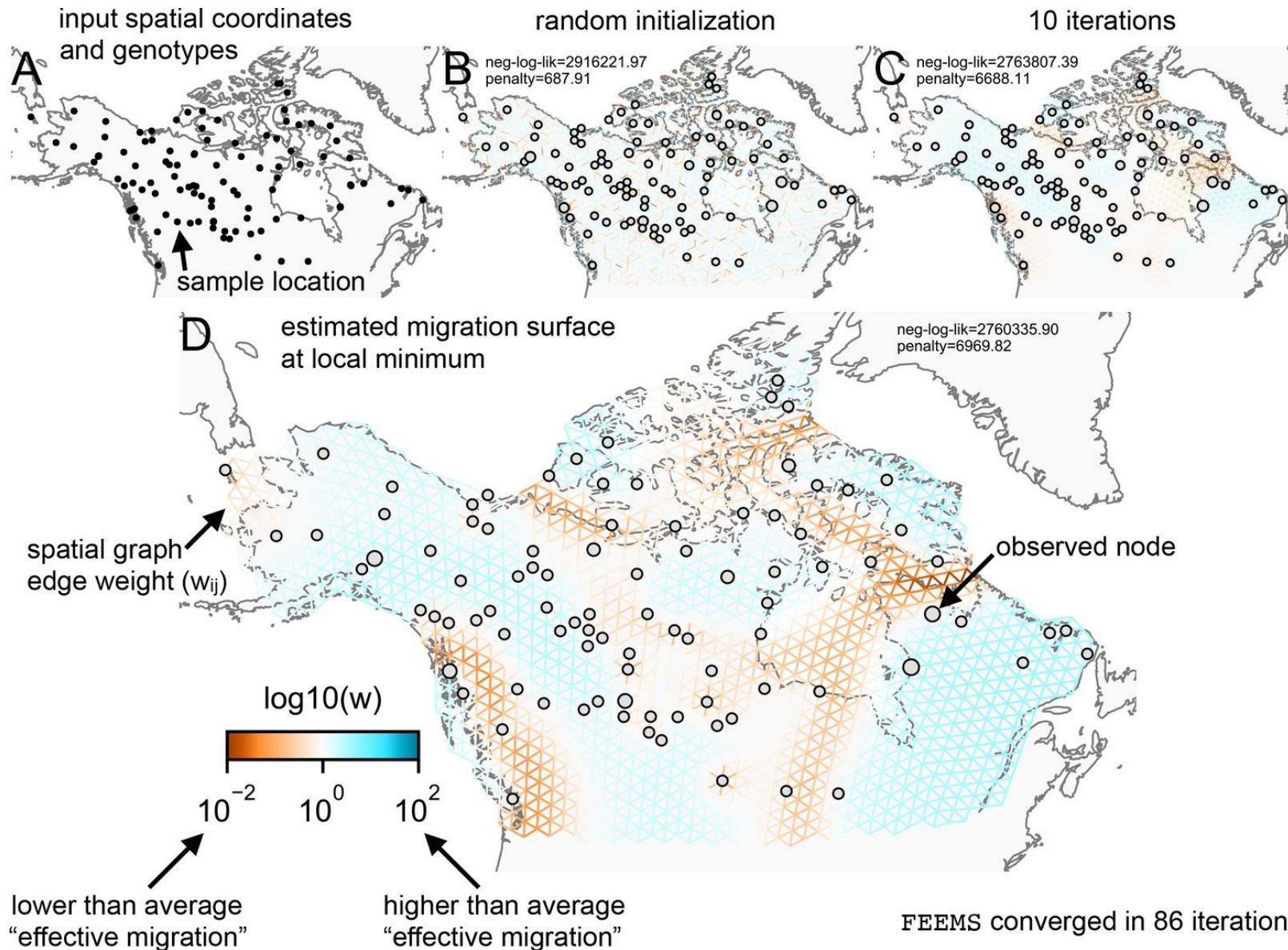
(a) Samples are collected at known locations across a two-dimensional habitat; green and orange represent two species of African elephant, forest and savanna, respectively.

(b) A dense triangular grid is chosen to span the habitat. (c) Each sample is assigned to the closest deme on the grid. (d–f) EEMS analysis. (d) Migration rates vary according to a Voronoi tessellation that partitions the habitat into ‘cells’ with constant migration rate; colors represent relative rates of migration, ranging from low (orange) to high (blue).

(e) Each edge has the same migration rate as the cell into which it falls. Cell locations and migration rates are adjusted, using Bayesian inference, so that expected genetic dissimilarities under the EEMS model match observed genetic dissimilarities. (f) The EEMS is a color contour plot produced by averaging draws from the posterior distribution of the migration rates, interpolating between grid points. Here and in all other figures, $\log(m)$ denotes the effective migration rate on a \log_{10} scale, relative to the overall migration rate across the habitat. (Thus, $\log(m) = 1$ corresponds to effective migration that is tenfold faster than the average.) The main feature of the EEMS for the African elephant is a barrier of low effective migration that separates the habitats of the two species: forest elephants to the west and savanna elephants to the north, south and east.

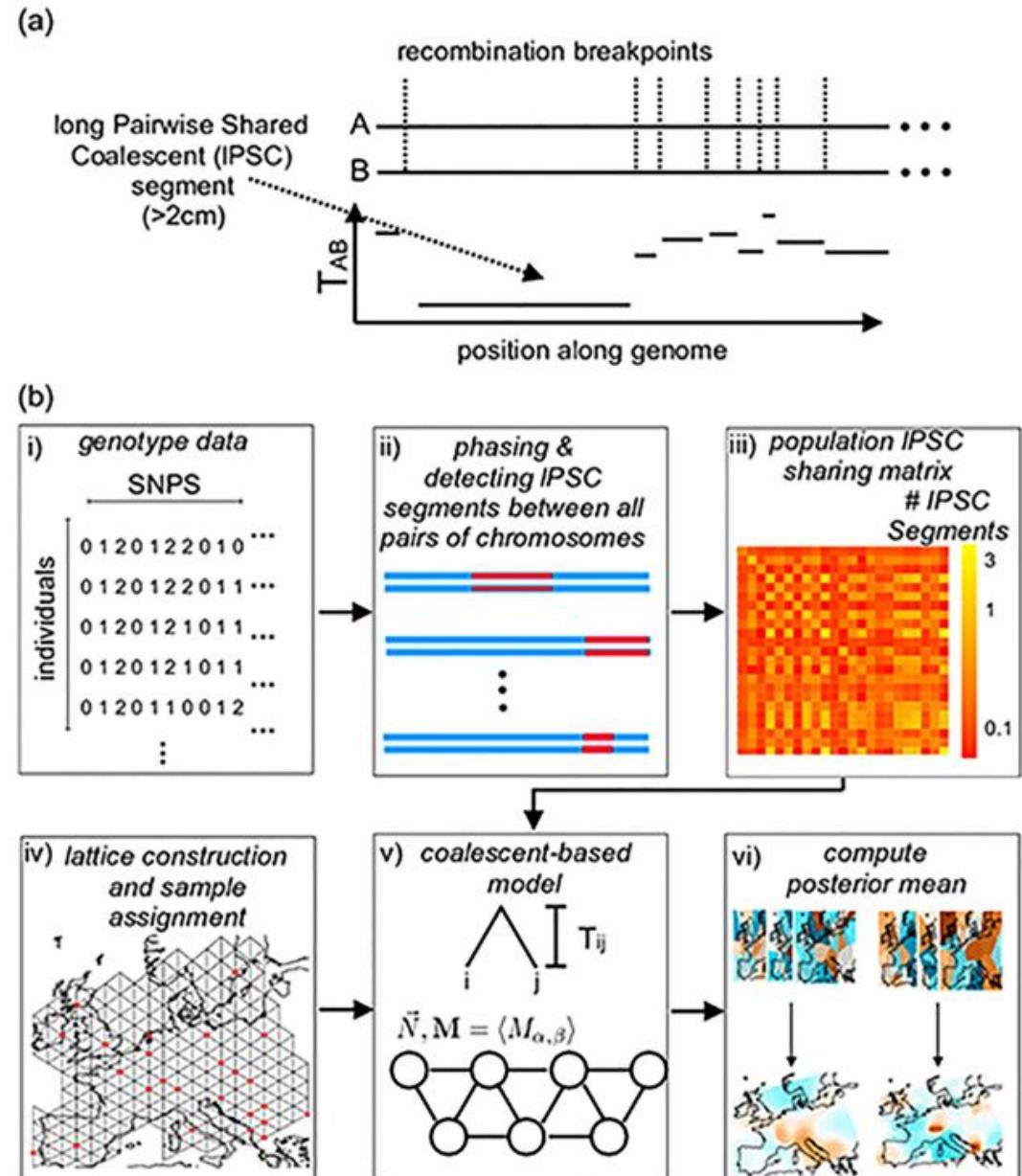


Barriers and corridors to gene flow



Dynamics in time and space: MAPS

- Requires information about tracks identical by descent (IBD, but not isolation-by-distance...)
- Not suitable for small sample sizes/species with little prior information (e.g. better to have recombination rates).



Dynamics in time and space: MAPS

- Requires information about tracks identical by descent (IBD, but not isolation-by-distance...)
- Not suitable for small sample sizes/species with little prior information (e.g. recombination rates).
- Very approximate times.

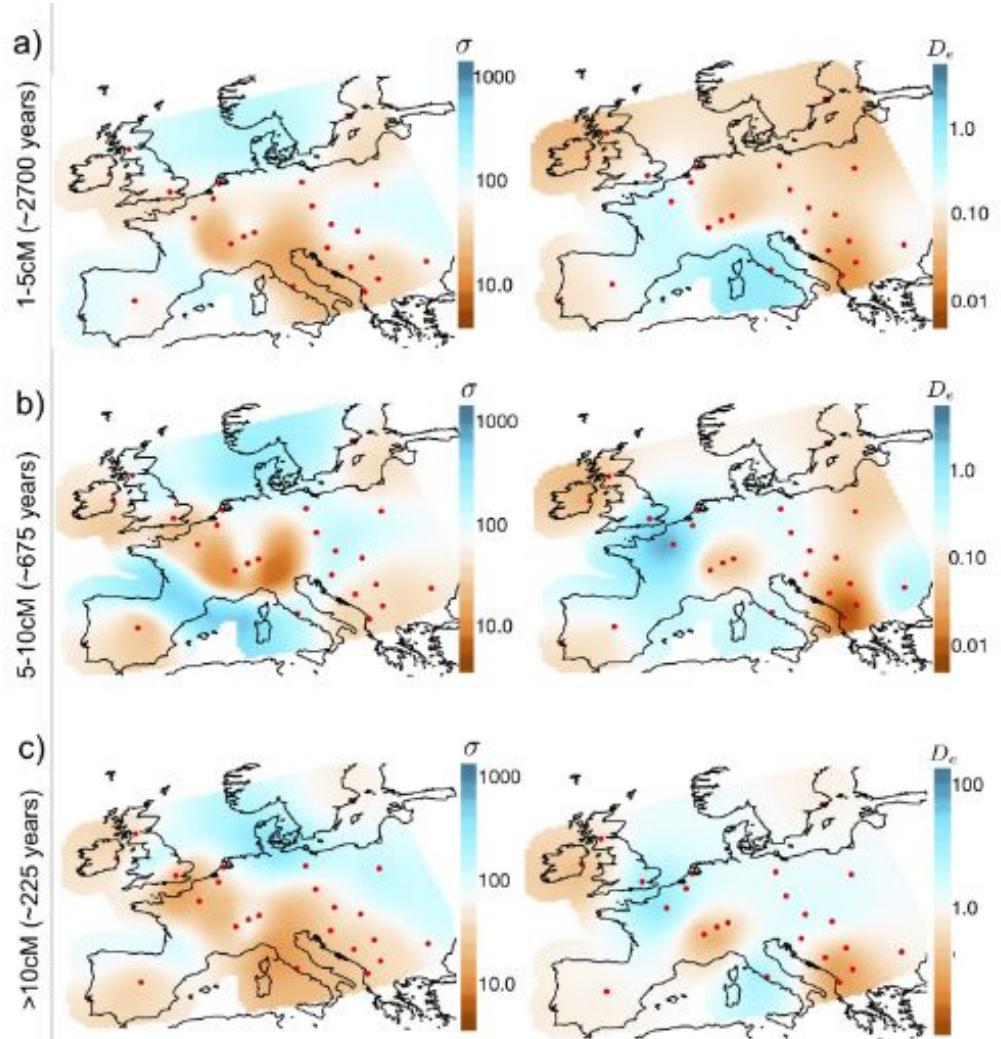
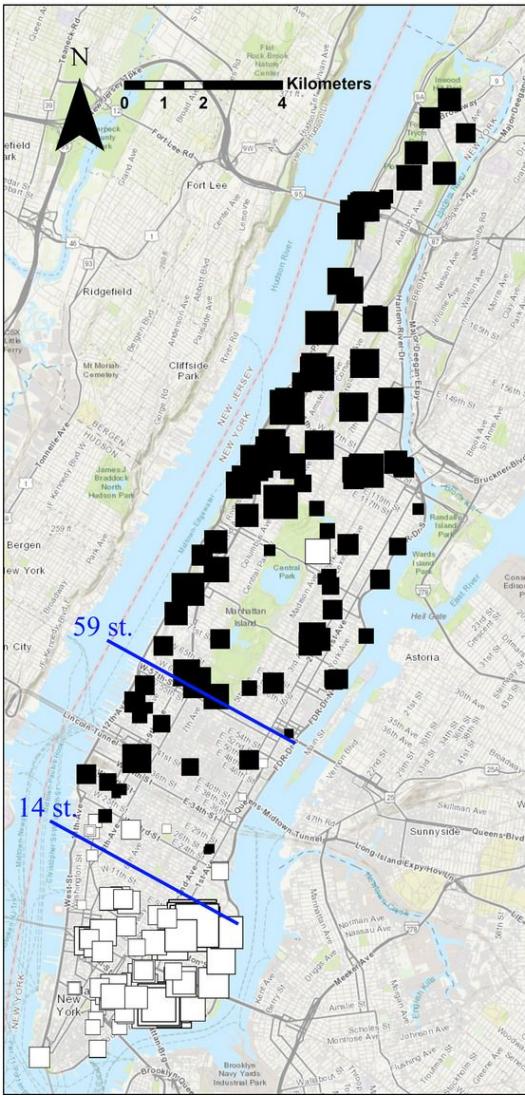


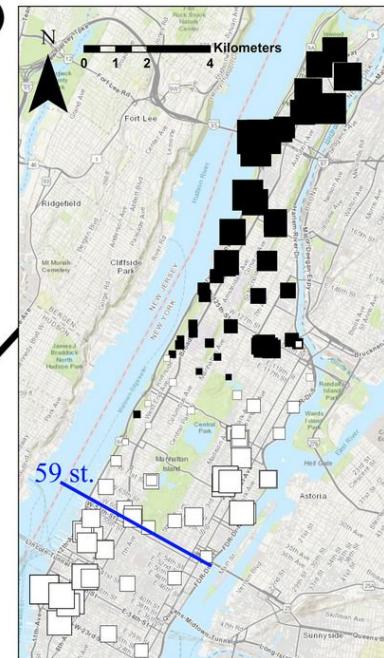
Fig 4. Inferred dispersal surfaces and population density surfaces over time for Europe. We apply MAPS to a European subset of POPRES [25] with 2,234 individuals and plot the inferred dispersal $\sigma(\bar{x})$ and population density $D_e(\bar{x})$ surfaces for PSC length bins (a) $>1\text{cM}$ (b) $5\text{-}10\text{cM}$ and (c) $>10\text{cM}$. We transform estimates of \bar{N} and M to estimates of $\sigma(\bar{x})$ and $D_e(\bar{x})$ by scaling the migration rates and population sizes by the grid step-size and area (see Eqs (17) and (18)). Generally, we observe the patterns of dispersal to be relatively constant over time periods, however, we see a sharp increase in population density in the most recent time scale ($>10\text{cM}$). Note the wider plotting limits in inferred densities in the most recent time scale.

An application: rats of New York.

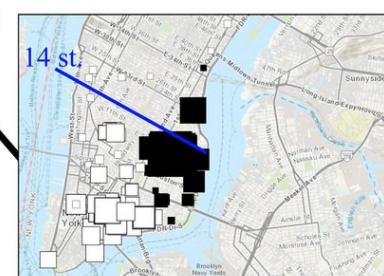
(a)



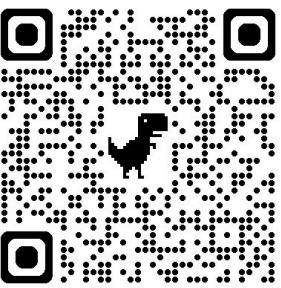
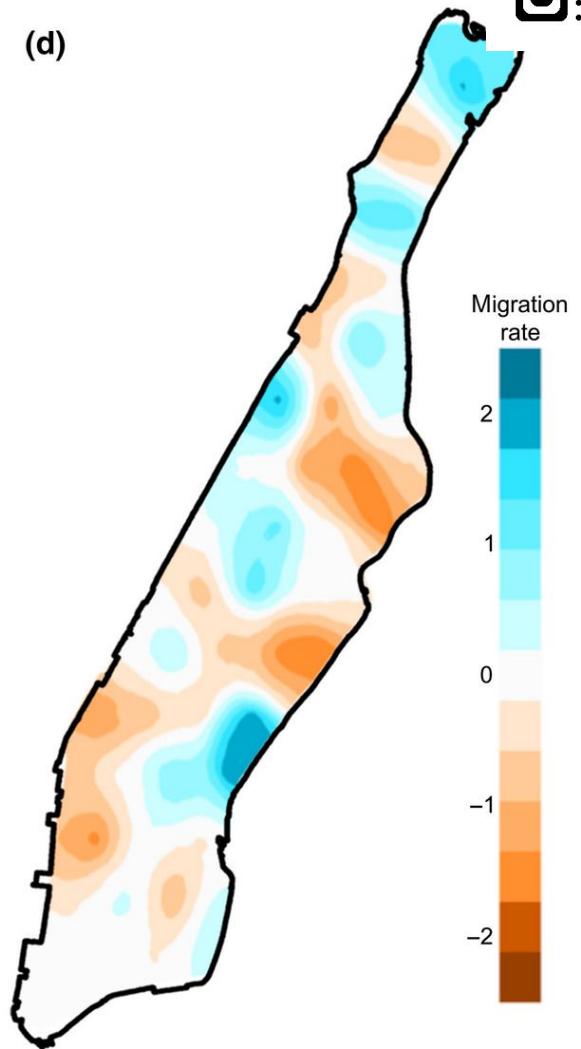
(b)



(c)

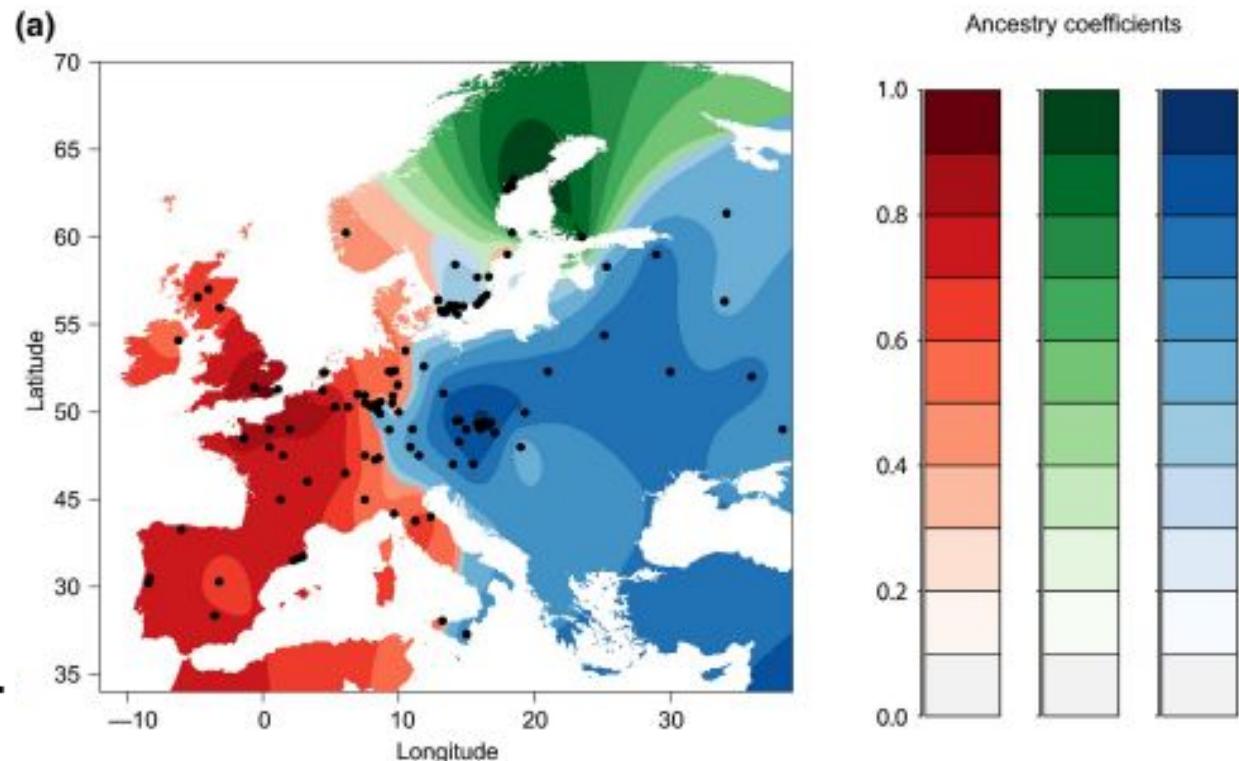


(d)



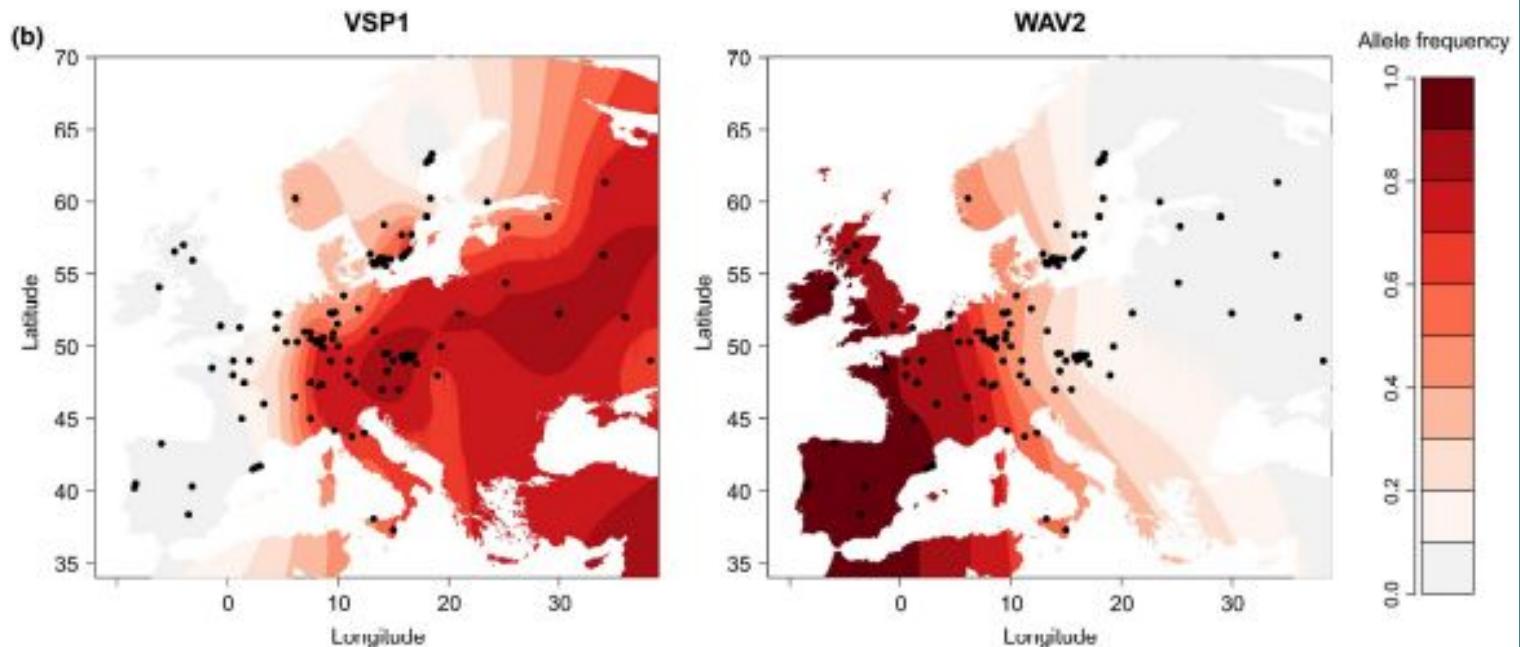
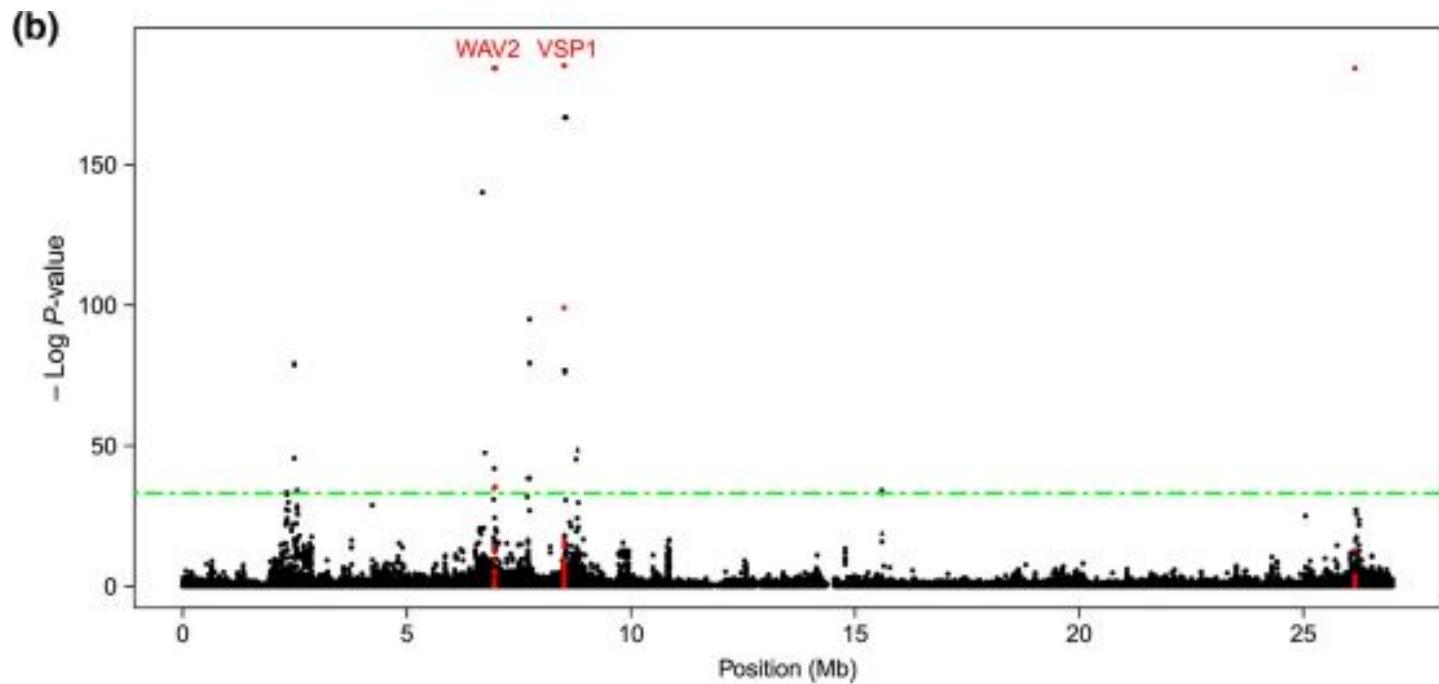
TESS3

- Spatialized clustering analysis
- ADMIXTURE in space, but model-free
- Spatially close individuals have a higher prior probability for being related.



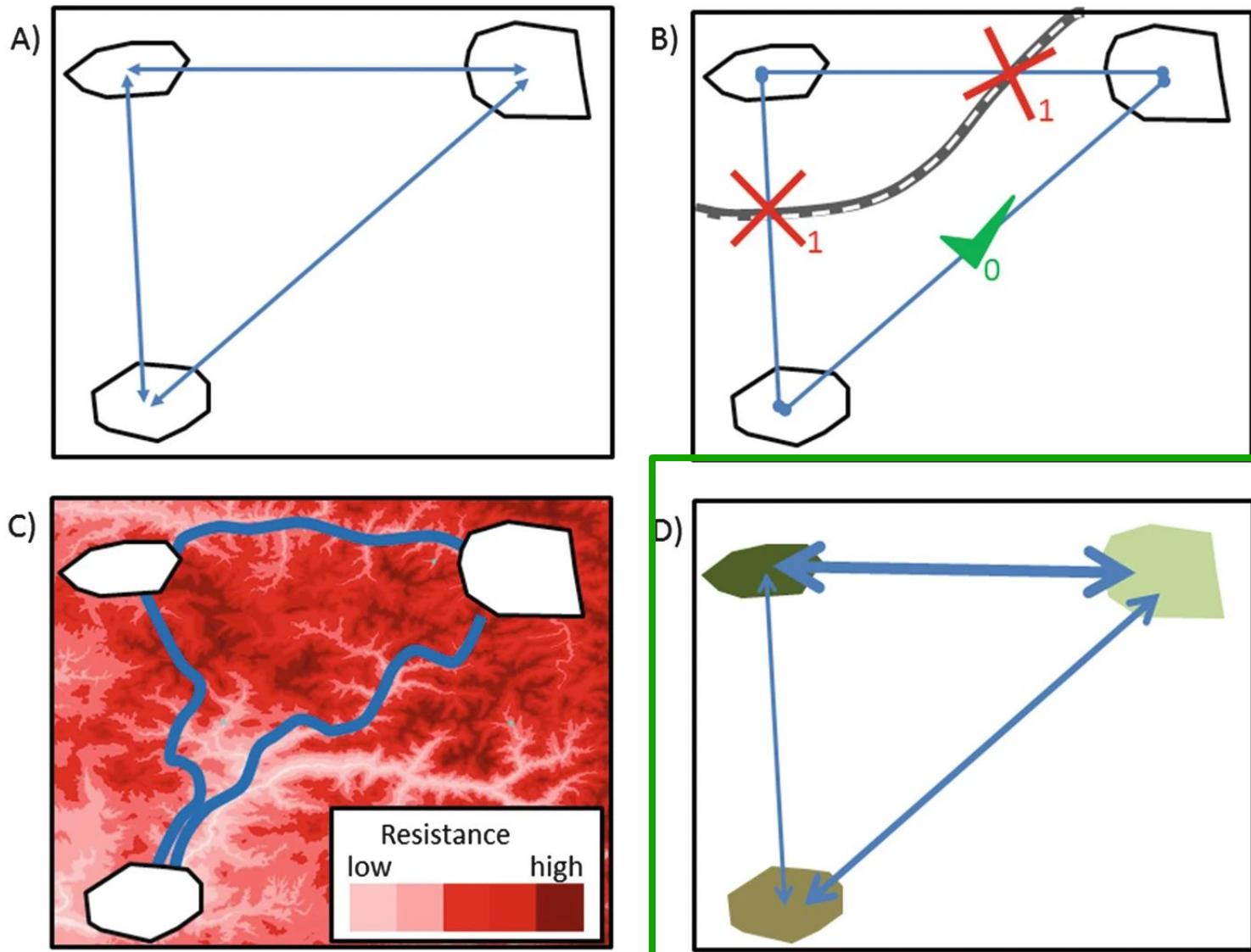
TESS3

- Can scan for loci deviating from genome-wide pattern
- Can be used to represent selected alleles in space



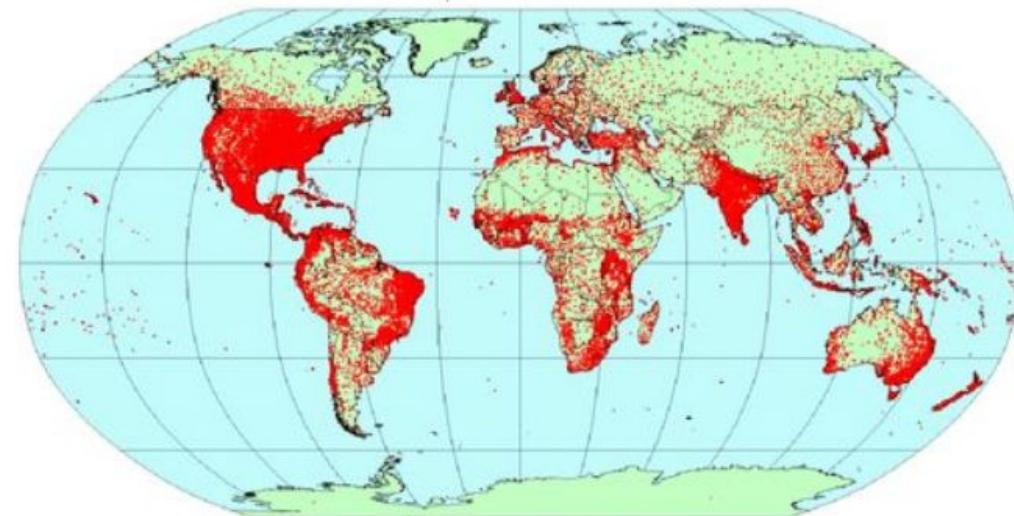


Isolation-by-Environment

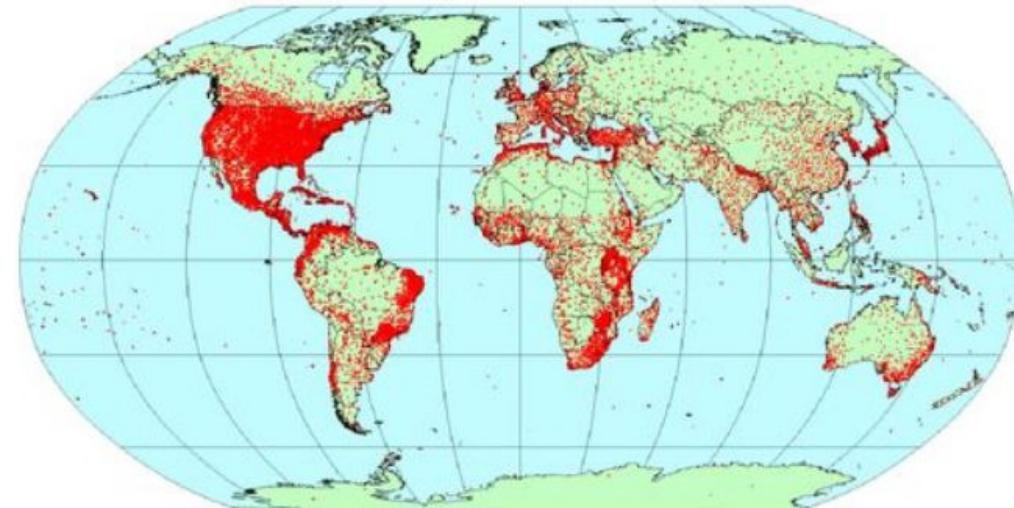


Including the environment: where can I find data?

- Worldclim
- Obtained from weather stations across the world
- **Interpolated** values obtained across a 1km grid.
- <https://www.worldclim.org/data/worldclim21.html>
- Abiotic variables (precipitations, temperatures).
- Past (LGM) and future conditions also available.



Locations of climate stations with precipitation data.



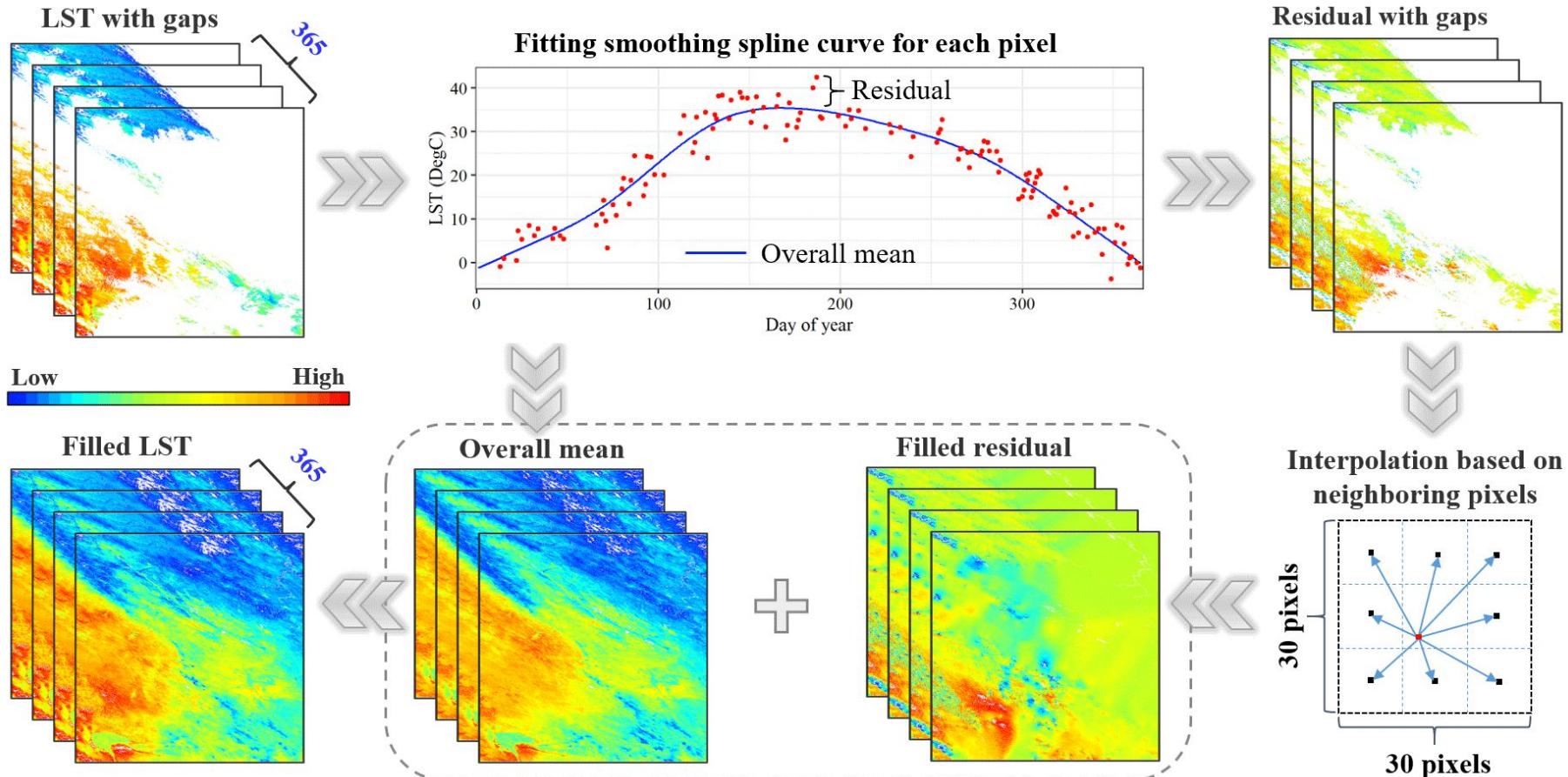
Locations of climate stations with mean temperature data.

The environment: where can I find data?

- Chelsa
- Obtained from weather stations across the world
- **Interpolated** values obtained across a 1km grid.
- May be more precise than Worldclim at small resolutions (check both over your study area!)
- Example of comparison:
<https://chelsa-climate.org/comparison-to-other-high-resolution-climate-products/>
- Abiotic variables (precipitations, temperatures).
- Past (LGM) and future conditions also available.

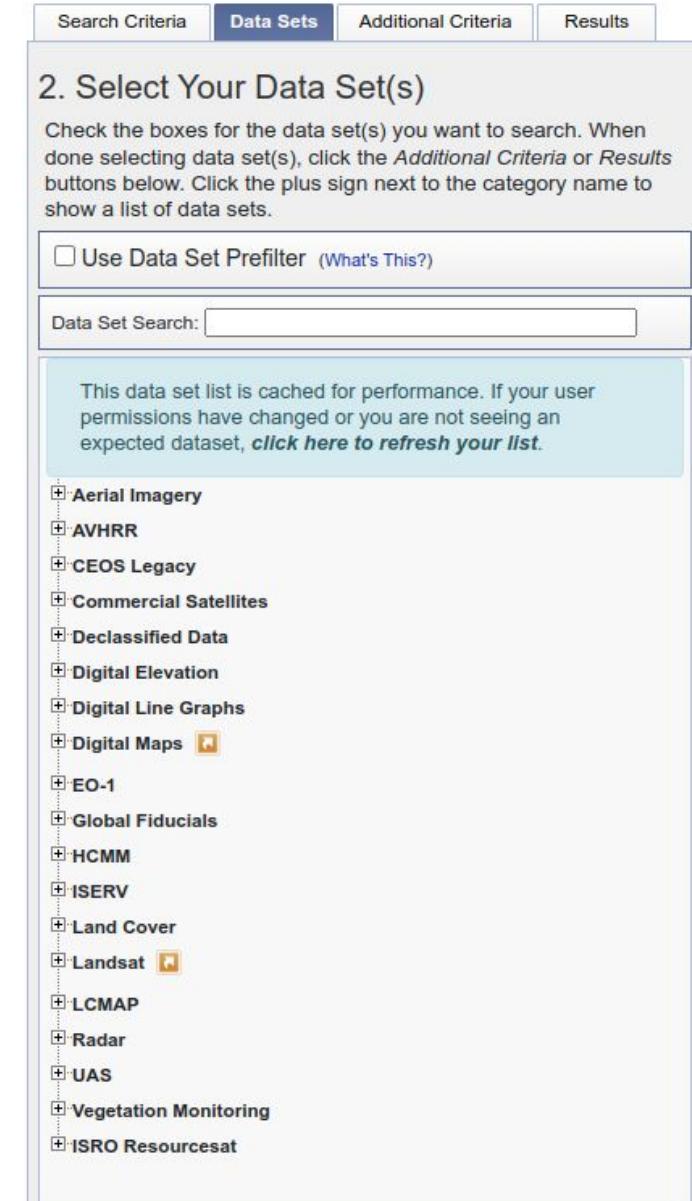
The environment: where can I find data?

- Keep in mind that these data are obtained through **interpolation of residuals**: the accuracy of the values therefore depend on the underlying model and density of stations in the study area



The environment: where can I find data?

- USGS
<https://earthexplorer.usgs.gov/>
- Satellite imagery: no interpolation here
- Resolution may remain limited at short spatial scales, for regions with permanent cloud covers, or mountain slopes.



The screenshot shows the 'Data Sets' tab selected in the top navigation bar. Below it, a section titled '2. Select Your Data Set(s)' contains instructions and two input fields: 'Use Data Set Prefilter' (checkbox) and 'Data Set Search' (text input). A note below the search field says: 'This data set list is cached for performance. If your user permissions have changed or you are not seeing an expected dataset, [click here to refresh your list](#)'. A large list of data sets is displayed as a tree structure with plus signs:

- + Aerial Imagery
- + AVHRR
- + CEOS Legacy
- + Commercial Satellites
- + Declassified Data
- + Digital Elevation
- + Digital Line Graphs
- + Digital Maps
- + EO-1
- + Global Fiducials
- + HCMM
- + ISERV
- + Land Cover
- + Landsat
- + LCMAP
- + Radar
- + UAS
- + Vegetation Monitoring
- + ISRO Resourcesat

The environment: where can I find data?

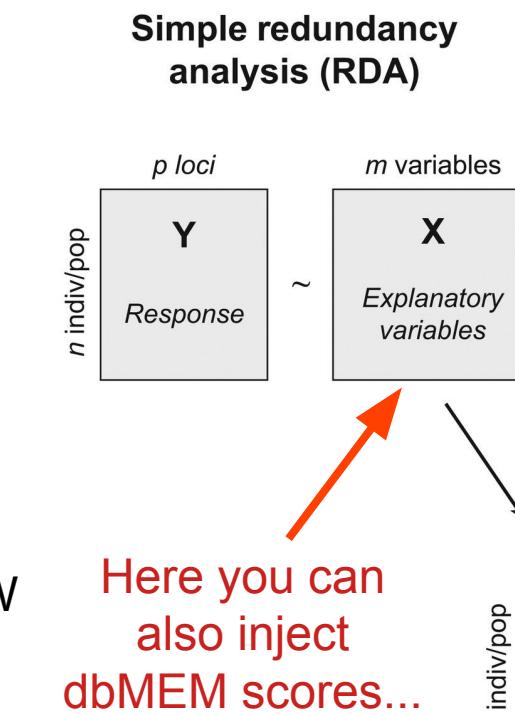
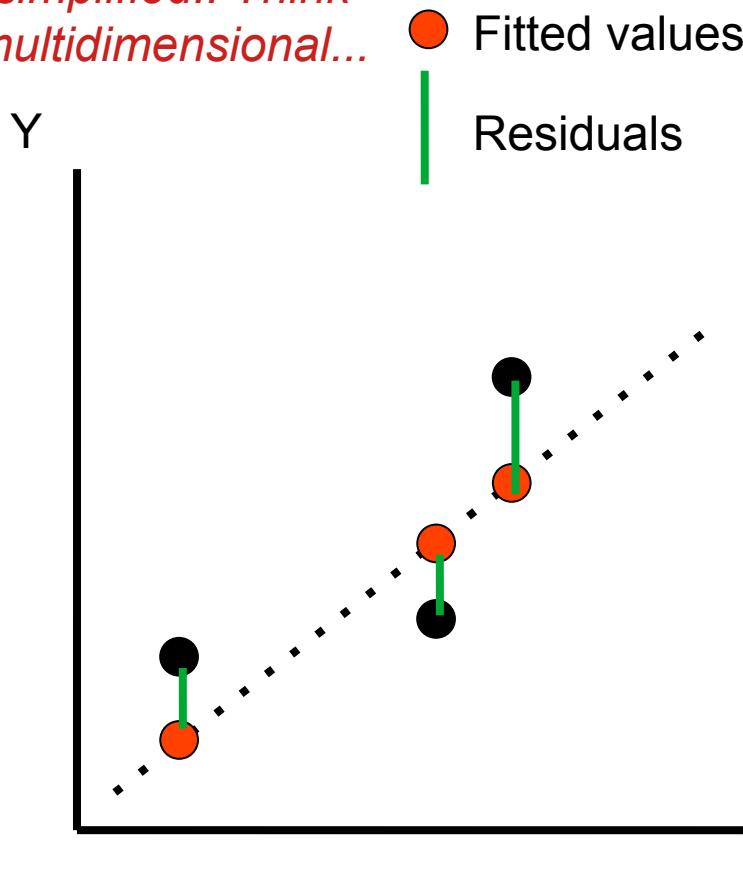
- USDS
- Satellite imagery: no interpolation here
- Resolution may remain limited at short spatial scales.
- National platforms may also provide resources.

Models for the earth	SRTM3	ASTER GDEM
Acquisition date	11 days in 2000	2000-2011
Spatial resolution	3 arcsec ($\approx 90m$)	1 arcsec ($\approx 30m$)
DEM accuracy (stdev.)	$\pm 16m$	$\pm 12.6m$ (v.2)
Coverage	60°N – 56°S	83°N – 83°S
Aquisition method	Radar interferometry	Extraction of corresponding points between images by pattern matching
Comment	Topographically steep areas causing radar shadow	Available for steep mountainous regions Missing data for regions under constant cloud cover

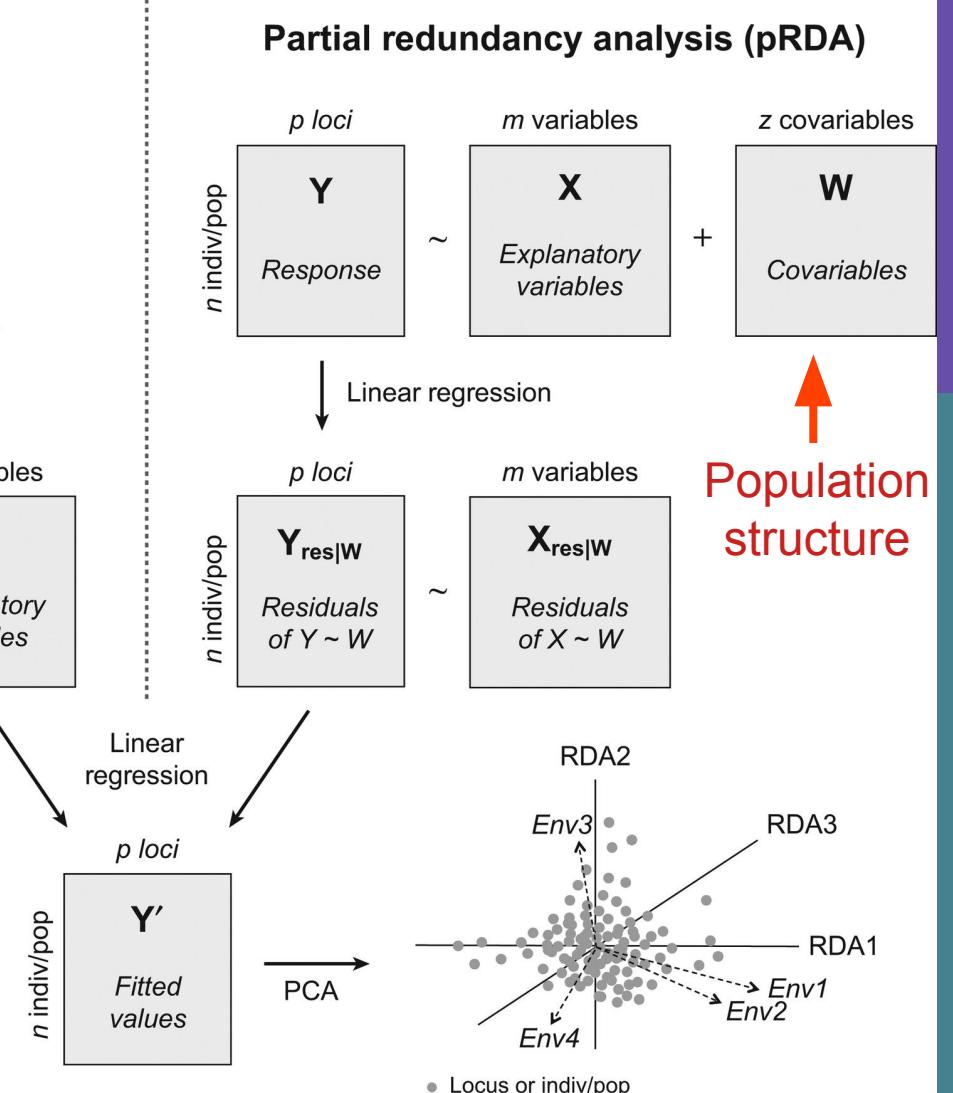


Redundancy analyses

This is outrageously simplified!! Think multidimensional...



2 steps: regression and ordination (constrained PCA)



Redundancy analyses

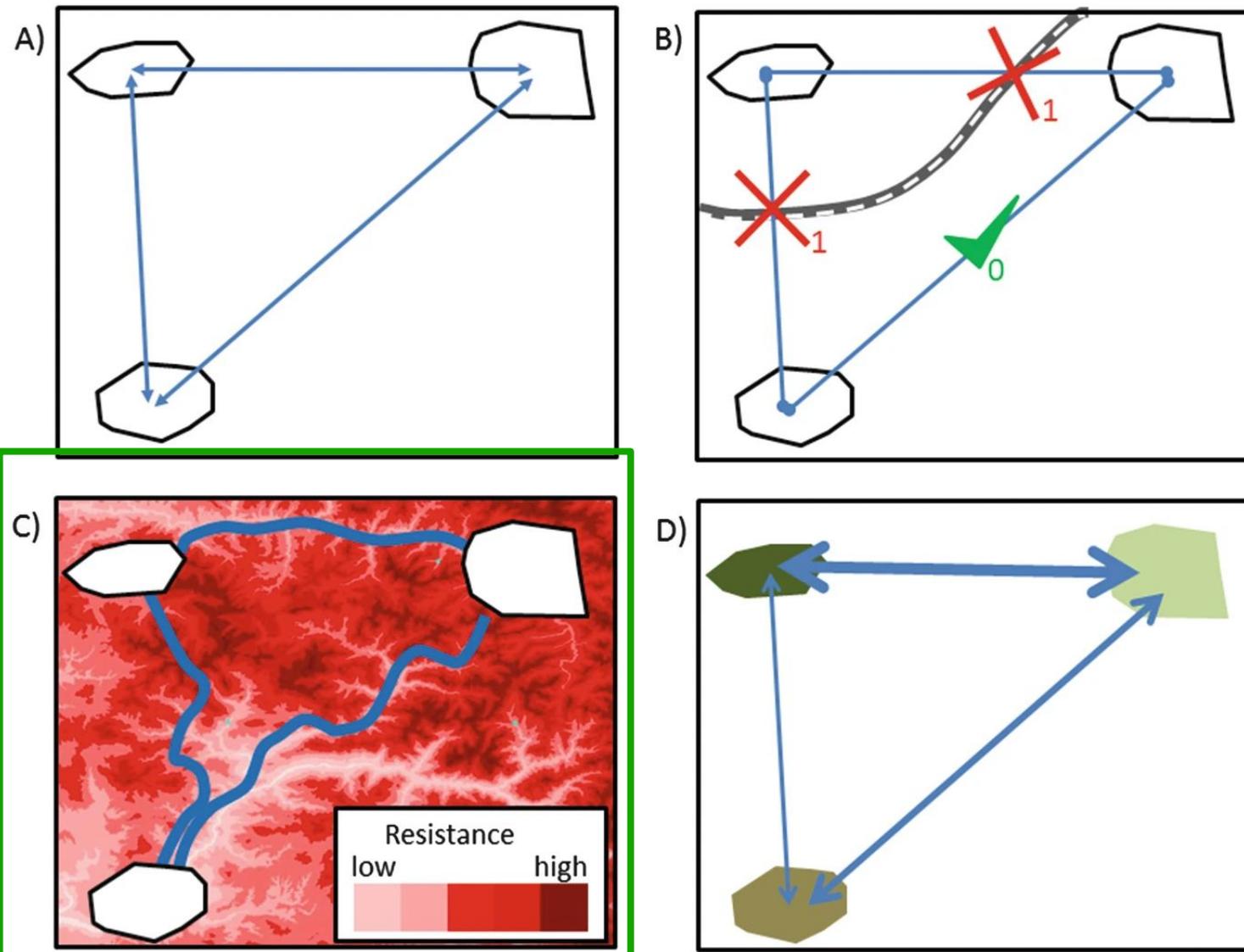
- Can be used to decide which environmental variables matter:
- - First fit an empty model $Y \sim \text{intercept}$.
 - Sequentially complexify the model until no significant increase of % of variance explained.
- Can also be used to identify loci that covary the most with the environment (SNP loadings).

Redundancy analyses

- Optimizes variance explained.
- What explains the most variance is not necessarily causal.
- Not popgen model-based, so deviations from population genetics expectations are not going to show.
- Default RDA expects a linear relationship between X and Y (but extensions can deviate from that).
- As usual: bad data enter, bad results come out.



To finish, Isolation-by-resistance

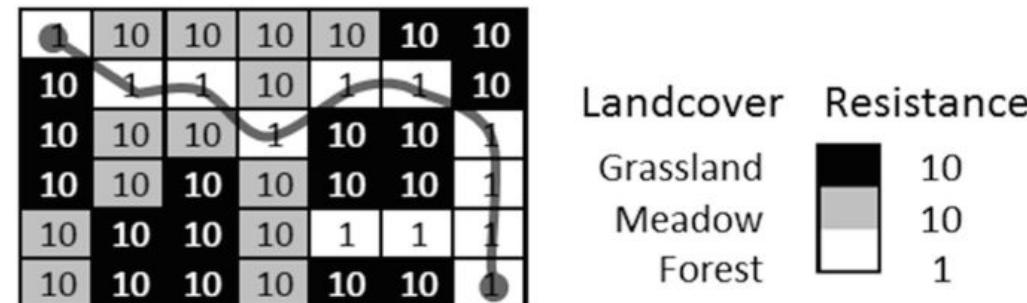
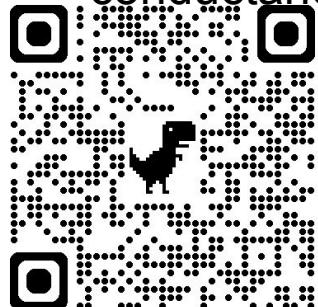


More on isolation by resistance

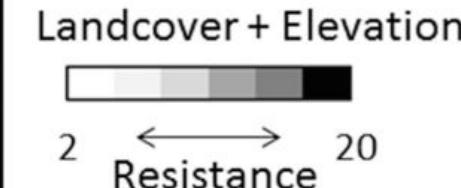
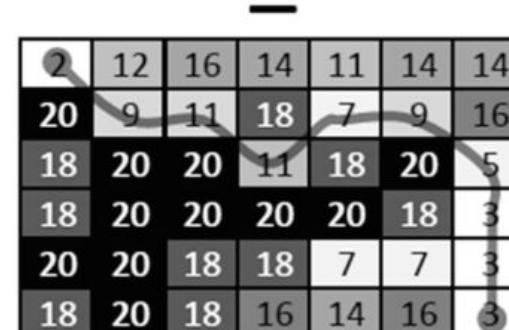
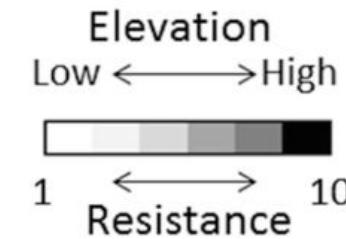
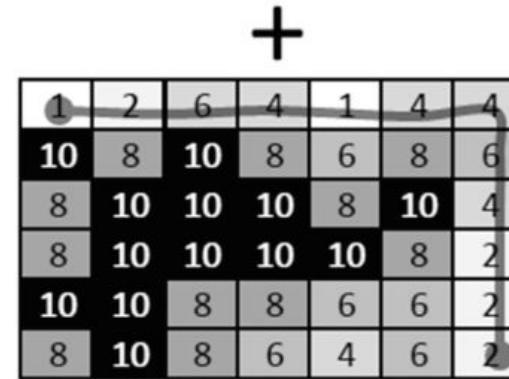
Create a combination of maps where each cell is given a resistance score depending on the feature it harbours

Calculate the path of least resistance (based on circuit theory)

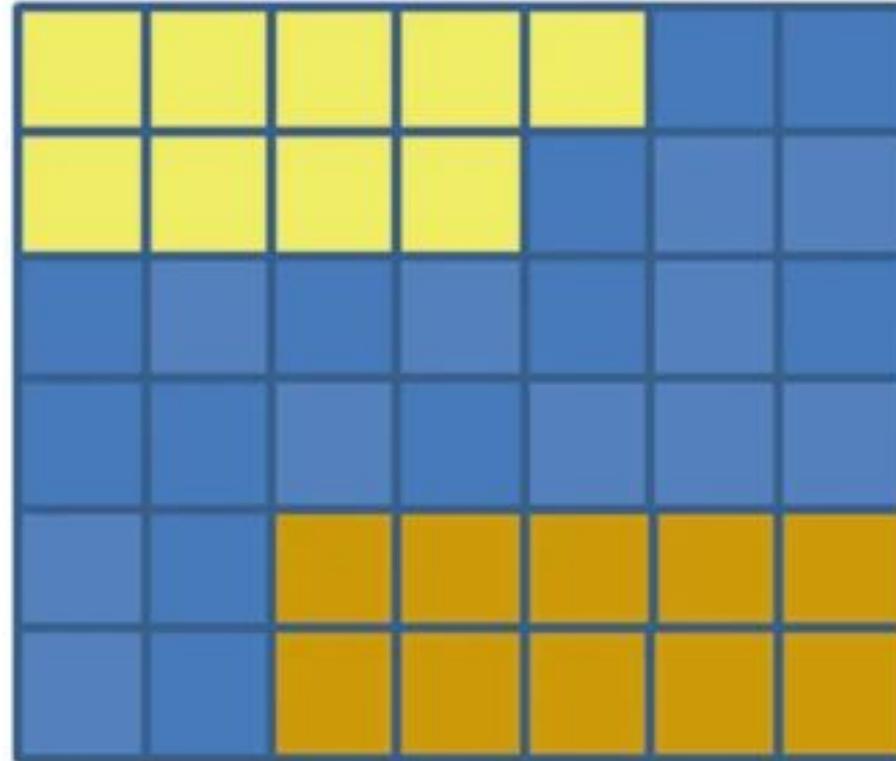
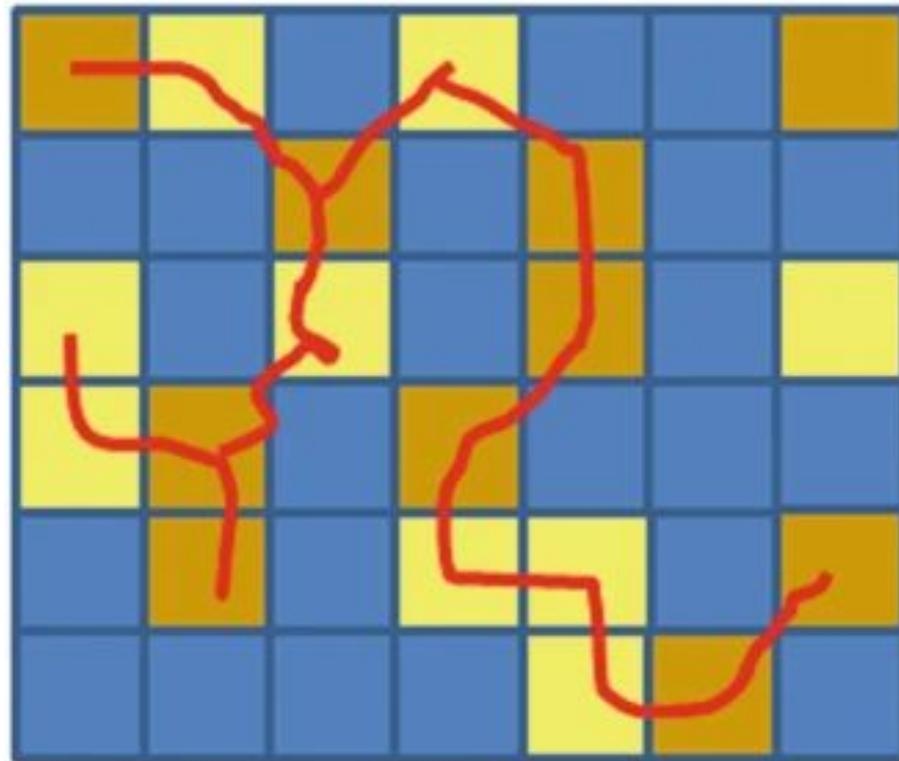
Test for correlation between genetic distance and distance of least resistance/higher conductance



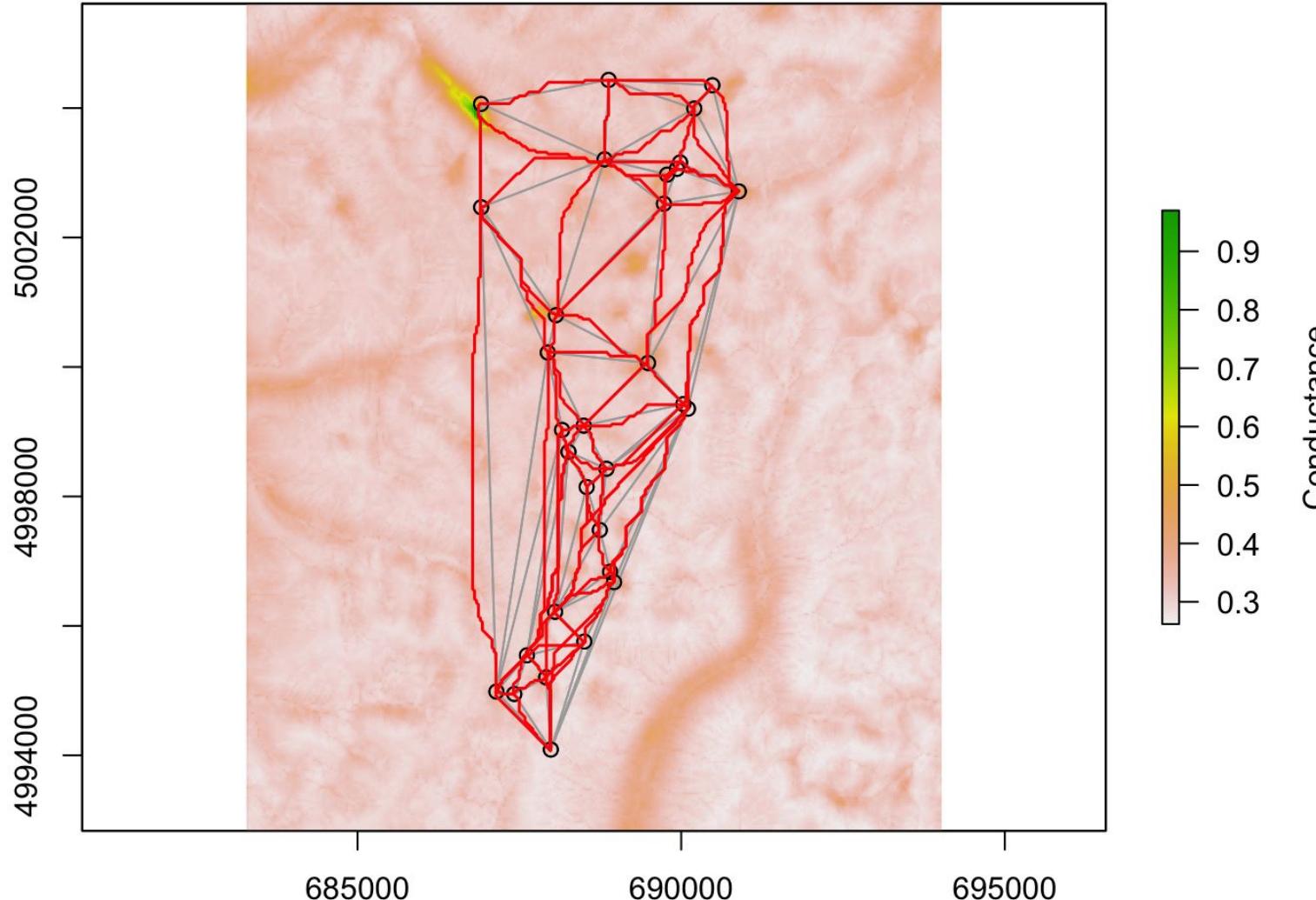
Landcover	Resistance
Grassland	10
Meadow	10
Forest	1



Functional connectivity ≠ landscape fragmentation.



Euclidean distance and shortest paths



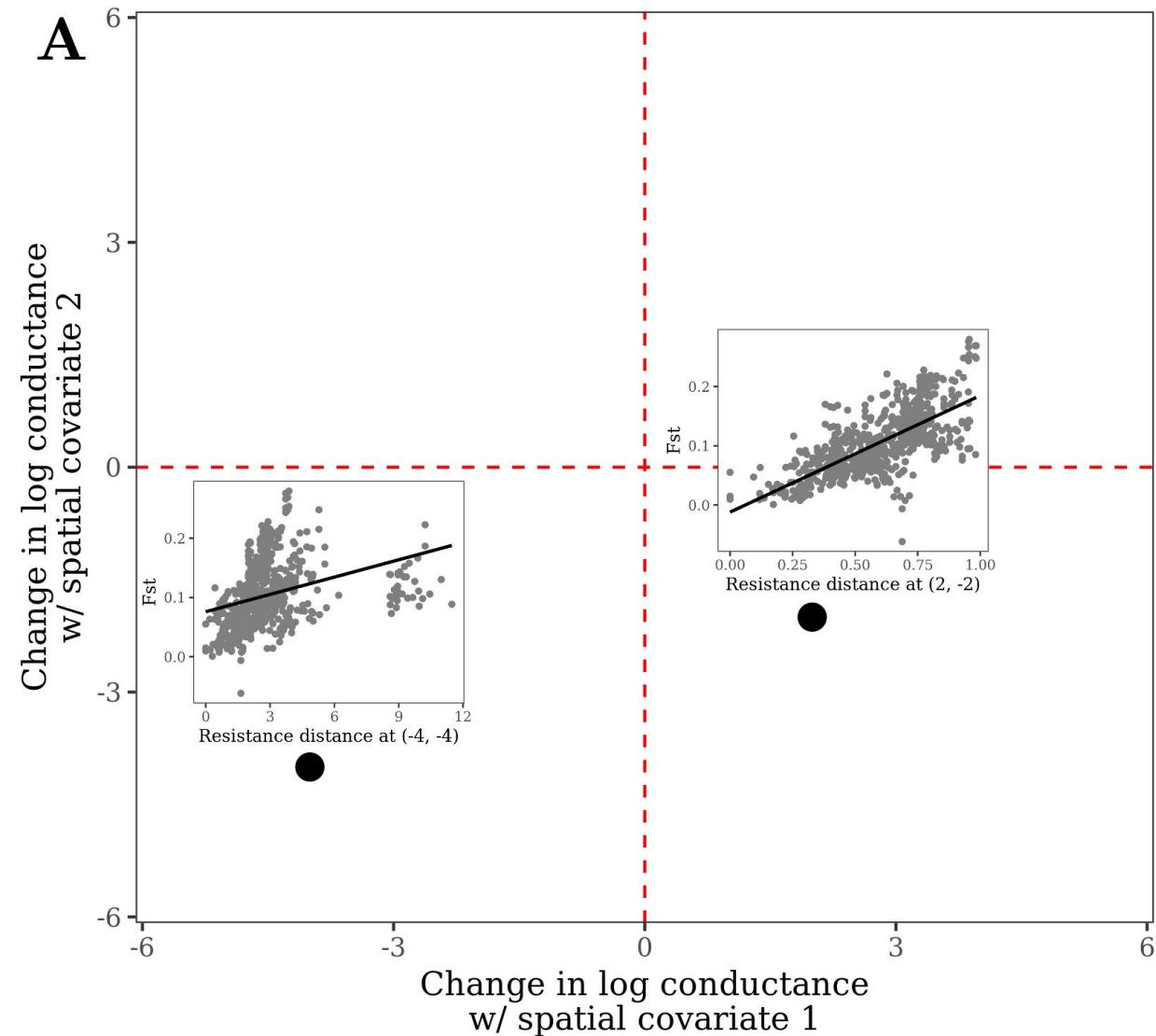
Testing isolation-by-resistance: how can we calculate resistance?

Step 1: vary the values of coefficients attributed to environmental covariates

Step 2: explore combinations of coefficients

Step 3: Choose the combination that maximizes the fit between resistance and genetic distance.

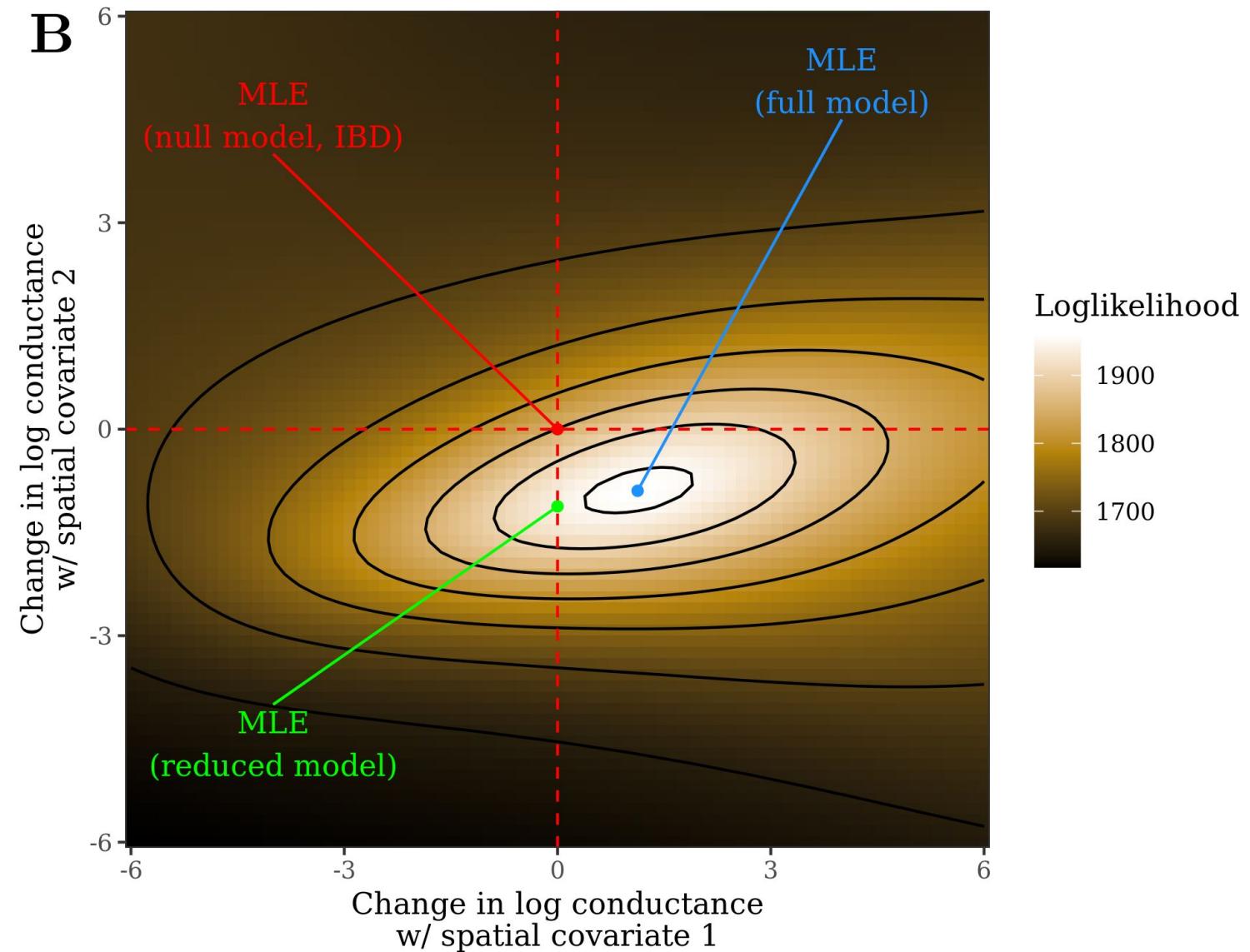
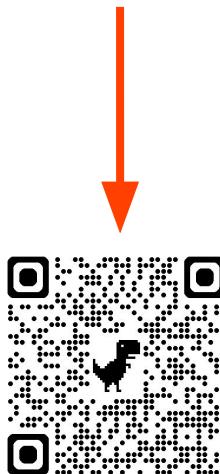
Excellent tutorial here:



Testing isolation-by-resistance

Step 4: Compare the likelihoods of different models

This last step is mathematically complex. The choice of optimizer in particular is still a matter for discussions.

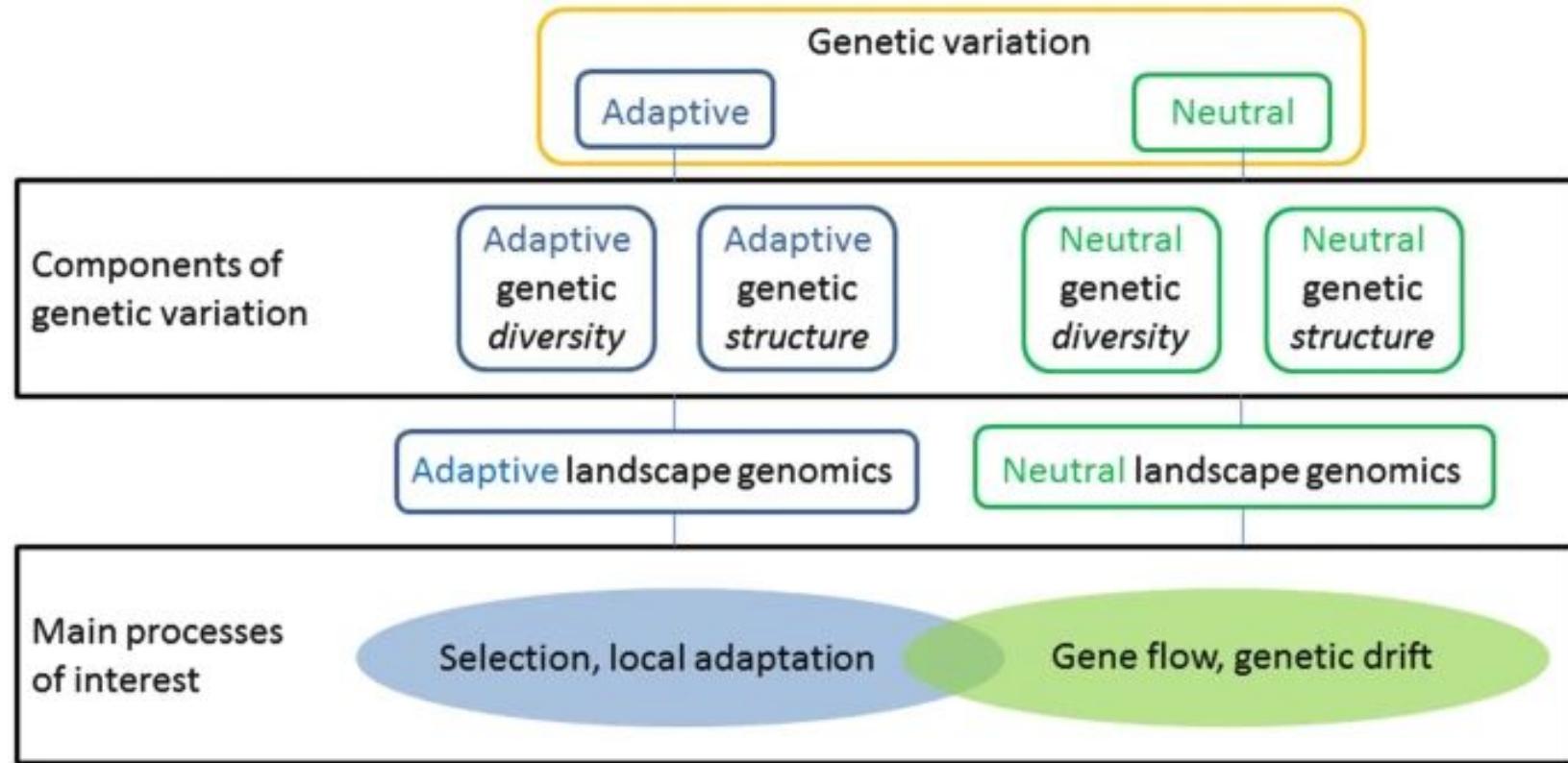


Resources for further training

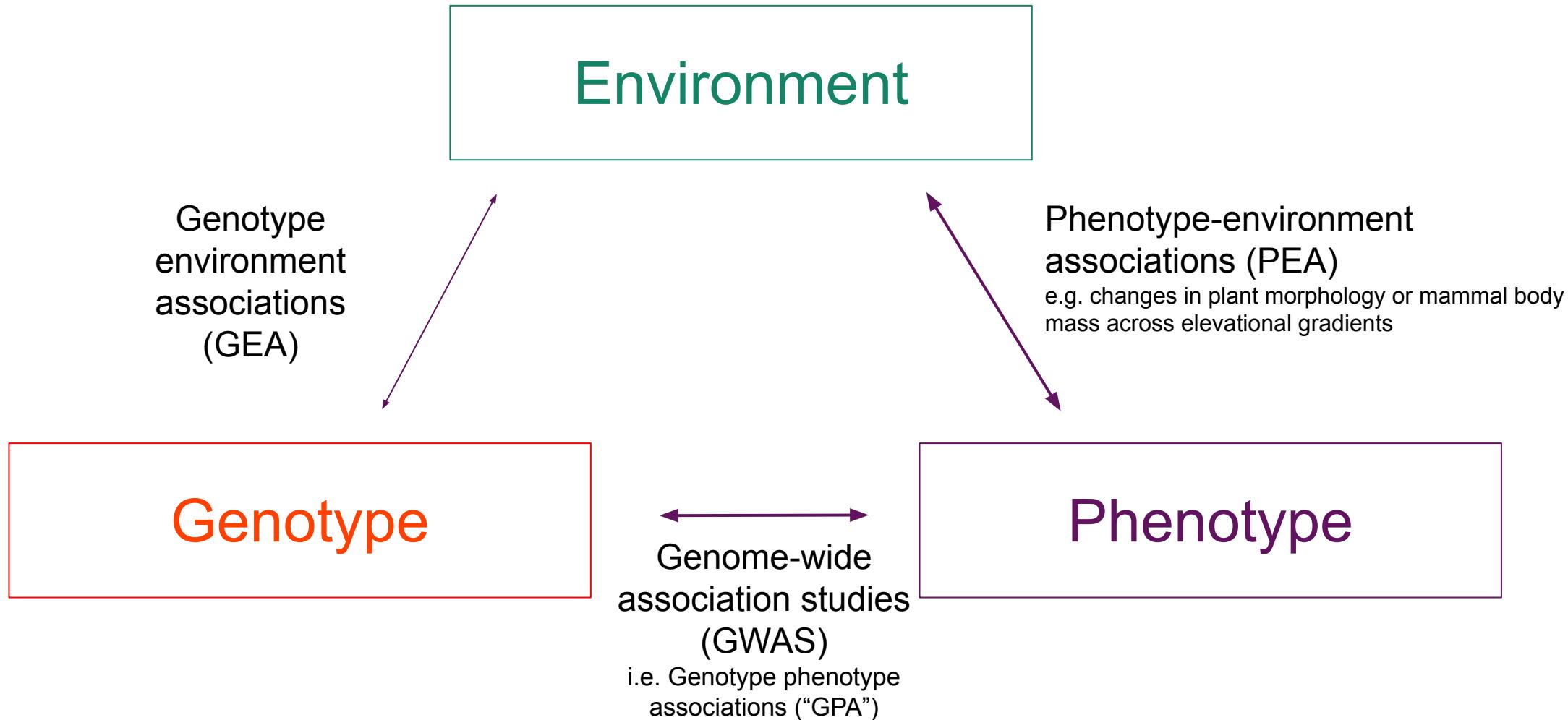
- https://bookdown.org/hhwagner1/LandGenCourse_book/
- <https://adegenet.r-forge.r-project.org/files/tutorial-spca.pdf>
- Balkenhol, N. et al. (2017). Landscape Genomics: Understanding Relationships Between Environmental Heterogeneity and Genomic Characteristics of Populations. In: Rajora, O. (eds) Population Genomics. Population Genomics. Springer, Cham.
https://doi.org/10.1007/13836_2017_2



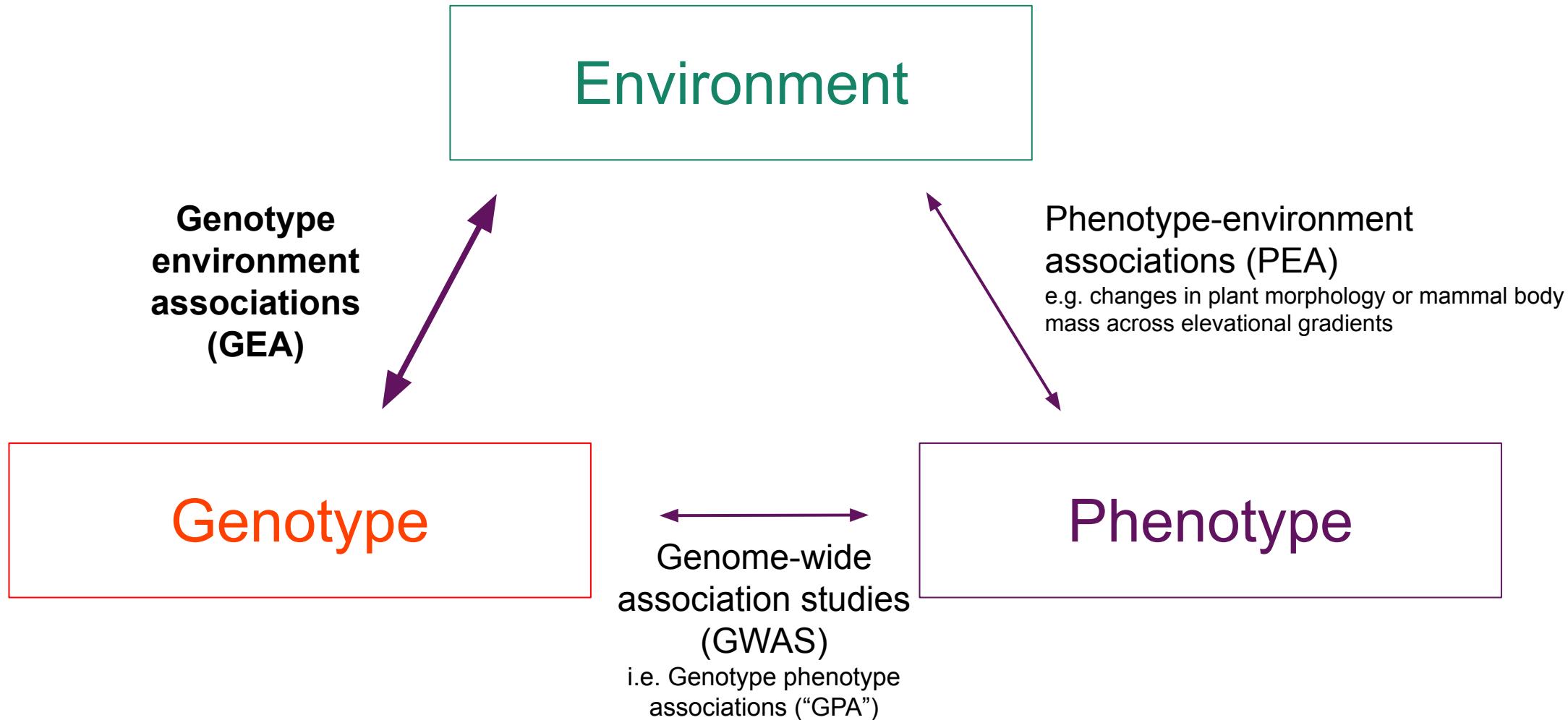
General framework



Genotype-Environment Associations (GEA)



Genotype-Environment Associations (GEA)



Genotype-Environment Associations (GEA)

A quite recent field, enabled by the advances in genome sequencing made over the last one or two decades (massive and relatively low-cost sequencing), such as Illumina sequencing:

Sequencing of one or few populations -> sequencing of many pop covering different environments/climates

Genotype-Environment Associations (GEA)

A quite recent field, enabled by the advances in genome sequencing made over the last one or two decades (massive and relatively low-cost sequencing), such as Illumina sequencing:

Sequencing of one or few populations -> sequencing of many pop covering different environments/climates

For most of the researchers, GEAs only refers to **correlations between allele frequency and environmental gradients** for some parameters associated with the environment of origin of populations (e.g. average temperature, precipitation sums, pH, etc.), with the objective to uncover locally adapted alleles and **identify the architecture/genetic bases of local adaptation**

Genotype-Environment Associations (GEA)

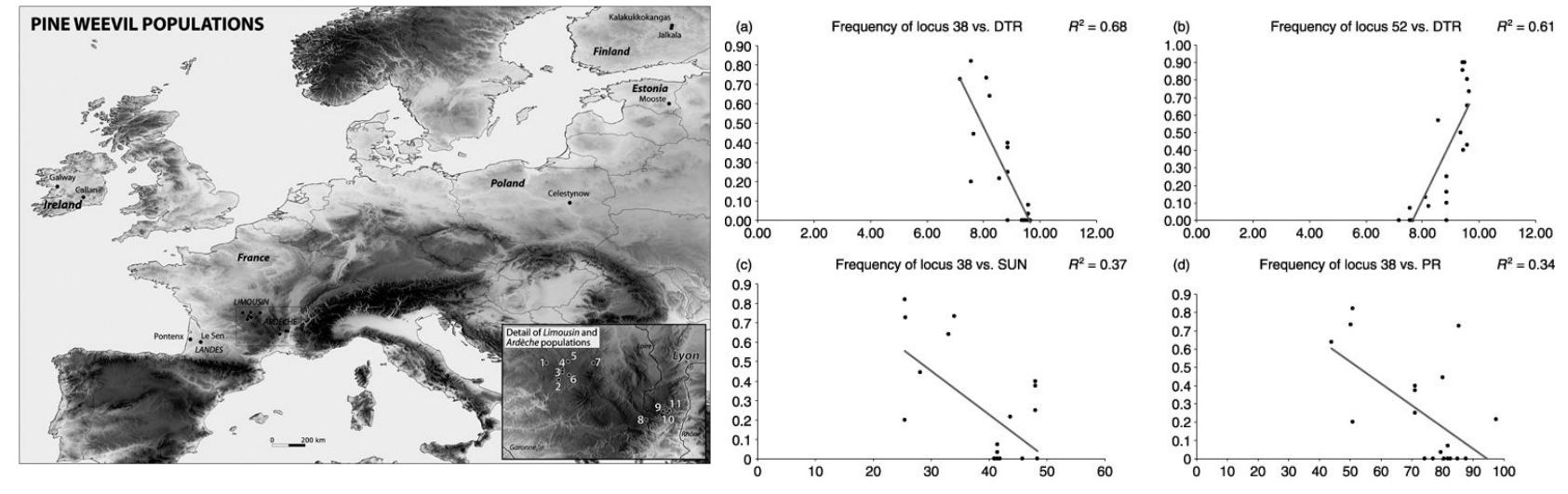
A quite recent field, enabled by the advances in genome sequencing made over the last one or two decades (massive and relatively low-cost sequencing), such as Illumina sequencing:

Sequencing of one or few populations -> sequencing of many pop covering different environments/climates

For most of the researchers, GEAs only refers to **correlations between allele frequency and environmental gradients** for some parameters associated with the environment of origin of populations (e.g. average temperature, precipitation sums, pH, etc.), with the objective to uncover locally adapted alleles and **identify the architecture/genetic bases of local adaptation**

Let's consider its simpler possible form: a linear cline of allele frequency along environmental gradients

Climate parameters (right):
DTR= diurnal temperature range
(difference between daily minimum and maximum temperature)
SUN= sunshine duration
PR = precipitation



BayPass core model

GENETICS | INVESTIGATION ■

Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates

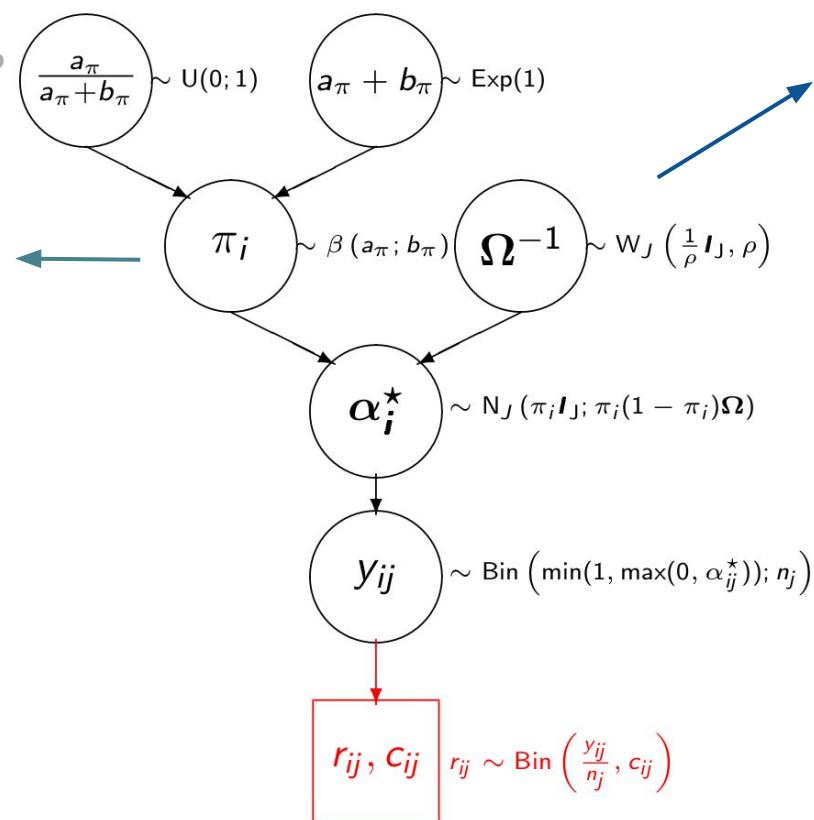
Mathieu Gautier¹

Developments of models that accounts for:

- Population structure (here omega matrix)
- more and more deviations from the linearity in the relationship between allele frequency and env. gradients

Priors (defined to take into account some SNP ascertainment bias)

Ancestral allele frequency (unobserved)



The general strategy developed in BayPass (Mathieu Gautier, 2015) ~ Bayenv2 (Günther & Coop, 2013)

Omega matrix:
Covariance matrix of allele frequencies
(genome-wide covariance due to admixture,
Isolation by distance, IBE, ...)

(allele counts at locus i for population j)

Gautier, 2015 Genetics

BayPass core model

GENETICS | INVESTIGATION ■

Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates

Mathieu Gautier¹

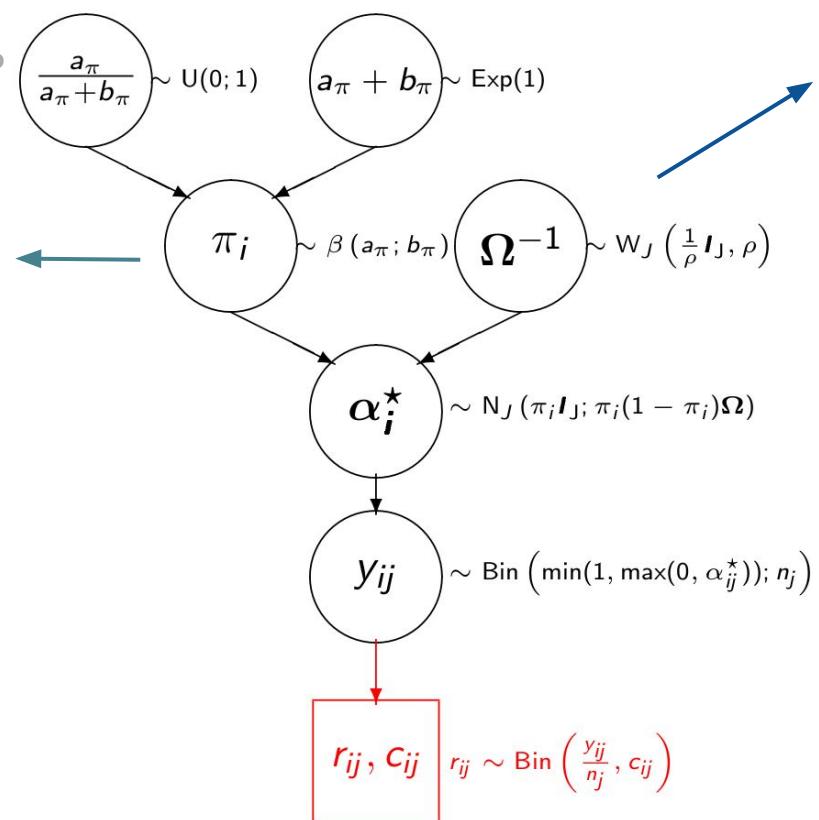
Developments of models that accounts for:

- Population structure (here omega matrix)
- more and more deviations from the linearity in the relationship between allele frequency and env. gradients

Priors (defined to take into account some SNP ascertainment bias)

Ancestral allele frequency (unobserved)

The general strategy developed in BayPass (Mathieu Gautier, 2015) ~ Bayenv2 (Günther & Coop, 2013)



Omega matrix:
Covariance matrix of allele frequencies
(genome-wide covariance due to
admixture, Isolation by distance, IBE, ...)

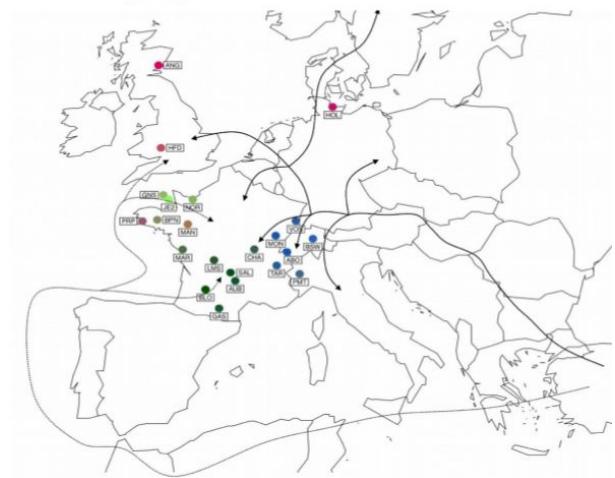
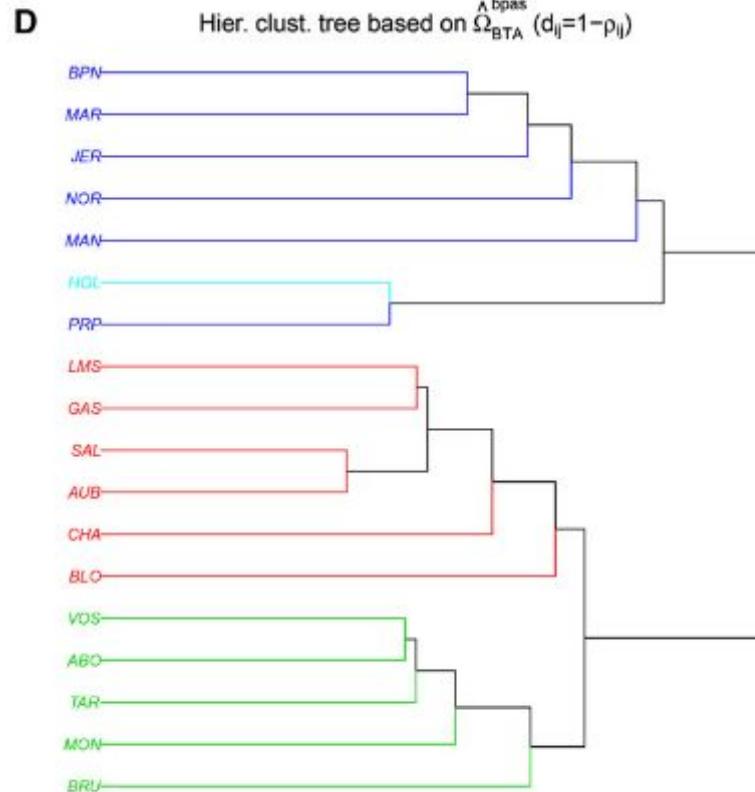
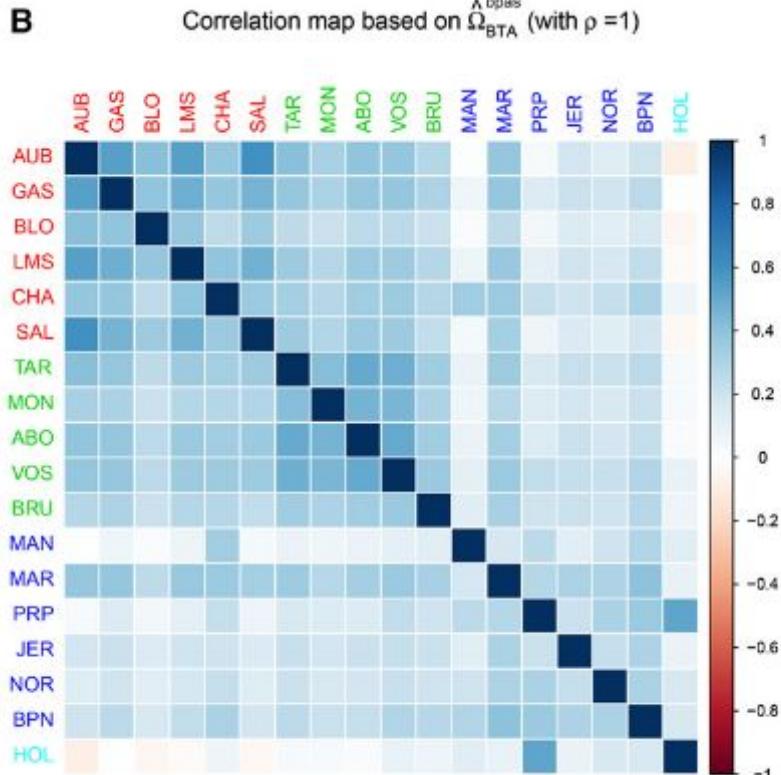
(allele counts at locus i
for population j)

Gautier, 2015 Genetics

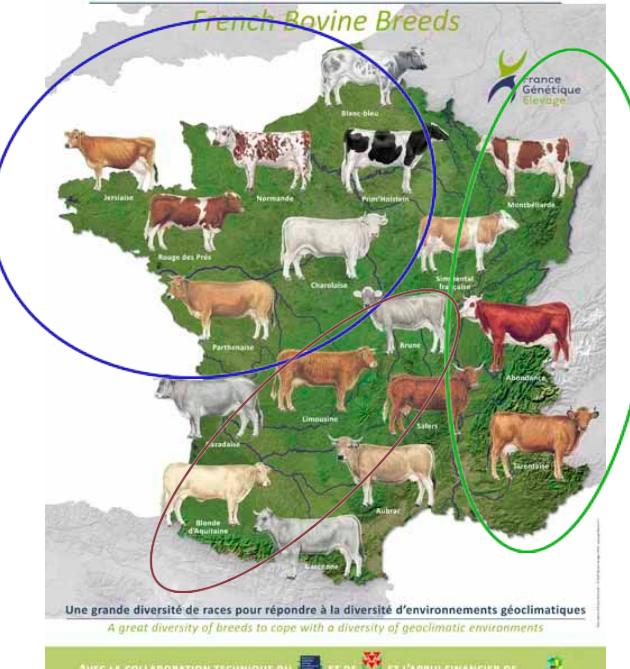
BayPass core model

omega matrix (ω , among-pop covariance in allele frequencies) captures the overall population structure in the data

453 individuals from 18 French cattle breeds genotyped at 42,056 SNPs



RACES BOVINES FRANCAISES



AVEC LA COLLABORATION TECHNIQUE DU ET DE ET L'APPUI FINANCIER DE

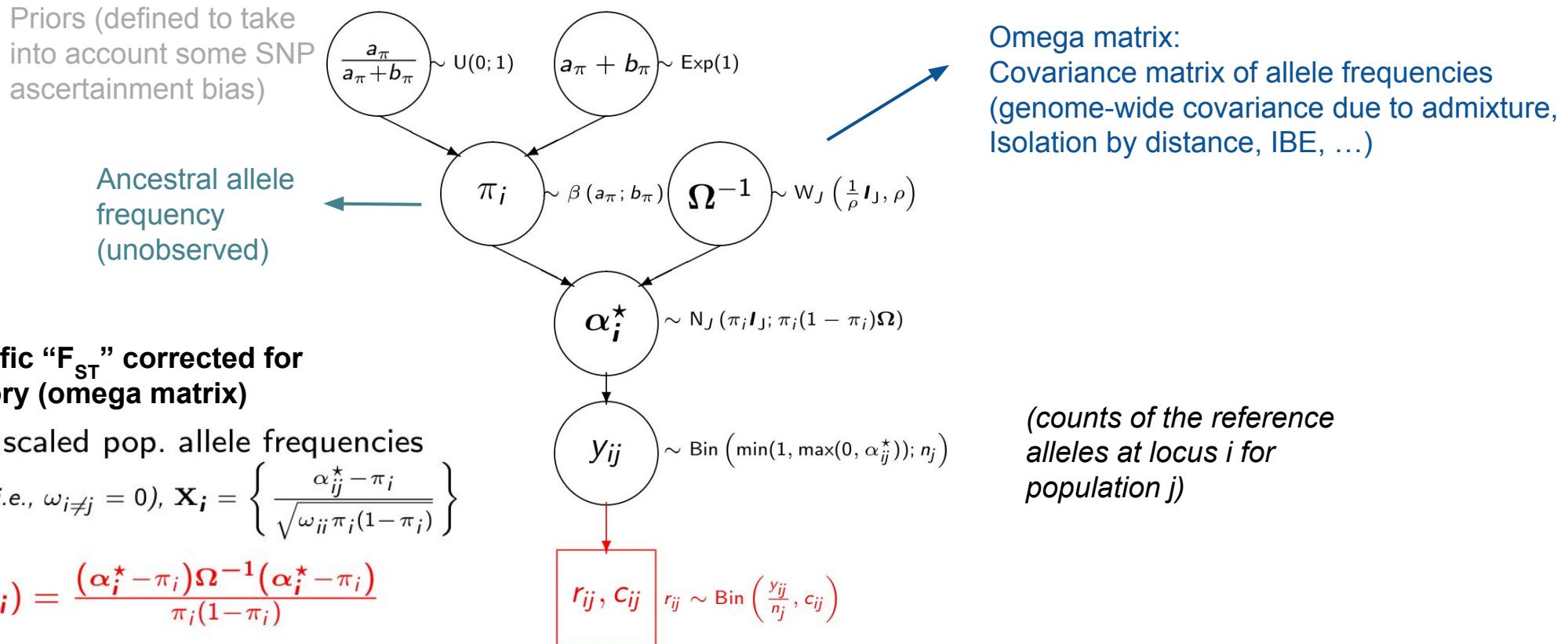
140 Rue de Berry 75012 Paris - France. +33 (0)1 40 04 52 02

France AgriMer

BayPass core model & selection scan

Developments of models that accounts for:

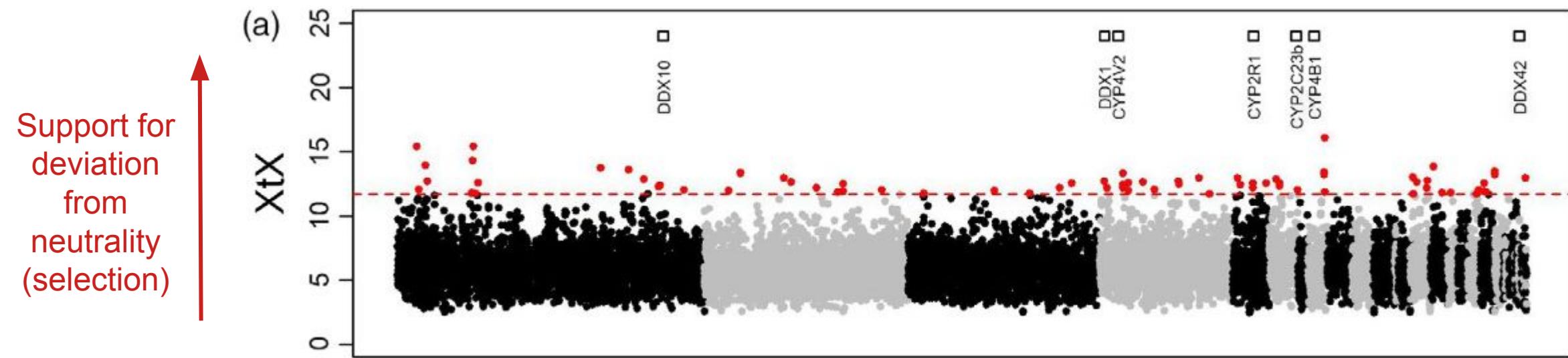
- Population structure (here omega matrix)
- more and more deviations from the linearity in the relationship between allele frequency and env. gradients



BayPass core model & selection scan

Signatures of adaptive divergence among populations of an avian species of conservation concern

Shawna J. Zimmerman¹  | Cameron L. Aldridge¹  | Kevin P. Oh²  |
Robert S. Cornman²  | Sara J. Oyler-McCance² 

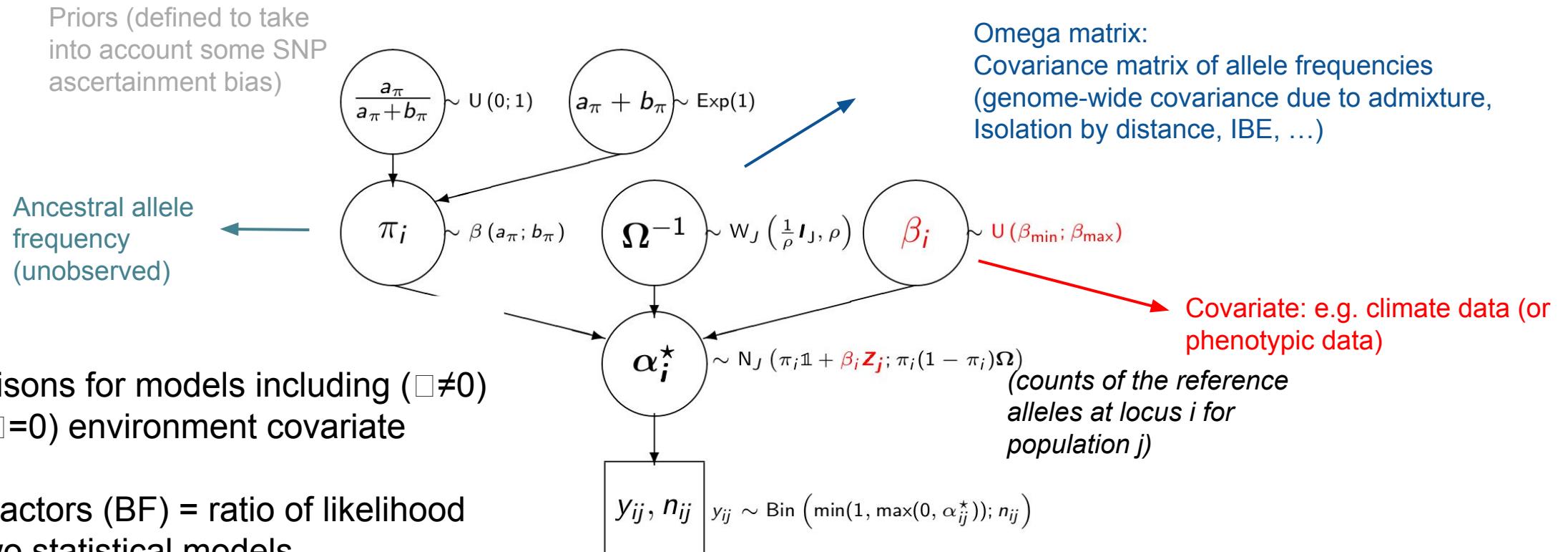


Xtx a F_{ST} -like metric accounting for population structure (omega matrix) that can be used to scan genomes for excess of differentiation

BayPass core model & detection of GEA

Developments of models that accounts for:

- Population structure
- more and more deviations from the linearity in the relationship between allele frequency and env. gradients



BayPass core model & detection of GEA



Evolution Letters, 2023, XX(XX), 1–11
<https://doi.org/10.1093/evlett/qrad043>
Letter

The genomics of adaptation to climate in European great tit (*Parus major*) populations

Joanne C. Stonehouse¹, Lewis G. Spurgin², Veronika N. Laine^{3,4,10}, Mirte Bosse^{5,6}, The Great Tit HapMap Consortium, Martien A.M. Groenen^{5,10}, Kees van Oers³, Ben C. Sheldon⁷, Marcel E. Visser³, Jon Slate¹

¹School of Biosciences, University of Sheffield, Sheffield, United Kingdom

²School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, United Kingdom

³Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands

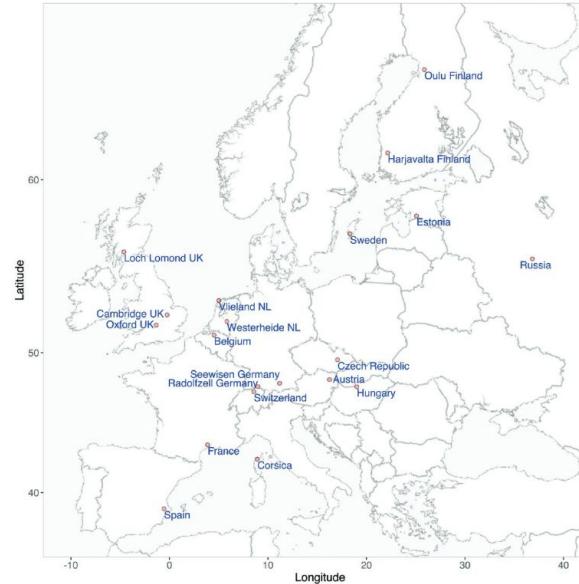
⁴Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland

⁵Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands

⁶Amsterdam Institute for Life and Environment (A-LIFE), Section Ecology and Evolution, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands

⁷Edward Grey Institute, Department of Biology, University of Oxford, Oxford, United Kingdom

Corresponding author: School of Biosciences, University of Sheffield, Western Bank, Sheffield S10 2TN, United Kingdom. Email: j slate@sheffield.ac.uk

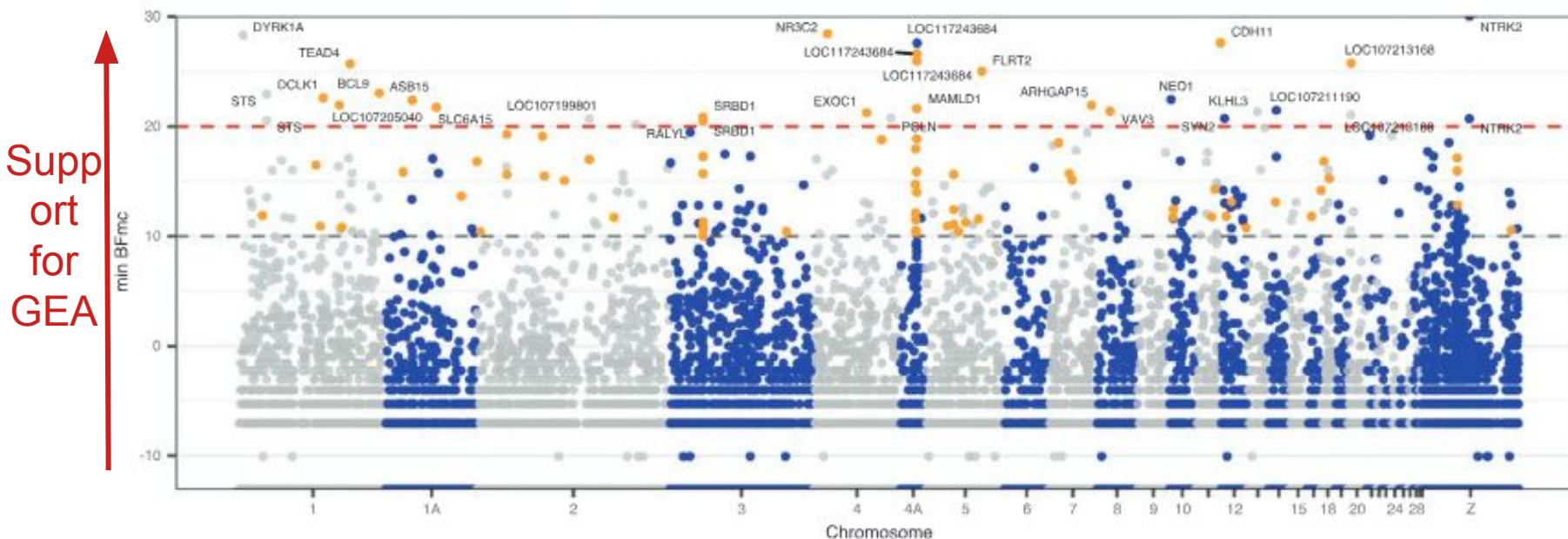


20 populations across europe

22 climate parameters:

19 bioclimate variables (Worldclim) + solar radiation
+ wind speed + water vapor pressure

-> all summarized using PCA (2 PCs)



Jeffreys' scale for Bayes Factors

Bayes Factor (Alternative/Null)	Strength of Evidence
> 100	Decisive evidence for alternative
10 – 30	Very strong evidence for alternative
3 – 10	Substantial evidence for alternative
1 – 3	Anecdotal evidence for alternative

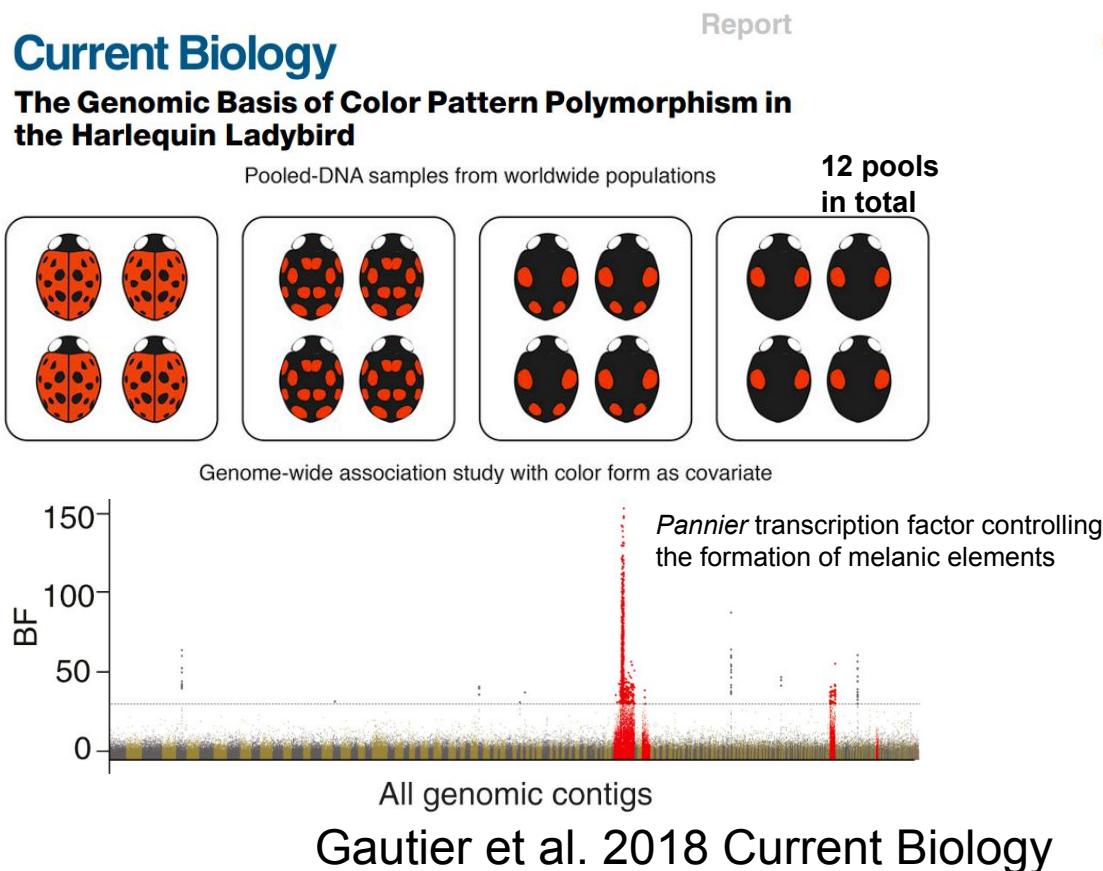
Empirically, in a GEA context, we often use $BF>15$ as a strong and $BF>20$ as a very strong support for GEA

→ However, it would be conceptually better to calibrate the results based on neutral simulations (-> practical)

“Population-GWAS”

GEA methods share the same rationale than GWAS methods, we are trying to identify associations between some parameters (phenotypes or environment-associated), but at the population level!

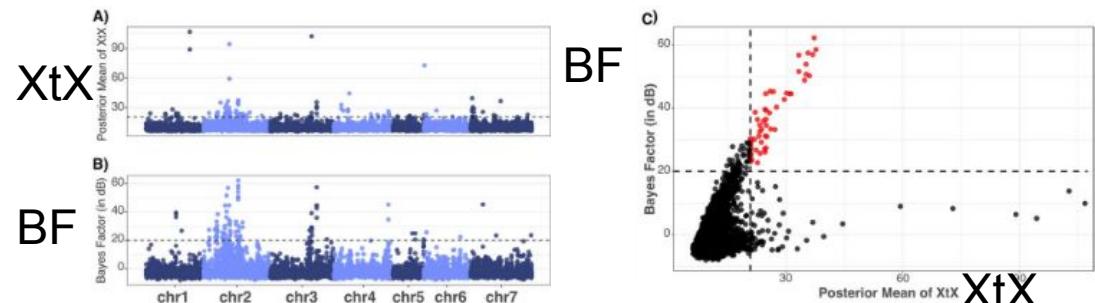
Consequently, GEA methods such as BayPass are sometimes used to perform “Population-GWAS”, the idea is to correlate allele frequency with a “mean phenotype” per population



OPEN Cost-effective detection of genome-wide signatures for 2,4-D herbicide resistance adaptation in red clover



Pooled Samples	Resistant (R) or Susceptible (S)	Number of Individuals	Raw Read Count	Illumina Platform	SRA Accession Number
Gen0	S	20	128,762,194	NextSeq 500	SRR8157540
Cherokee	S	20	117,929,104	NextSeq 500	SRR8157541
Southern Belle	S	26	144,553,898	NextSeq 500	SRR8157538
Gen6	R	48	100,041,080	NextSeq 500	SRR8157539
Kenland	S	27	34,717,334	HiSeq3000	SRR8157543
FL24DElite	R	24	39,236,182	HiSeq3000	SRR8157536
Early24D	R	28	37,229,808	HiSeq3000	SRR8157537
Late24D	R	27	41,948,810	HiSeq3000	SRR8157534
UK	R	30	35,126,246	HiSeq3000	SRR8157535
UKElite	R	26	38,124,186	HiSeq3000	SRR8157542



GEA & Genomic offsets

Back to GEA!

The objective of GEA is to identify genetic basis of adaptation to local climate/environment

- Assumes that you have uncovered genetic basis of adaptation
 - Set of adaptive loci (knowledge of allele frequency at adaptive loci in many pops)
- Assumes that you have precise information about how the environment is expected to change

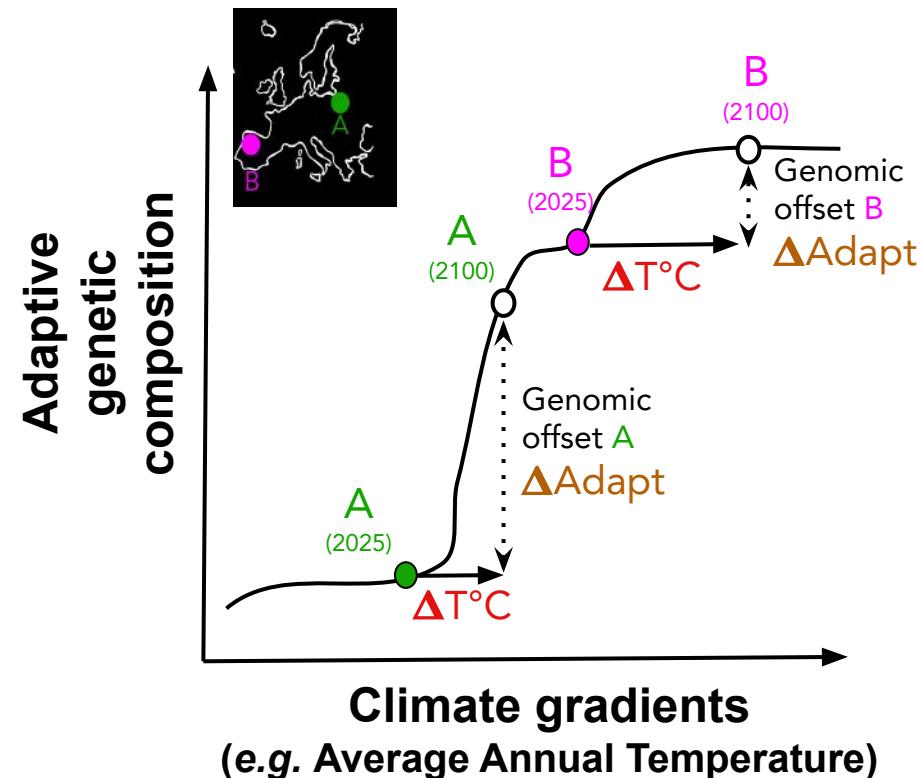
We could expect to be able to predict the change of allele frequency needed at adaptive loci for a population to persist in an environment (i.e. predict local maladaptation of a population to future environment)

- Emerging field, becoming especially important in the current climate change context.
- Global objective: predicting how populations will respond to climate change by estimating genomic offsets

GEA & Genomic offsets

Genomic offsets:

distance between the current and required genomic composition in a set of putatively adaptive loci under a future/changed environment



Climate gradients (x-axis)



Adaptive genetic composition (y-axis)

- Genetic bases for adaptation to climate
- Large present-day genome data (ideally at the landscape level)

A lot of future developments regarding genomic offsets expected in the next one or two decades...

Question 1

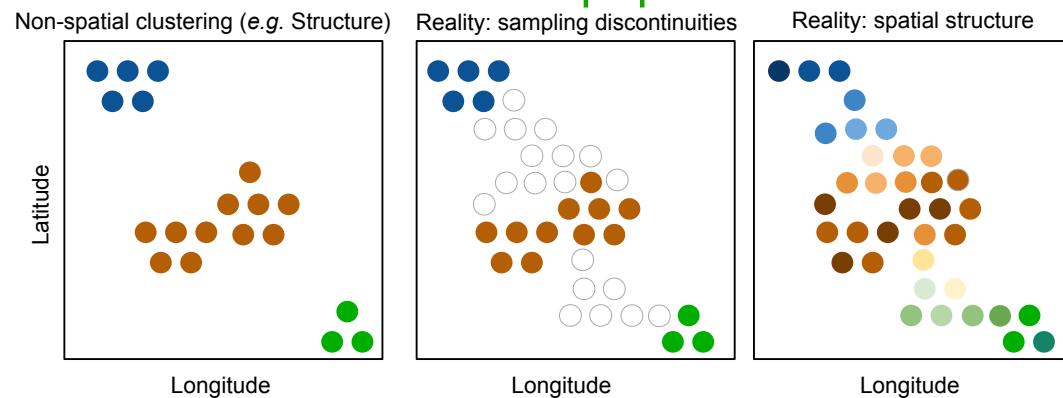
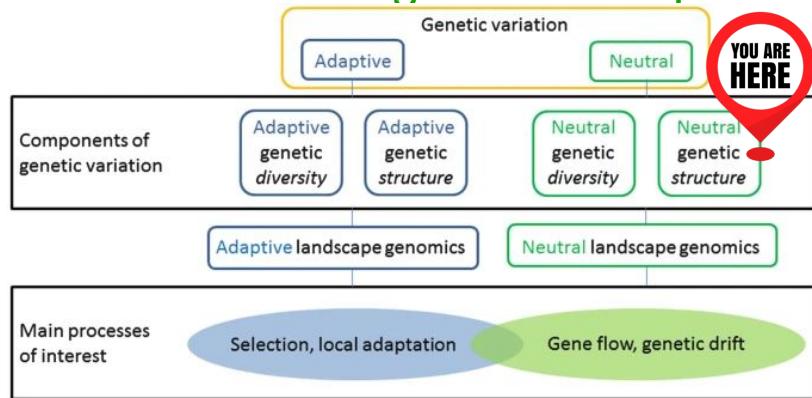
Why is it important to account for isolation by distance (IBD) when inferring population structure from genomic data?

- A - Because failing to control for IBD can lead to spurious signals of discrete population clusters
- B - Because IBD always reflects adaptive divergence between populations
- C - Because IBD ensures that migration rates are equal across all geographic distances
- D - Because controlling for IBD allows removal of all neutral variation from the data
- E - Because accounting for IBD helps distinguish clinal variation from true population boundaries

Question 1

Why is it important to account for isolation by distance (IBD) when inferring population structure from genomic data?

- A - Because failing to control for IBD can lead to spurious signals of discrete population clusters
- B - Because IBD always reflects adaptive divergence between populations
- C - Because IBD ensures that migration rates are equal across all geographic distances
- D - Because controlling for IBD allows removal of all neutral variation from the data
- E - Because accounting for IBD helps distinguish clinal variation from true population boundaries



Question 2

What are some advantages and limitations of using spatial PCA (sPCA) in landscape genomics?

- A - sPCA explicitly models spatial autocorrelation
- B - sPCA requires predefined population clusters as input
- C - sPCA can identify both global and local spatial structure in genetic variation
- D - Interpretation of sPCA axes may be challenging due to complex spatial signals
- E - sPCA is less informative than a simple Mantel test for detecting spatial patterns

Question 2

What are some advantages and limitations of using spatial PCA (sPCA) in landscape genomics?

A - sPCA explicitly models spatial autocorrelation

B - sPCA requires predefined population clusters as input

C - sPCA can identify both global and local spatial structure in genetic variation

D - Interpretation of sPCA axes may be challenging due to complex spatial signals

E - sPCA is less informative than a simple Mantel test for detecting spatial patterns

Question 3

What factors can influence the accuracy and interpretation of isolation-by-resistance (IBR) and isolation-by-environment (IBE) models in landscape genomics?

- A - The quality and resolution of environmental data layers used in resistance surfaces
- B - The assumption that the environment is static over time
- C - The long-term evolutionary history of populations
- D - Strict isolation-by-distance models cannot be tested against IBE and IBR
- E - The spatial sampling design

Question 3

What factors can influence the accuracy and interpretation of isolation-by-resistance (IBR) and isolation-by-environment (IBE) models in landscape genomics?

- A - The quality and resolution of environmental data layers used in resistance surfaces
- B - The assumption that the environment is static over time
- C - The long-term evolutionary history of populations
- D - Strict isolation-by-distance models cannot be tested against IBE and IBR
- E - The spatial sampling design

Question 4

What is the main purpose of genotype–environment association (GEA) methods in population genomics, and how can they be effectively combined with other approaches?

A - GEA methods can be used to identify correlations between phenotypic traits and environmental variables

B - GEA methods are designed to detect associations between allele frequencies and environmental gradients, suggesting potential local adaptation

C - Combining GEA methods with F_{ST} or F_{ST} -like genome scans helps identify candidate loci by detecting those that are both highly differentiated and environmentally associated

D - GEA methods are only effective when population structure is completely absent

E - GEA methods primarily estimate historical gene flow rates between populations under selection

Question 4

What is the main purpose of genotype–environment association (GEA) methods in population genomics, and how can they be effectively combined with other approaches?

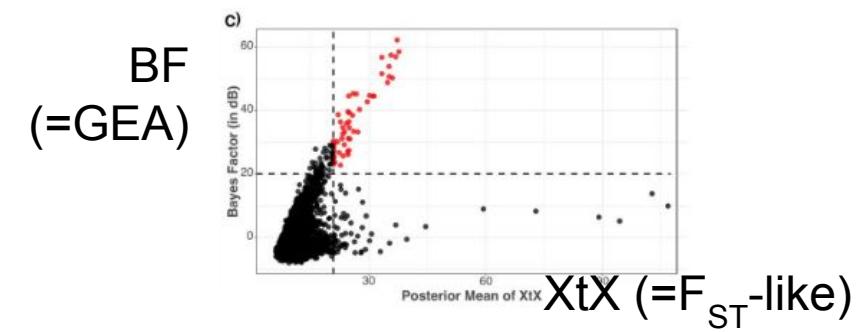
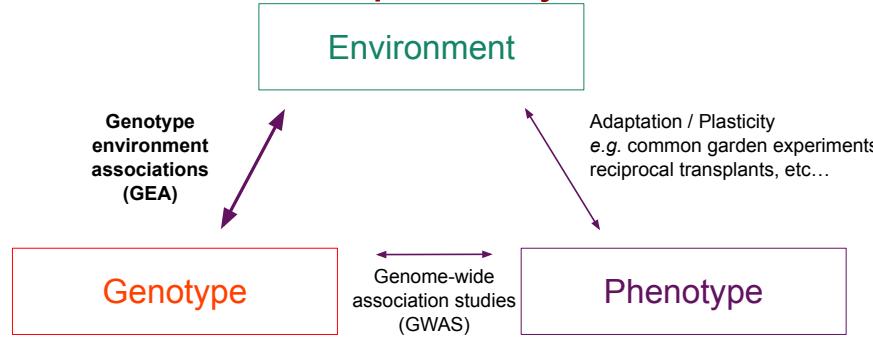
A - GEA methods can be used to identify correlations between phenotypic traits and environmental variables

B - GEA methods are designed to detect associations between allele frequencies and environmental gradients, suggesting potential local adaptation

C - Combining GEA methods with F_{ST} or F_{ST} -like genome scans helps identify candidate loci by detecting those that are both highly differentiated and environmentally associated

D - GEA methods are only effective when population structure is completely absent

E - GEA methods primarily estimate historical gene flow rates between populations under selection



Question 5

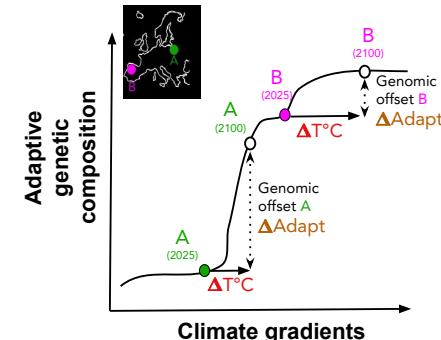
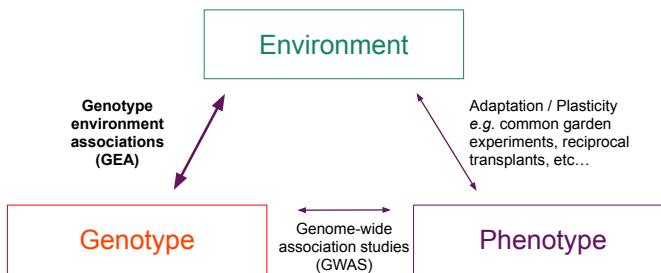
What is the promise of genomic offset methods, and how are they related to GEA analyses?

- A - Genomic offset methods aim to quantify how much species distribution range will shift with climate change
- B - Genomic offset quantifies how much the environment must change before gene flow breaks down among populations
- C - Genomic offset methods predict the potential maladaptation of populations under future environmental changes
- D - Genomic offset methods identify individuals with the greatest adaptive potential under climate change using phenotypic data
- E - Genomic offset relies on the associations inferred by GEA to estimate how allele frequencies would need to change to maintain adaptation under novel conditions

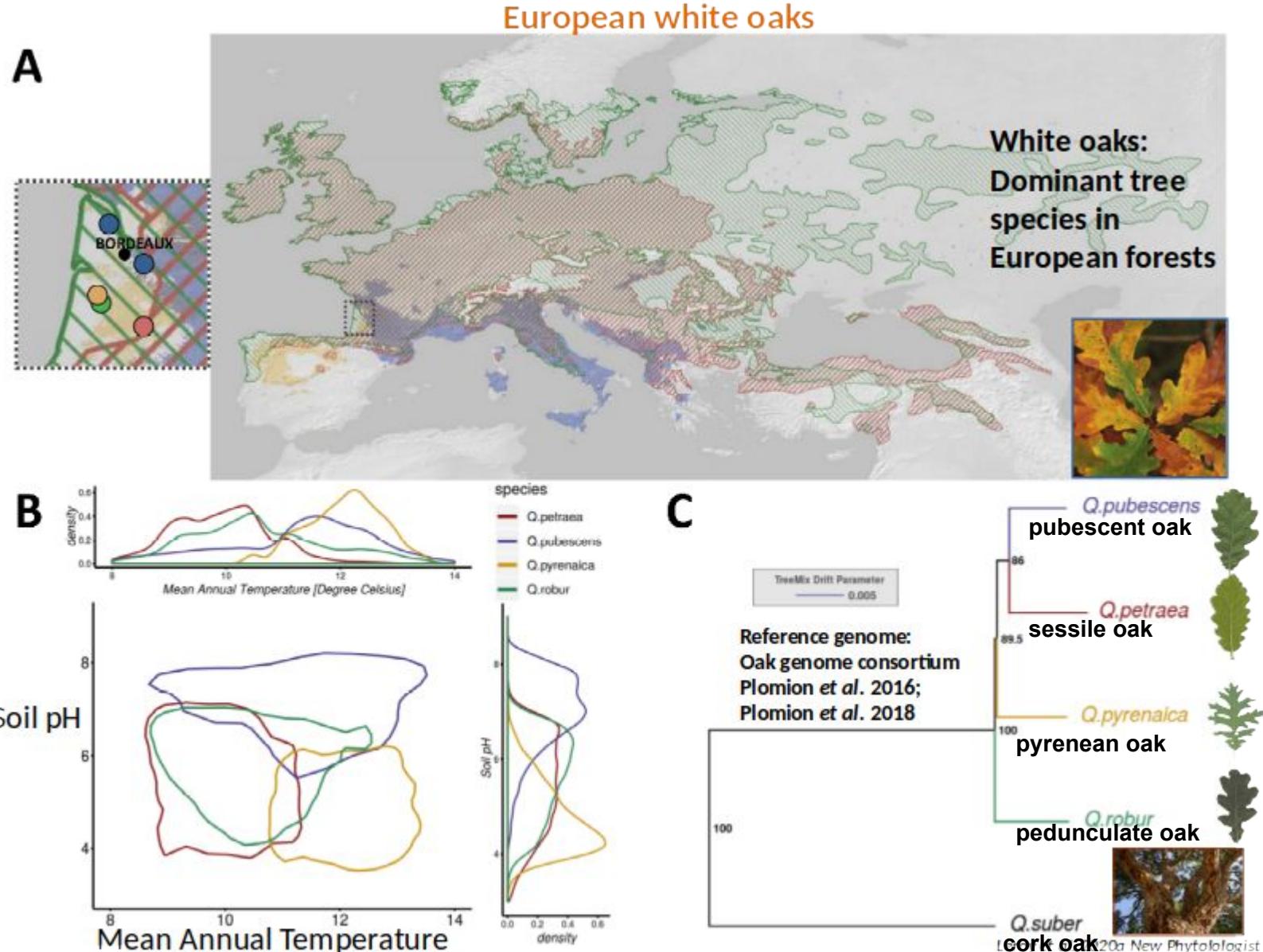
Question 5

What is the promise of genomic offset methods, and how are they related to GEA analyses?

- A - Genomic offset methods aim to quantify how much species distribution range will shift with climate change
- B - Genomic offset quantifies how much the environment must change before gene flow breaks down among populations
- C - Genomic offset methods predict the potential maladaptation of populations under future environmental changes
- D - Genomic offset methods identify individuals with the greatest adaptive potential under climate change using phenotypic data
- E - Genomic offset relies on the associations inferred by GEA to estimate how allele frequencies would need to change to maintain adaptation under novel conditions



Introducing the practical: European white oaks



Introducing the practical: European white oaks

- Genomic data (Pool-seq): 18 pops (sessile oak, *Q. petraea*)

- 10 populations at low elevation (25 ind/pool)
(7 in France, 2 in Germany, 1 in Ireland)
- 8 Pyrenean populations from low to quite high elevation
(up to 1630m; 10-20 ind/pool)

Synchronized mpileup format
(Step 8 from Monday's practical)

A:T:C:G:N:*

```
Sc0000000 3338 T 0:37:0:0:0:0 0:25:0:0:0:0 0:26:0:0:0:0 0:29:0:0:0:0 0:22:0:0:0:0 0:14:0:0:0:0  
0:25:0:0:0:0 0:31:0:0:0:0 0:42:0:0:0:0 0:43:0:0:0:0 0:23:0:0:0:0 0:16:0:0:0:0 0:20:0:0:0:0 0:14:0:0:0:0  
0:23:0:0:0:0 0:32:0:0:0:0 0:24:0:0:0:0 0:39:0:0:0:0  
Sc0000000 4078 n 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0  
0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0  
Sc0000000 14220 A 196:0:0:1:0:0 227:0:0:0:0:0 159:0:0:0:0:0 299:0:0:0:0:0 253:0:0:0:0:0 145:0:0:0:0:0 209:0:0:0:0:0  
304:0:0:0:0:0 300:0:0:0:0:0 397:0:1:0:0:0 223:0:0:0:0:0 157:0:0:0:0:0 194:0:0:0:0:0 262:0:0:0:0:0 279:1:0:0:0:0 199:0:0:0:0:0  
235:0:0:0:0:0 243:0:0:0:0:0  
Sc0000000 26478 C 0:0:113:0:0:0 0:0:134:0:0:0 0:0:141:0:0:0 0:0:171:0:0:0 0:0:153:0:0:0 0:0:138:0:0:0 0:0:138:0:0:0  
0:0:142:0:0:0 0:0:178:0:0:0 0:0:211:0:0:0 0:0:141:0:0:0 0:0:84:0:0:0 0:0:115:0:0:0 0:0:131:0:0:0 0:0:114:0:0:0 0:0:126:0:0:0  
0:0:119:0:0:0 0:0:106:0:0:0  
Sc0000000 26848 T 0:113:0:0:0:0 0:152:0:0:0:0 0:126:0:0:0:0 0:159:0:0:0:0 0:177:0:0:0:0 0:123:0:0:0:0 0:127:0:0:0:0  
0:170:0:0:0:0 0:219:0:0:0:0 0:219:0:0:0:0 0:148:0:0:0:0 0:106:0:0:0:0 0:170:0:0:0:0 0:114:0:0:0:0 0:148:0:0:0:0 0:174:0:0:0:0  
0:151:0:0:0:0 0:155:0:0:0:0
```

(continued)

Introducing the practical: European white oaks

- Genomic data (Pool-seq): 18 pops (sessile oak, *Q. petraea*)

- 10 populations at low elevation (25 ind/pool)
(7 in France, 2 in Germany, 1 in Ireland)
- 8 Pyrenean populations from low to quite high elevation
(up to 1630m; 10-20 ind/pool)

- Climate data (1950-2000, worldclim data):

Mean annual temperature & precipitation sums

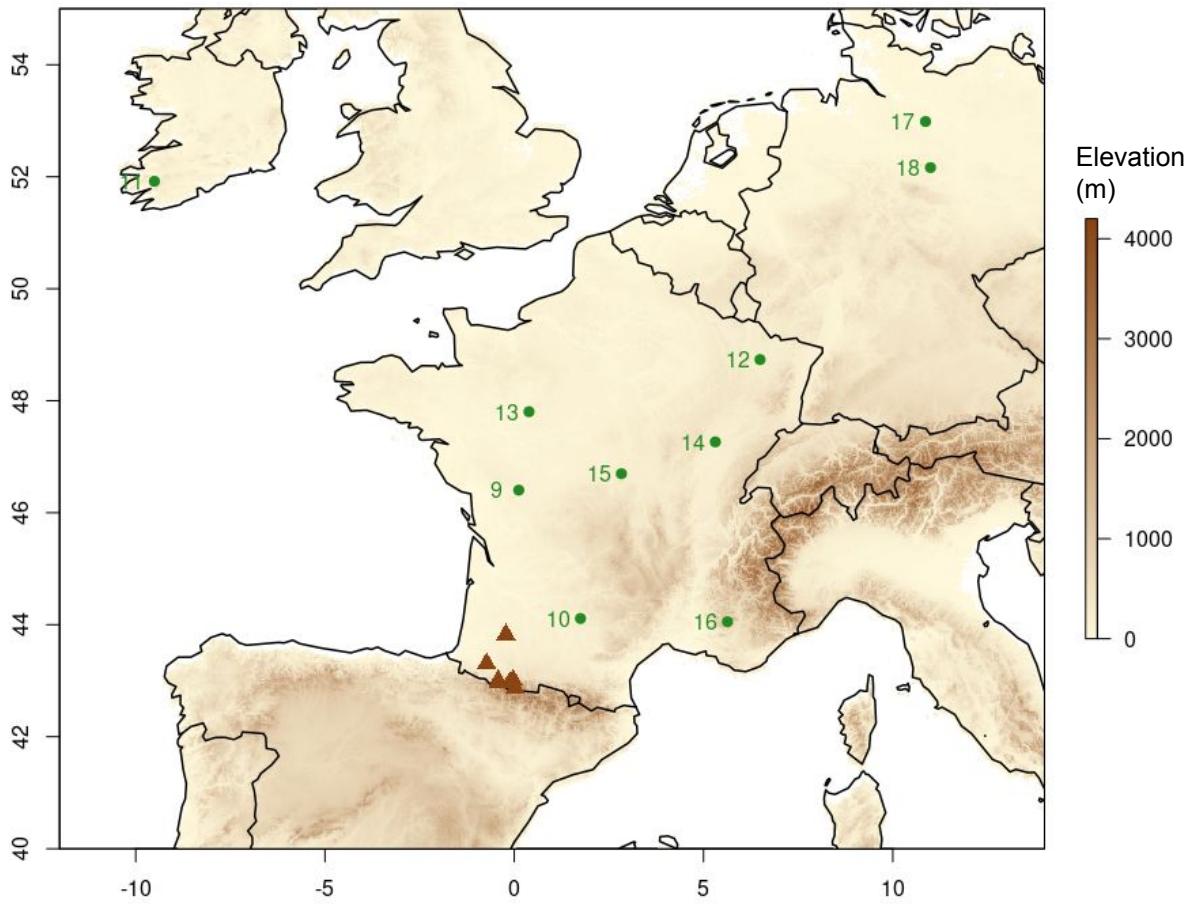
- Phenotypic data:

leaf unfolding in common gardens

Table 1 Geographic and climatic data for the *Quercus petraea* populations studied.

Code	Location	Elevation (m)	Latitude	Longitude	Temperature	Precipitation (mm yr ⁻¹)	Leaf unfolding	Sample size
Elevational gradient (French Pyrenees)								
L1	Laveyron, Luz Valley, France	131	43.75	-0.22	12.33	901	-1.333	20
L8	Chèze, Luz Valley, France	803	42.92	-0.03	9.27	914	0.817	20
L12	Gèdre, Luz Valley, France	1235	42.78	0.02	7.12	1016	1.011	20
L16	Péguyères, Luz Valley, France	1630	42.87	-0.12	6.58	982	1.724	18
O1	Josbaig, Ossau Valley, France	259	43.22	-0.73	12.24	979	-1.309	20
O8	Le Hourcq, Ossau Valley, France	841	42.90	-0.43	9.16	933	-0.324	20
O12	Gabas, Ossau Valley, France	1194	42.88	-0.42	7.35	1031	0.036	20
O16	Artouste, Ossau Valley, France	1614	42.88	-0.40	5.16	1164	0.427	10
Latitudinal gradient								
9	Saint Sauvant, France	155	46.38	0.12	11.78	786	-0.166	25
97	Grésigne, France	310	44.04	1.75	12.05	791	-1.139	25
124	Killamey, Ireland	50	52.01	-9.50	9.96	1362	4.084	25
204	Bézanges, France	275	48.76	6.49	9.50	751	0.371	25
217	Berdé, France	165	47.81	0.39	10.65	698	0.434	25
218	Longchamp, France	235	47.26	5.31	10.59	801	-0.920	22
219	Tronçais, France	245	46.68	2.83	10.63	742	1.350	25
233	Vachères, France	650	43.98	5.63	10.22	797	-1.532	25
253	Görde, Germany	85	53.10	10.86	8.30	635	0.953	25
256	Lappwald, Germany	180	52.26	10.99	8.50	597	0.650	25

Date of leaf unfolding expressed as standardized values for common gardens (see the Materials and Methods section). Negative values indicate early flushing, and positive values indicate late flushing.



Introducing the practical: European white oaks

- Genomic data (Pool-seq): 18 pops (sessile oak, *Q. petraea*)

- 10 populations at low elevation (25 ind/pool)
(7 in France, 2 in Germany, 1 in Ireland)
- 8 Pyrenean populations from low to quite high elevation
(up to 1630m; 10-20 ind/pool)

- Climate data (1950-2000, worldclim data):

Mean annual temperature & precipitation sums

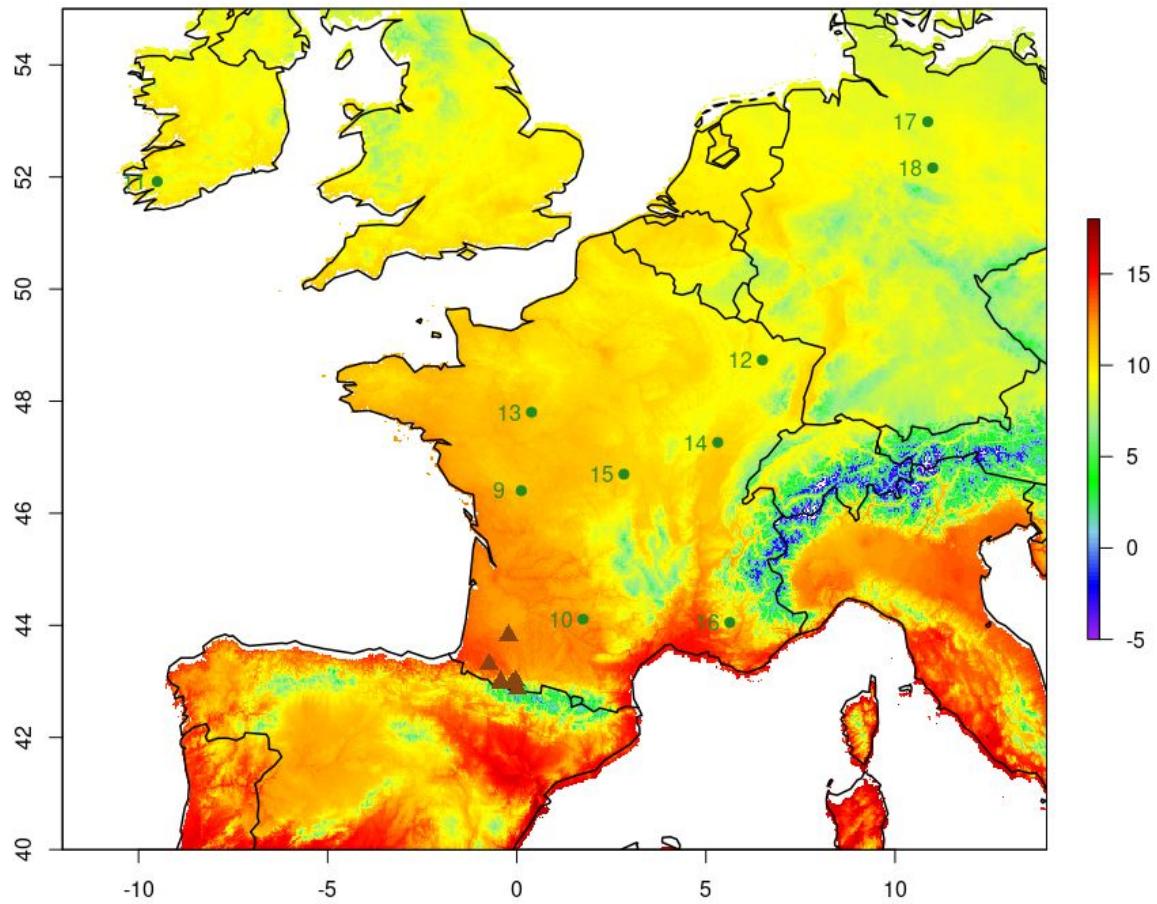
- Phenotypic data:

leaf unfolding in common gardens

Table 1 Geographic and climatic data for the *Quercus petraea* populations studied.

Code	Location	Elevation (m)	Latitude	Longitude	Temperature	Precipitation (mm yr ⁻¹)	Leaf unfolding	Sample size
Elevational gradient (French Pyrenees)								
L1	Laveyron, Luz Valley, France	131	43.75	-0.22	12.33	901	-1.333	20
L8	Chèze, Luz Valley, France	803	42.92	-0.03	9.27	914	0.817	20
L12	Gèdre, Luz Valley, France	1235	42.78	0.02	7.12	1016	1.011	20
L16	Péguères, Luz Valley, France	1630	42.87	-0.12	6.58	982	1.724	18
O1	Josbaig, Ossau Valley, France	259	43.22	-0.73	12.24	979	-1.309	20
O8	Le Hourcq, Ossau Valley, France	841	42.90	-0.43	9.16	933	-0.324	20
O12	Gabas, Ossau Valley, France	1194	42.88	-0.42	7.35	1031	0.036	20
O16	Artouste, Ossau Valley, France	1614	42.88	-0.40	5.16	1164	0.427	10
Latitudinal gradient								
9	Saint Sauvant, France	155	46.38	0.12	11.78	786	-0.166	25
97	Grésigne, France	310	44.04	1.75	12.05	791	-1.139	25
124	Killamey, Ireland	50	52.01	-9.50	9.96	1362	4.084	25
204	Bézanges, France	275	48.76	6.49	9.50	751	0.371	25
217	Berdé, France	165	47.81	0.39	10.65	698	0.434	25
218	Longchamp, France	235	47.26	5.31	10.59	801	-0.920	22
219	Tronçais, France	245	46.68	2.83	10.63	742	1.350	25
233	Vachères, France	650	43.98	5.63	10.22	797	-1.532	25
253	Görde, Germany	85	53.10	10.86	8.30	635	0.953	25
256	Lappwald, Germany	180	52.26	10.99	8.50	597	0.650	25

Date of leaf unfolding expressed as standardized values for common gardens (see the Materials and Methods section). Negative values indicate early flushing, and positive values indicate late flushing.



Introducing the practical: European white oaks

- Genomic data (Pool-seq): 18 pops (sessile oak, *Q. petraea*)

- 10 populations at low elevation (25 ind/pool)
(7 in France, 2 in Germany, 1 in Ireland)
- 8 Pyrenean populations from low to quite high elevation
(up to 1630m; 10-20 ind/pool)

- Climate data (1950-2000, worldclim data):

Mean annual temperature & precipitation sums

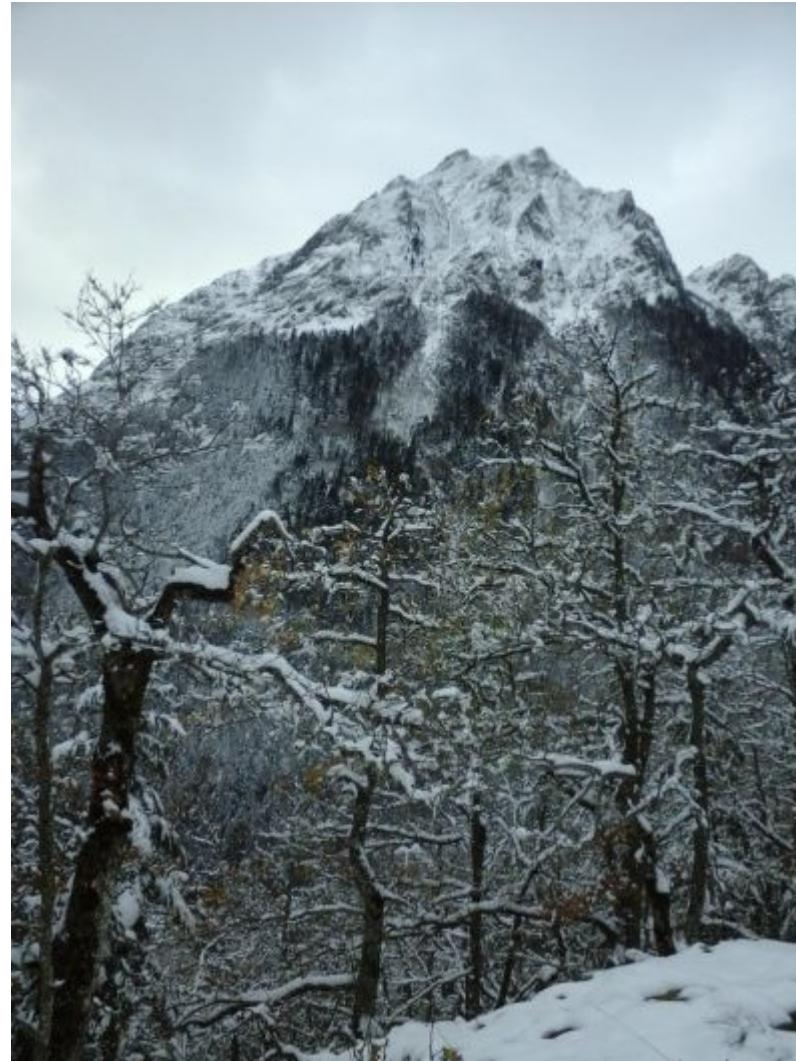
- Phenotypic data:

leaf unfolding in common gardens

Table 1 Geographic and climatic data for the *Quercus petraea* populations studied.

Code	Location	Elevation (m)	Latitude	Longitude	Temperature	Precipitation (mm yr ⁻¹)	Leaf unfolding	Sample size
Elevational gradient (French Pyrenees)								
L1	Laveyron, Luz Valley, France	131	43.75	-0.22	12.33	901	-1.333	20
L8	Chèze, Luz Valley, France	803	42.92	-0.03	9.27	914	0.817	20
L12	Gèdre, Luz Valley, France	1235	42.78	0.02	7.12	1016	1.011	20
L16	Péguères, Luz Valley, France	1630	42.87	-0.12	6.58	982	1.724	18
O1	Josbaig, Ossau Valley, France	259	43.22	-0.73	12.24	979	-1.309	20
O8	Le Hourcq, Ossau Valley, France	841	42.90	-0.43	9.16	933	-0.324	20
O12	Gabas, Ossau Valley, France	1194	42.88	-0.42	7.35	1031	0.036	20
O16	Artouste, Ossau Valley, France	1614	42.88	-0.40	5.16	1164	0.427	10
Latitudinal gradient								
9	Saint Sauvant, France	155	46.38	0.12	11.78	786	-0.166	25
97	Grésigne, France	310	44.04	1.75	12.05	791	-1.139	25
124	Killamey, Ireland	50	52.01	-9.50	9.96	1362	4.084	25
204	Bézanges, France	275	48.76	6.49	9.50	751	0.371	25
217	Berdé, France	165	47.81	0.39	10.65	698	0.434	25
218	Longchamp, France	235	47.26	5.31	10.59	801	-0.920	22
219	Tronçais, France	245	46.68	2.83	10.63	742	1.350	25
233	Vachères, France	650	43.98	5.63	10.22	797	-1.532	25
253	Görhrde, Germany	85	53.10	10.86	8.30	635	0.953	25
256	Lappwald, Germany	180	52.26	10.99	8.50	597	0.650	25

Date of leaf unfolding expressed as standardized values for common gardens (see the Materials and Methods section). Negative values indicate early flushing, and positive values indicate late flushing.



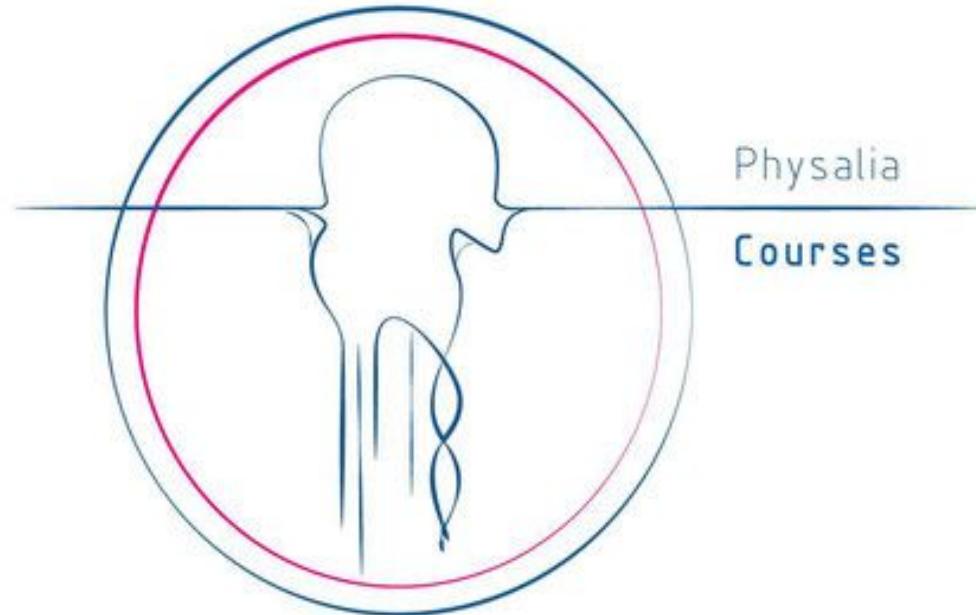
Leroy et al. 2020b New Phytologist

Landscape genomics

30/05/2025

Physalia course

Yann Bourgeois, Thibault Leroy



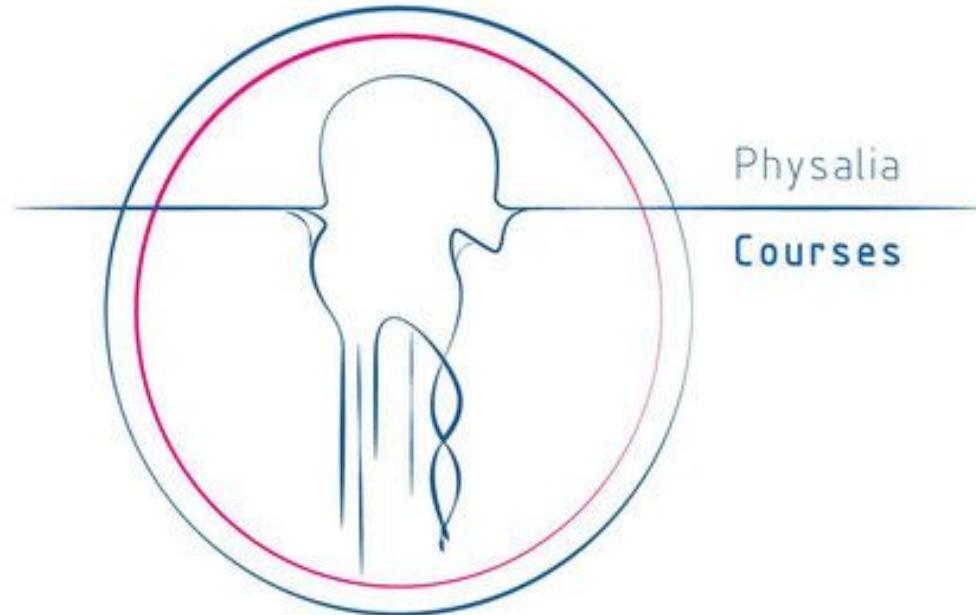
Landscape genomics

Recap - Practical

30/05/2025

Physalia course

Yann Bourgeois, Thibault Leroy



Sessile oak data

-> Day 1

- Genomic data (Pool-seq):

10 populations at low elevation (25 ind/pool)
(7 in France, 2 in Germany, 1 in Ireland)

8 Pyrenean populations from low to quite high elevation
(up to 1630m; 10-20 ind/pool)

more BNP-BHW_18pops_v2.3.pileup.200k.sync

```
Sc000000003338 T 0:37:0:0:0:0 0:25:0:0:0:0 0:26:0:0:0:0 0:29:0:0:0:0 0:22:0:0:0:0 0:14:0:0:0:0 0:25:0:0:0:0
    0:31:0:0:0:0 0:42:0:0:0:0 0:43:0:0:0:0 0:23:0:0:0:0 0:16:0:0:0:0 0:20:0:0:0:0 0:14:0:0:0:0 0:23:0:0:0:0
    0:32:0:0:0:0 0:24:0:0:0:0 0:39:0:0:0:0
Sc000000004078 n 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0
0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0 0:0:0:0:0:0
```

less BNP-BHW_18pops_v2.3.pileup.200k.sync | wc -l

213319

```
pooldata=popsync2pooldata(sync.file="BNP-BHW_18pops_v2.3.pileup.200k.sync",
+ poolsizes=c(50,50,50,50,50,44,50,50,50,50,40,40,40,40,40,20,36),
+ poolnames=c("9","97","124","204","217","218","219","233","253","256",
+ "L1","O1","L8","O8","O12","L12","O16","L16"),
+ min.rc=10,min.cov.per.pool = 50, max.cov.per.pool = 250,
+ min.maf=0.02,noindel=TRUE)
```

0.213319 millions lines processed in 0.2 min.; 10443 SNPs found

Data consists of **10443** SNPs for 18 Pools

$10443 \div 213319 \sim 0.049$

$1/0.049 = 20.4$

1 SNP every 20.4 bp on average
on this subset

Whole-genome:
 $\pi \sim 0.01$
Quite diverse species

Map & climate data

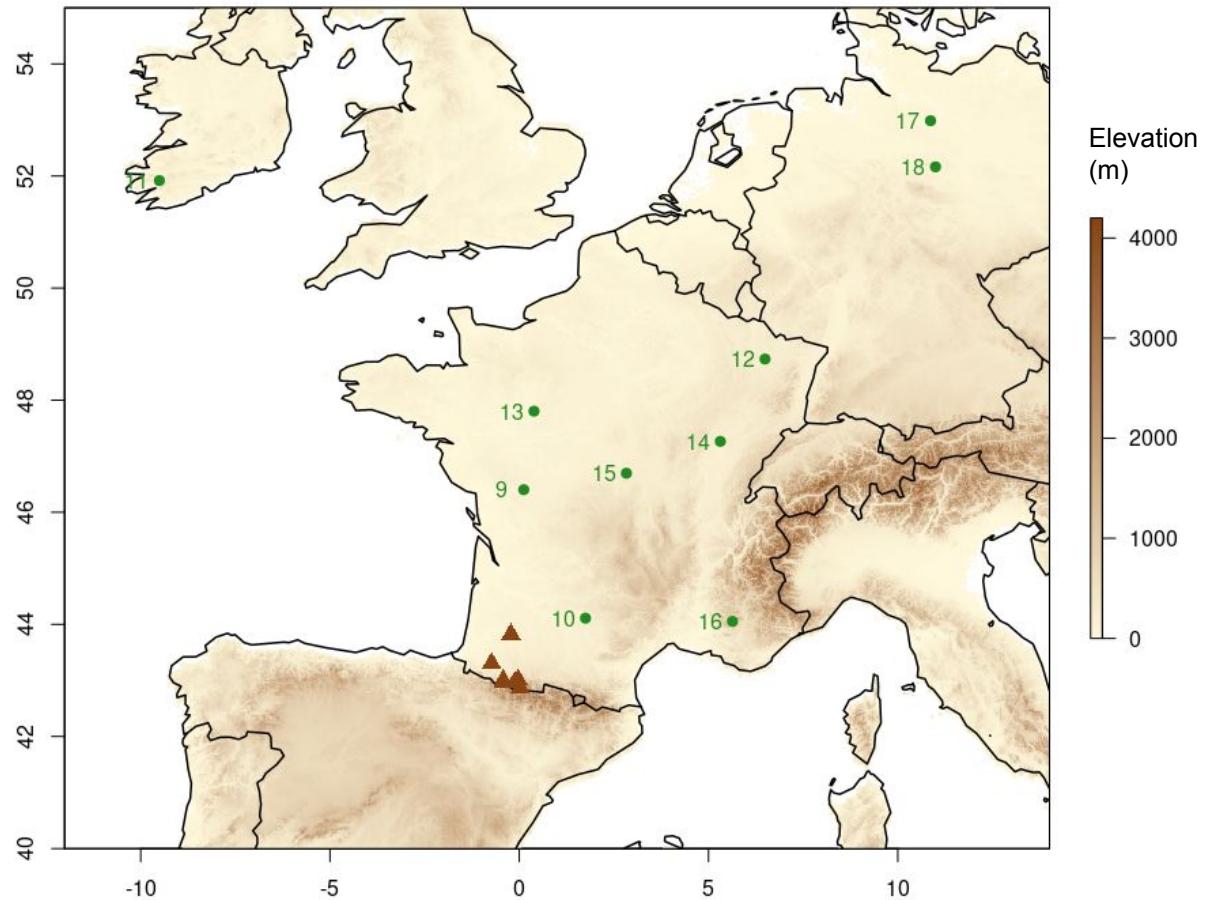
- Climate data (1950-2000, worldclim data):
Mean annual temperature & precipitation sums

- Phenotypic data:
leaf unfolding in common gardens

Table 1 Geographic and climatic data for the *Quercus petraea* populations studied.

Code	Location	Elevation (m)	Latitude	Longitude	Temperature	Precipitation (mm yr^{-1})	Leaf unfolding	Sample size
Elevational gradient (French Pyrenees)								
L1	Laveyron, Luz Valley, France	131	43.75	-0.22	12.33	901	-1.333	20
L8	Chèze, Luz Valley, France	803	42.92	-0.03	9.27	914	0.817	20
L12	Gèdre, Luz Valley, France	1235	42.78	0.02	7.12	1016	1.011	20
L16	Péguyères, Luz Valley, France	1630	42.87	-0.12	6.58	982	1.724	18
O1	Josbaig, Ossau Valley, France	259	43.22	-0.73	12.24	979	-1.309	20
O8	Le Hourcq, Ossau Valley, France	841	42.90	-0.43	9.16	933	-0.324	20
O12	Gabas, Ossau Valley, France	1194	42.88	-0.42	7.35	1031	0.036	20
O16	Artouste, Ossau Valley, France	1614	42.88	-0.40	5.16	1164	0.427	10
Latitudinal gradient								
9	Saint Sauvant, France	155	46.38	0.12	11.78	786	-0.166	25
97	Grésigne, France	310	44.04	1.75	12.05	791	-1.139	25
124	Killamey, Ireland	50	52.01	-9.50	9.96	1362	4.084	25
204	Bézanges, France	275	48.76	6.49	9.50	751	0.371	25
217	Bercé, France	165	47.81	0.39	10.65	698	0.434	25
218	Longchamp, France	235	47.26	5.31	10.59	801	-0.920	22
219	Tronçais, France	245	46.68	2.83	10.63	742	1.350	25
233	Vachères, France	650	43.98	5.63	10.22	797	-1.532	25
253	Görhde, Germany	85	53.10	10.86	8.30	635	0.953	25
256	Lappwald, Germany	180	52.26	10.99	8.50	597	0.650	25

Date of leaf unfolding expressed as standardized values for common gardens (see the Materials and Methods section). Negative values indicate early flushing, and positive values indicate late flushing.



Map & climate data

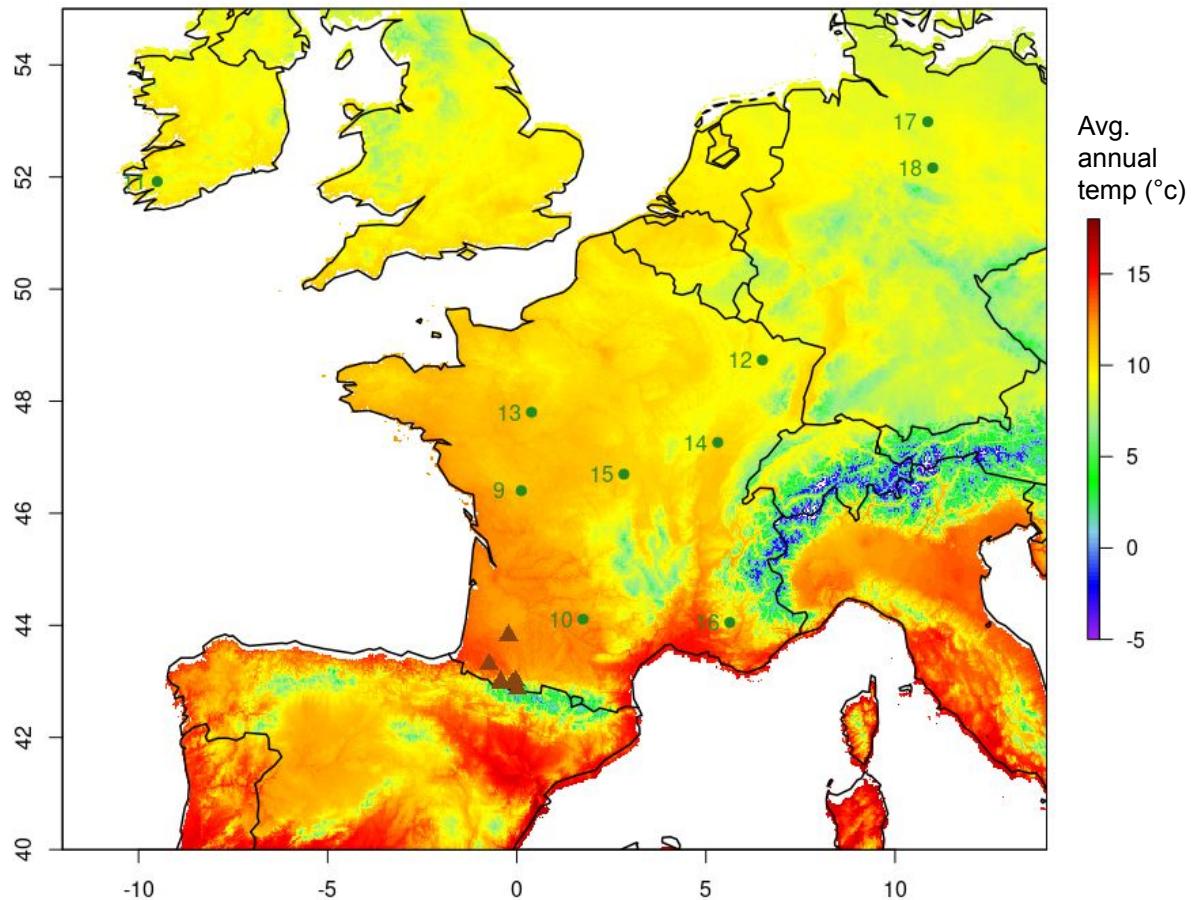
- Climate data (1950-2000, worldclim data):
Mean annual temperature & precipitation sums

- Phenotypic data:
leaf unfolding in common gardens

Table 1 Geographic and climatic data for the *Quercus petraea* populations studied.

Code	Location	Elevation (m)	Latitude	Longitude	Temperature	Precipitation (mm yr^{-1})	Leaf unfolding	Sample size
Elevational gradient (French Pyrenees)								
L1	Laveyron, Luz Valley, France	131	43.75	-0.22	12.33	901	-1.333	20
L8	Chèze, Luz Valley, France	803	42.92	-0.03	9.27	914	0.817	20
L12	Gèdre, Luz Valley, France	1235	42.78	0.02	7.12	1016	1.011	20
L16	Péguyères, Luz Valley, France	1630	42.87	-0.12	6.58	982	1.724	18
O1	Josbaig, Ossau Valley, France	259	43.22	-0.73	12.24	979	-1.309	20
O8	Le Hourcq, Ossau Valley, France	841	42.90	-0.43	9.16	933	-0.324	20
O12	Gabas, Ossau Valley, France	1194	42.88	-0.42	7.35	1031	0.036	20
O16	Artouste, Ossau Valley, France	1614	42.88	-0.40	5.16	1164	0.427	10
Latitudinal gradient								
9	Saint Sauvant, France	155	46.38	0.12	11.78	786	-0.166	25
97	Grésigne, France	310	44.04	1.75	12.05	791	-1.139	25
124	Killamey, Ireland	50	52.01	-9.50	9.96	1362	4.084	25
204	Bézanges, France	275	48.76	6.49	9.50	751	0.371	25
217	Bercé, France	165	47.81	0.39	10.65	698	0.434	25
218	Longchamp, France	235	47.26	5.31	10.59	801	-0.920	22
219	Tronçais, France	245	46.68	2.83	10.63	742	1.350	25
233	Vachères, France	650	43.98	5.63	10.22	797	-1.532	25
253	Görhde, Germany	85	53.10	10.86	8.30	635	0.953	25
256	Lappwald, Germany	180	52.26	10.99	8.50	597	0.650	25

Date of leaf unfolding expressed as standardized values for common gardens (see the Materials and Methods section). Negative values indicate early flushing, and positive values indicate late flushing.



Population structure -> Day 2

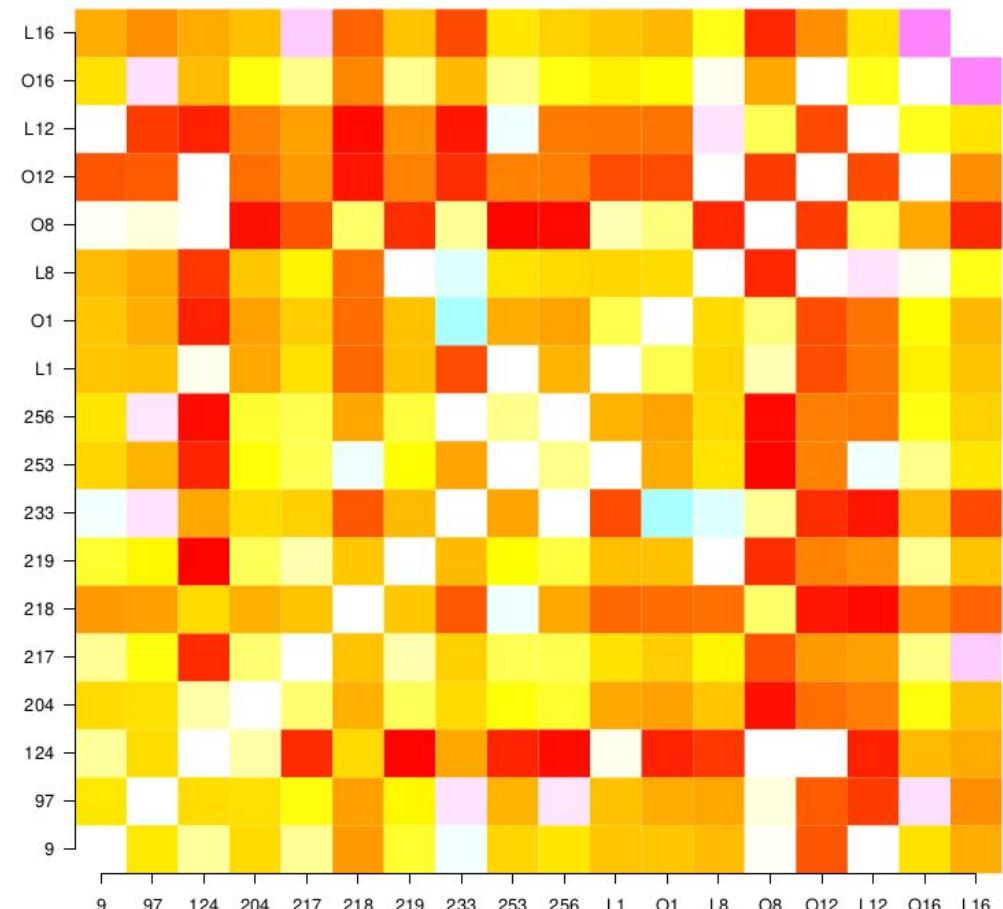
```
pairwise_fst_results <- compute.pairwiseFST(pooldata, method = "Anova",output.snp.values = FALSE)
Fstmatrix <- pairwise_fst_results@PairwiseFSTmatrix

pdf("Fst_matrix_res.pdf",width=10,height=10)

image(Fstmatrix,axes=FALSE,col=ifelse(Fstmatrix<0,rev(cm.colors(200)),rev(heat.colors(200))))
axis(1, at = seq(0, 1, length = nrow(Fstmatrix)), labels = rownames(Fstmatrix))
axis(2, at = seq(0, 1, length = ncol(Fstmatrix)), labels = colnames(Fstmatrix),las=2)

dev.off()
```

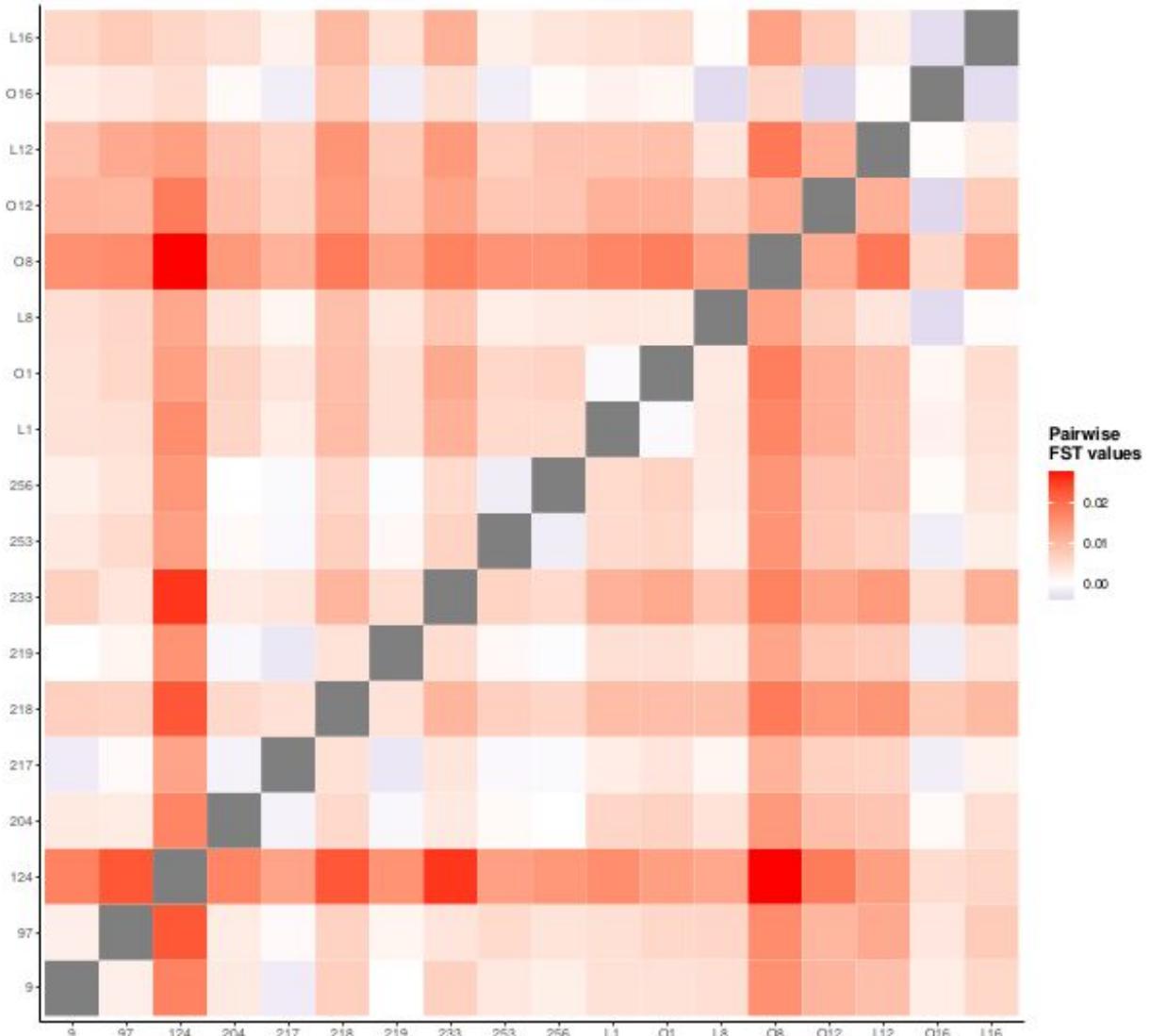
**A first evidence of the most differentiated populations
(124 vs. others, O12 & L12, ...)**



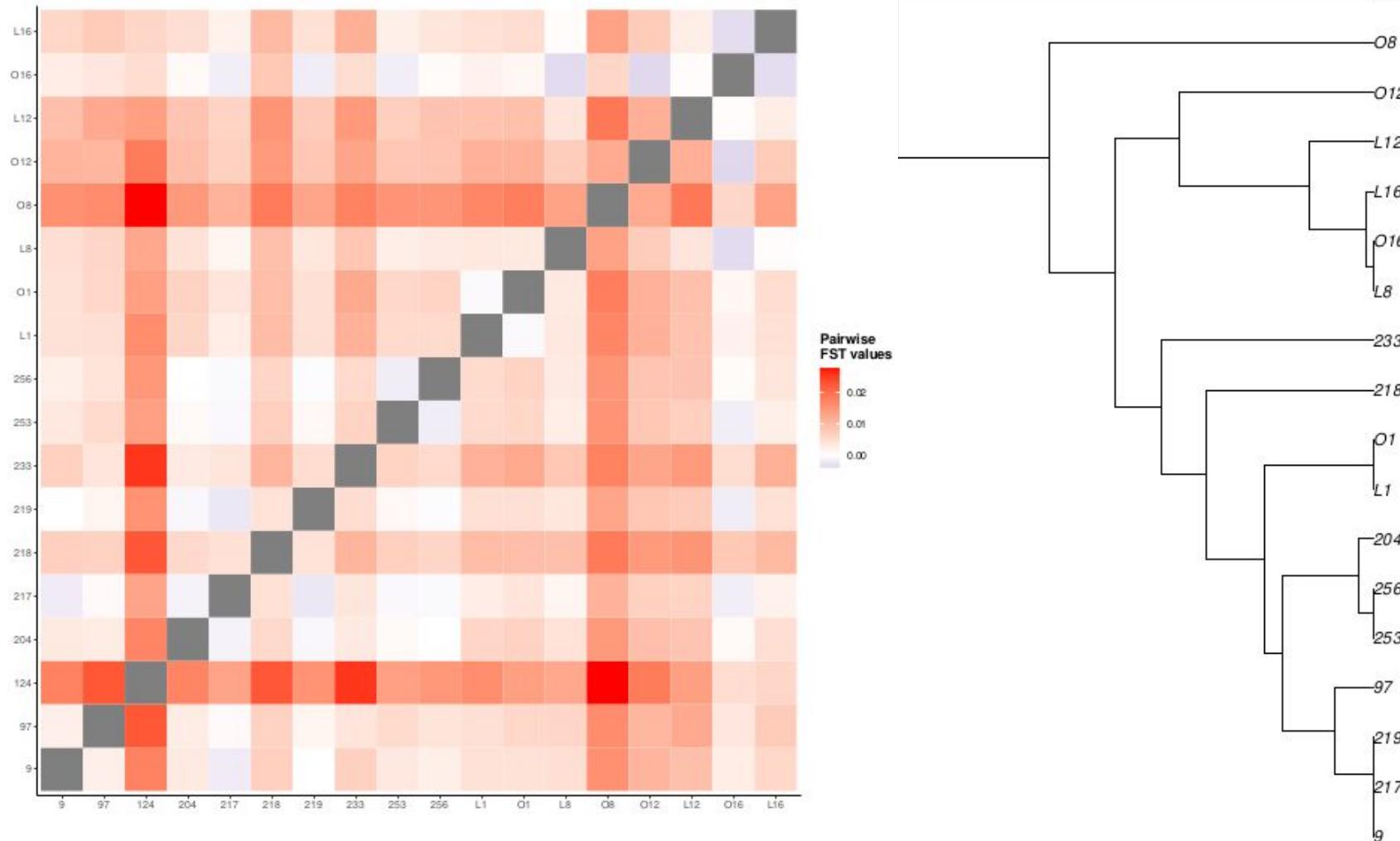
Population structure

-> Day 2

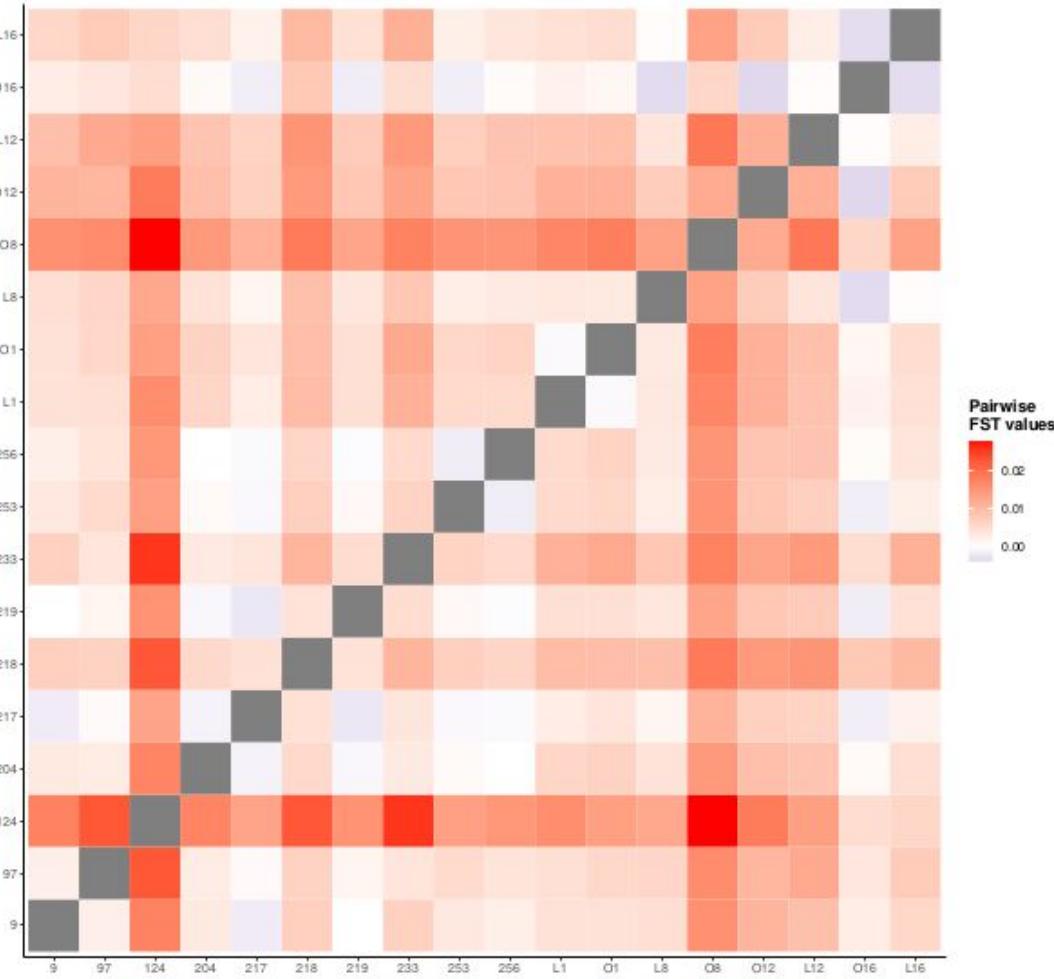
```
ggplot(data = melted_Fstmatrix, aes(x=X1, y=X2, fill=value))+  
  geom_tile() +  
  scale_fill_gradient2("Pairwise\nFST values", low = "navyblue",  
  mid = "white", high = "red", midpoint = 0) +  
  xlab("") + ylab("") +  
  scale_x_discrete(labels=c("Pool1" = "9", "Pool2" = "97",  
  "Pool3" = "124", "Pool4" = "204", "Pool5" = "217",  
  "Pool6" = "218", "Pool7" = "219", "Pool8" = "233",  
  "Pool9" = "253", "Pool10" = "256", "Pool11" = "L1",  
  "Pool12" = "O1", "Pool13" = "L8", "Pool14" = "O8",  
  "Pool15" = "O12", "Pool16" = "L12", "Pool17" = "O16",  
  "Pool18" = "L16")) +  
  scale_y_discrete(labels=c("Pool1" = "9", "Pool2" = "97",  
  "Pool3" = "124", "Pool4" = "204", "Pool5" = "217",  
  "Pool6" = "218", "Pool7" = "219", "Pool8" = "233",  
  "Pool9" = "253", "Pool10" = "256", "Pool11" = "L1",  
  "Pool12" = "O1", "Pool13" = "L8", "Pool14" = "O8",  
  "Pool15" = "O12", "Pool16" = "L12", "Pool17" = "O16",  
  "Pool18" = "L16")) +  
  theme_bw() + theme(panel.border = element_blank(),  
  panel.grid.major = element_blank(), panel.grid.minor = element_blank(),  
  axis.line = element_line(colour = "black")) + theme(legend.key  
  = element_blank()) +  
  theme(legend.title=element_text(size=12, face="bold"))
```



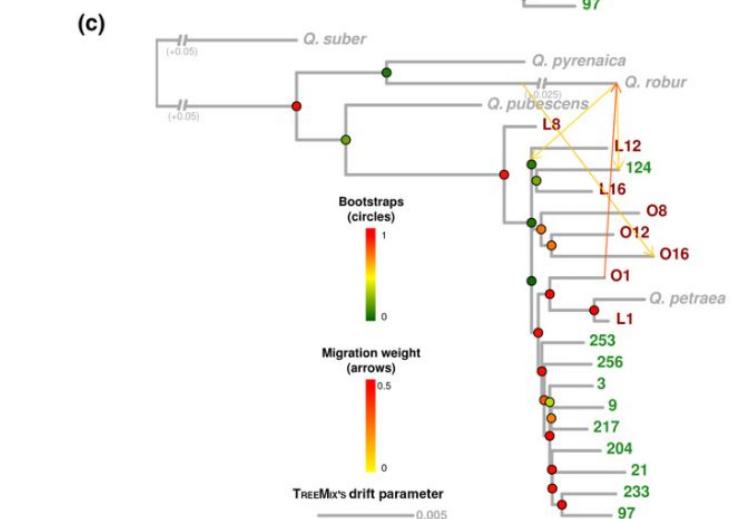
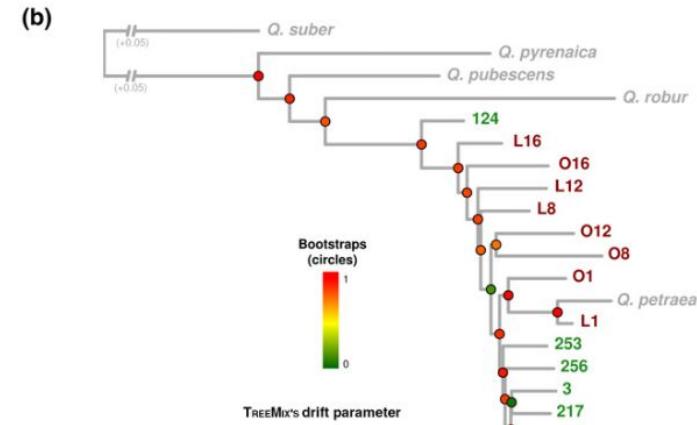
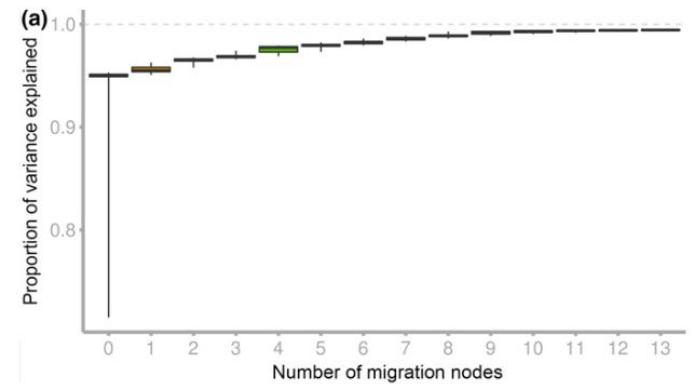
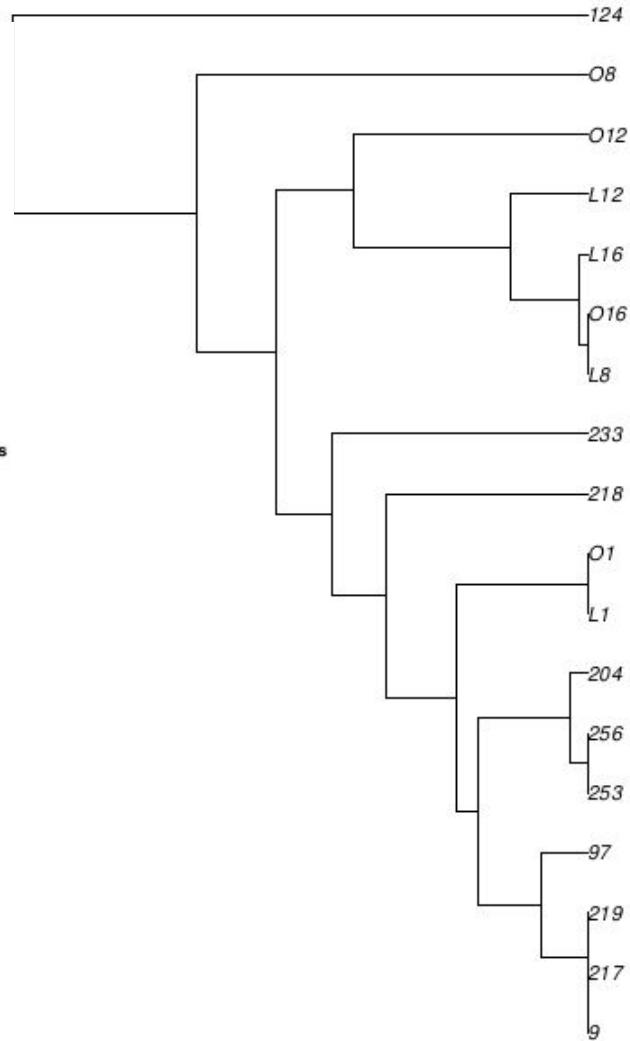
Population structure → Day 2



Population structure → Day 2



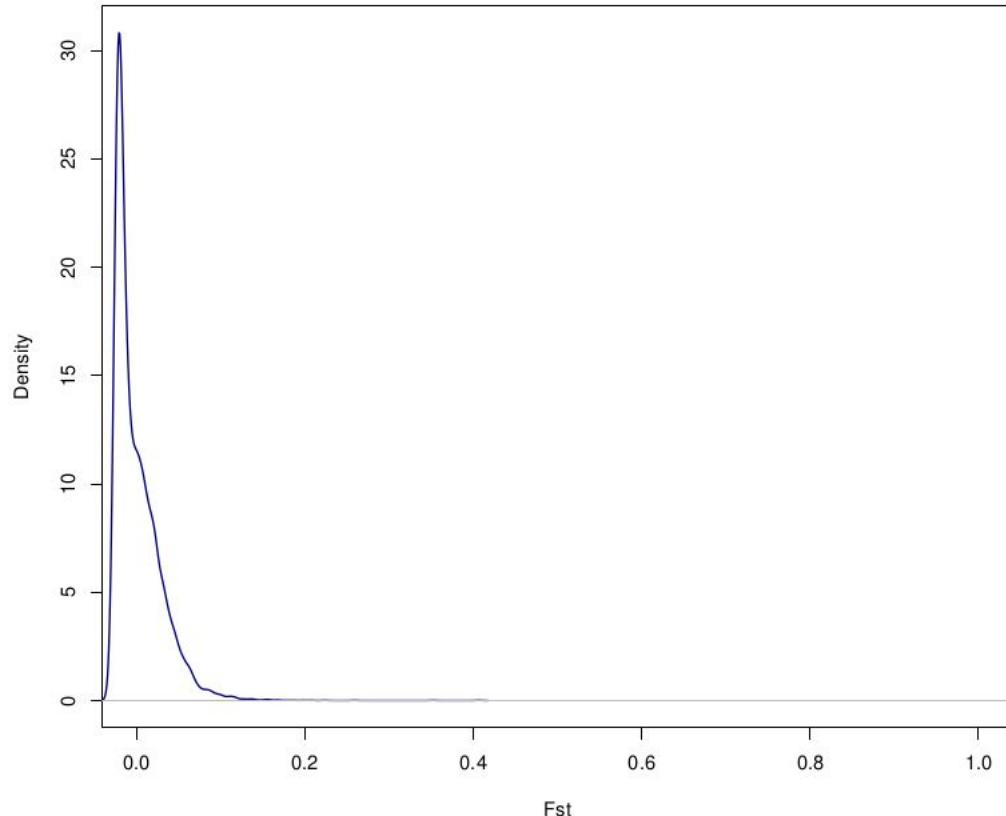
TreeMix



Global variation in F_{ST} -> (Day3 &) Day 4

```
SNPvalues=computeFST(pooldata, method="Anova")$snp.FST  
SNPvalues=as.data.frame(SNPvalues)
```

```
pdf("Distribution_Fstloci_allpops.pdf",width=9,height=8)  
plot(density(SNPvalues$SNPvalues),xlim=c(0,1),lwd=1.5,col="navyblue",xlab="Fst",main="")  
dev.off()
```



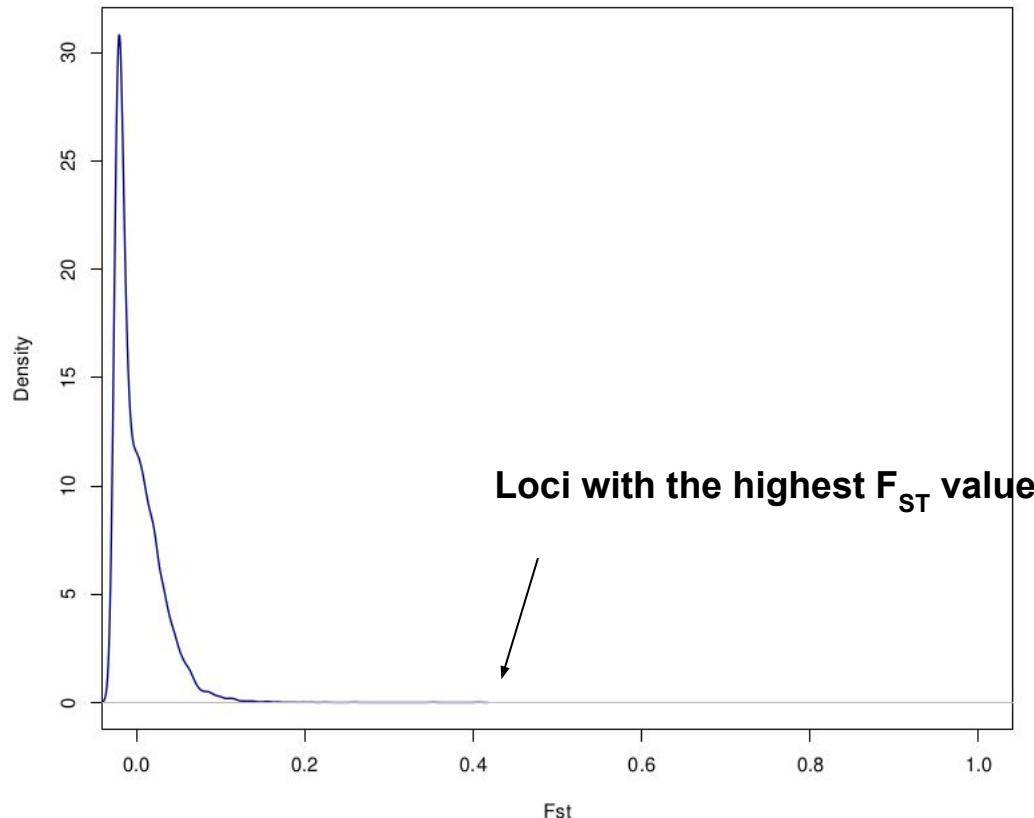
**Distribution of the global (among all pop)
 F_{ST} values (all loci) among all pools**

Global variation in F_{ST}

-> (Day3 &) Day 4

```
SNPvalues=computeFST(pooldata, method="Anova")$snp.FST  
SNPvalues=as.data.frame(SNPvalues)
```

```
pdf("Distribution_Fstloci_allpops.pdf",width=9,height=8)  
plot(density(SNPvalues$SNPvalues),xlim=c(0,1),lwd=1.5,col="navyblue",xlab="Fst",main="")  
dev.off()
```



Distribution of the global (among all pop) F_{ST} values (all loci) among all pools

```
print(paste("The greatest level of FST observed  
is:",max(SNPvalues$SNPvalues)))
```

```
[1] "The greatest level of FST observed is:  
0.406512841482787"
```

```
which(SNPvalues$SNPvalues==max(SNPvalues))  
[1] 6280
```

```
pooldata@snp.info[which(SNPvalues$SNPvalues==max(SNPvalues)),]  
Chromosome Position RefAllele AltAllele  
6280 Sc0000260 43938 A T
```

Global variation in F_{ST}

-> (Day3 &) Day 4

```
SNPvalues=computeFST(pooldata, method="Anova")$snp.FST  
SNPvalues=as.data.frame(SNPvalues)
```

```
pdf("Distribution_Fstloci_allpops.pdf",width=9,height=8)  
plot(density(SNPvalues$SNPvalues),xlim=c(0,1),lwd=1.5,col="navyblue",xlab="Fst",main="")  
dev.off()
```

```
pooldata@refallele.readcount[which(SNPvalues$SNPvalues==max(SNPvalues)),]  
[1] 19 61 64 4 5 30 1 21 13 9 143 120 15 7 15 9 10 5  
  
pooldata@readcoverage[which(SNPvalues$SNPvalues==max(SNPvalues)),]  
[1] 108 127 125 149 132 119 106 116 154 138 169 142 88 108 89 97 120 88  
  
pooldata@refallele.readcount[which(SNPvalues$SNPvalues==max(SNPvalues)),]/  
+ pooldata@readcoverage[which(SNPvalues$SNPvalues==max(SNPvalues)),]  
[1] 0.175925926 0.480314961 0.512000000 0.026845638 0.037878788 0.252100840  
[7] 0.009433962 0.181034483 0.084415584 0.065217391 0.846153846 0.845070423  
[13] 0.170454545 0.064814815 0.168539326 0.092783505 0.083333333 0.056818182  
  
>max(pooldata@refallele.readcount[which(SNPvalues$SNPvalues==max(SNPvalues)),]/pooldata@readcoverage[which(SNPvalues$SNPvalues==max(SNPvalues)),])  
[1] 0.8461538      (i.e. Pop L1)  
>  
min(pooldata@refallele.readcount[which(SNPvalues$SNPvalues==max(SNPvalues)),]/pooldata@readcoverage[which(SNPvalues$SNPvalues==max(SNPvalues)),])  
[1] 0.009433962   (i.e. Pop 219)
```

Distribution of the global (among all pop) F_{ST} values (all loci) among all pools

```
print(paste("The greatest level of FST observed  
is:",max(SNPvalues$SNPvalues)))
```

```
[1] "The greatest level of FST observed is:  
0.406512841482787"
```

```
which(SNPvalues$SNPvalues==max(SNPvalues))  
[1] 6280
```

```
pooldata@snp.info[which(SNPvalues$SNPvalues==max(SNPvalues)),]  
Chromosome Position RefAllele AltAllele  
6280 Sc0000260 43938 A T
```

Identifying most differentiated SNPs (all pops)

-> Day 4

```
quantile(SNPvalues$SNPvalues, probs = 0.99, na.rm = TRUE)
```

99%

0.09092675 ← 1% of SNPs have F_{ST} values exceeding this threshold (hypothetically under positive selection)

```
which(SNPvalues$SNPvalues > quantile(SNPvalues$SNPvalues, probs = 0.99, na.rm = TRUE))
```

```
[1] 27 60 74 282 300 383 404 432 908 1129 1337 1517  
[13] 1537 1766 1829 1955 1968 2129 2167 2177 2181 2193 2341 2431  
[25] 2452 2494 2582 2654 2893 3091 3351 3673 3708 3896 3901 3992  
[37] 3996 4061 4150 4303 4507 4508 4931 5075 5117 5126 5254 5318  
[49] 5416 5425 5476 5683 5903 5973 6041 6280 6281 6292 6348 6349  
[61] 6482 6504 6680 6844 6884 6887 7007 7029 7134 7184 7217 7354  
[73] 7438 7448 7462 7581 7594 7659 7691 7696 7750 7838 7844 8082  
[85] 8087 8101 8242 8442 8796 8921 8959 8999 9120 9142 9185 9226  
[97] 9399 9514 9548 9573 9822 9923 10083 10206 10425
```

Data consists of **10443** SNPs for 18 Pools

```
length(which(SNPvalues$SNPvalues > quantile(  
SNPvalues$SNPvalues, probs = 0.99, na.rm =  
TRUE)))
```

105 SNPs (logical, no?)

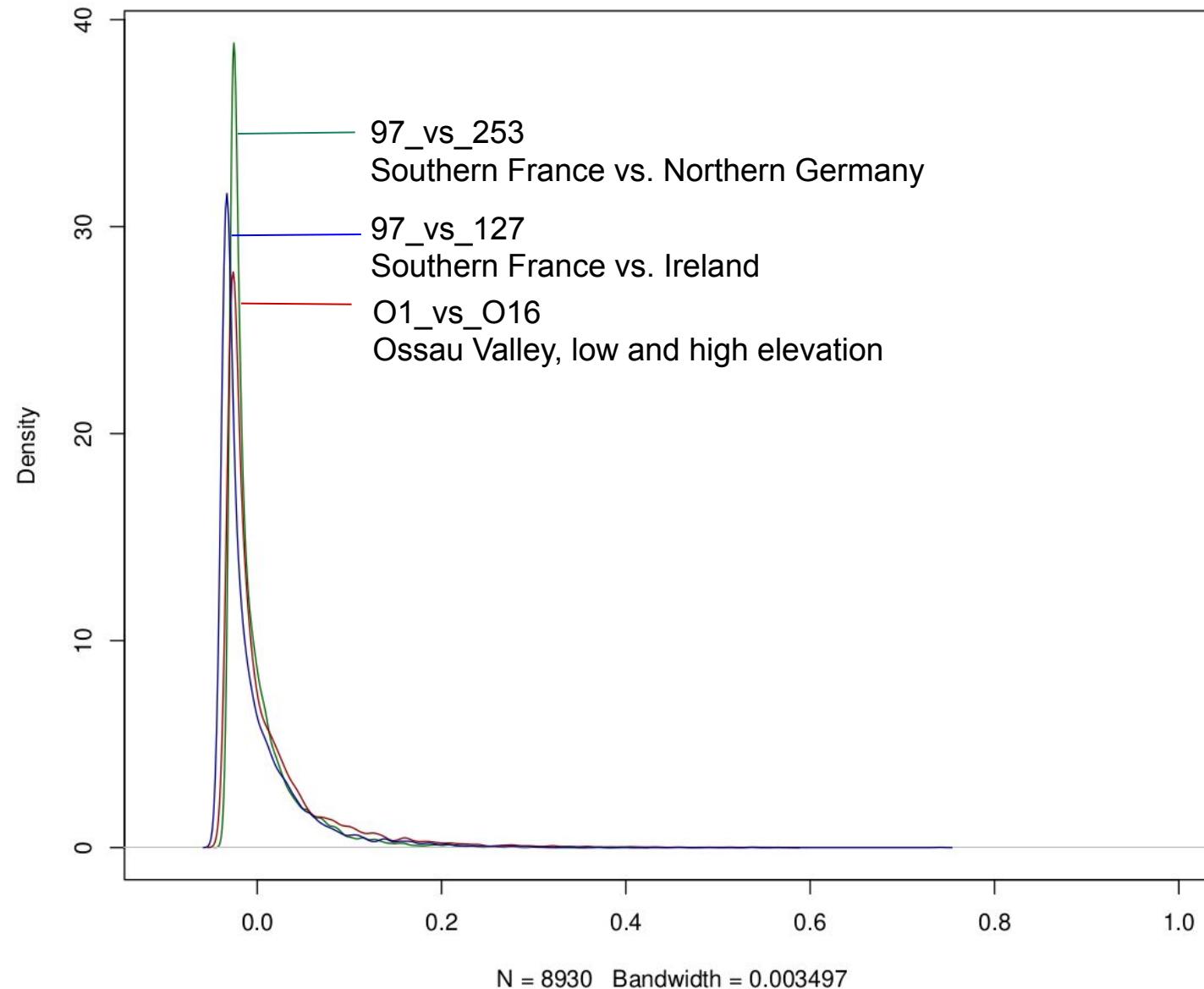
```
pooldata@snp.info[which(SNPvalues$SNPvalues > quantile(SNPvalues$SNPvalues, probs = 0.99, na.rm = TRUE)),]
```

Chromosome Position RefAllele AltAllele

	Chromosome	Position	RefAllele	AltAllele
27	Sc0000000	1986962	A	C
60	Sc0000000	3615264	A	G
74	Sc0000000	4659176	T	C
282	Sc0000004	2265588	T	C
300	Sc0000004	3816366	A	T
383	Sc0000006	1636275	A	T

... (continued)

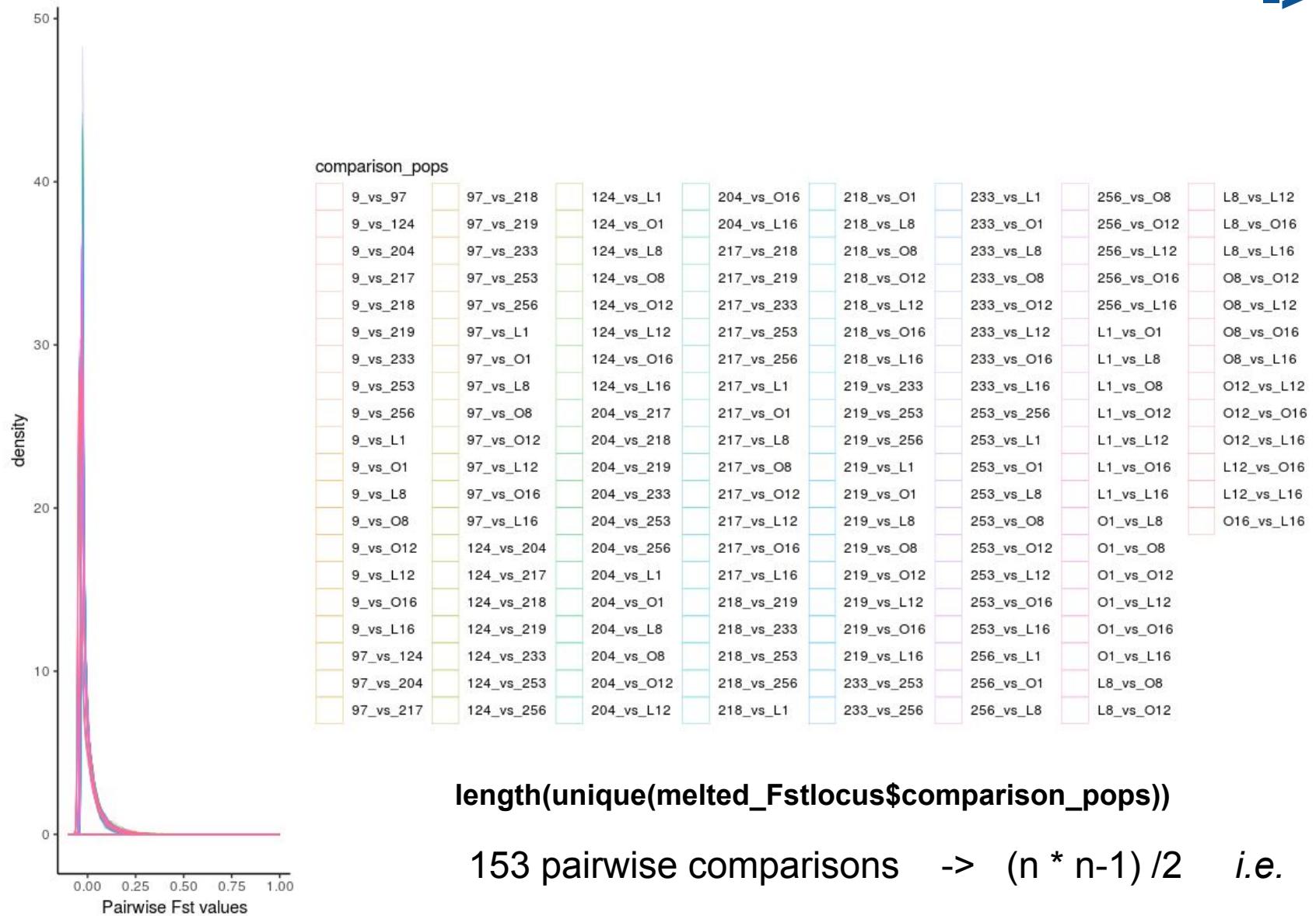
Empirical distribution of F_{ST} over pairs of pop -> Day4



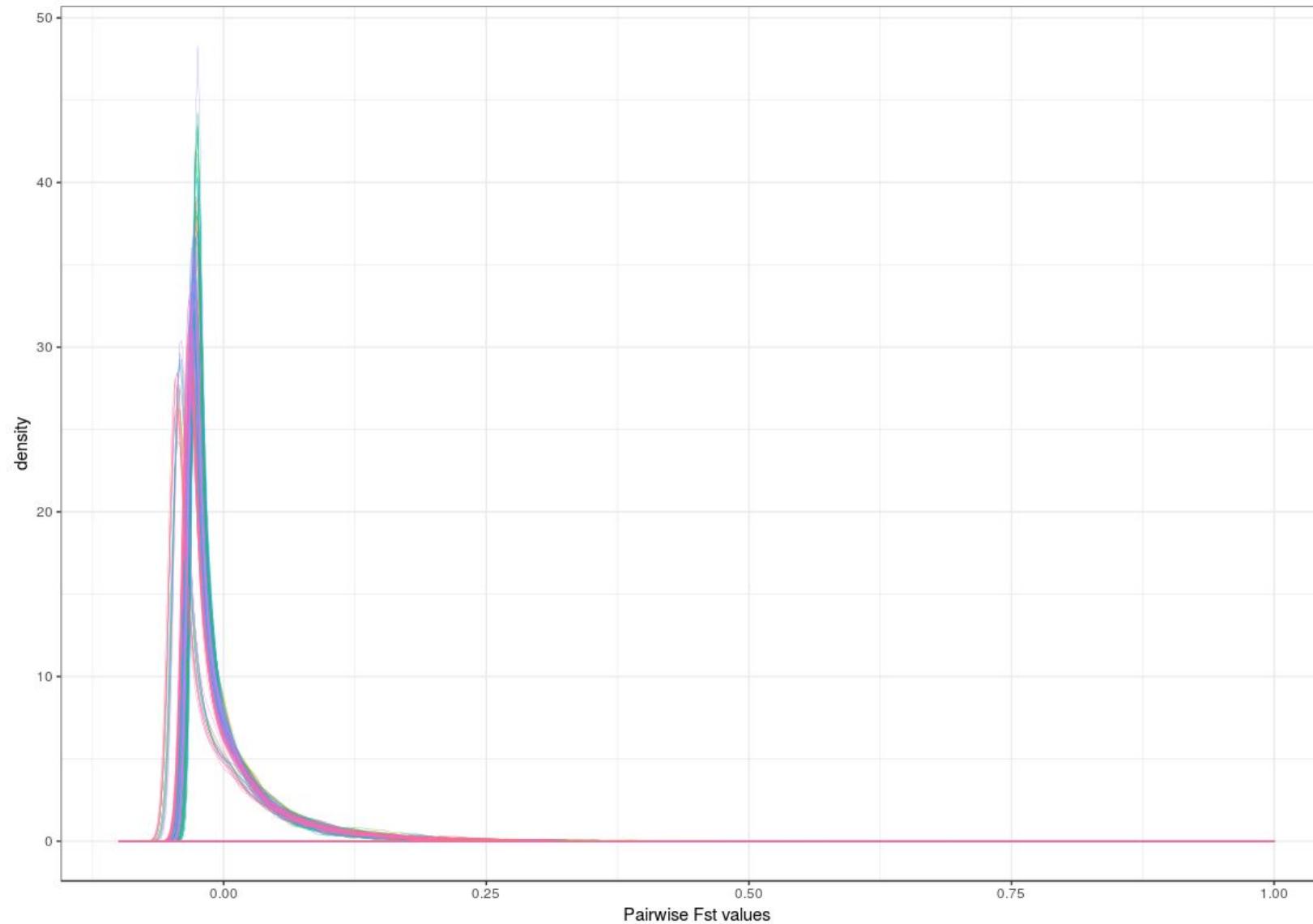
Near-zero F_{ST} at most loci,
even between populations at
quite different latitude

-> “Recent” post-glacial
recolonisation from Southern
refugia
+ large population sizes

Empirical distribution of F_{ST} over pairs of pop -> Day4



Empirical distribution of F_{ST} over pairs of pop -> Day4



Similar true for all pairs!

Long tail of distribution
(SNPs potentially under positive selection)

-> Low differentiation = excellent contrast

Top 1% F_{ST} outliers for some selected pairs

-> Day4

```
OUTLIERS_threshold_9_vs_124 <- quantile(melted_Fstlocus$Fstvalues[melted_Fstlocus$comparison_pops=="9_vs_124"], probs = 0.99, na.rm = TRUE)
nrow(pooldata@snp.info[which(melted_Fstlocus$Fstvalues[melted_Fstlocus$comparison_pops=="9_vs_124"]>=OUTLIERS_threshold_9_vs_124),])
-> 88
```

```
OUTLIERS_threshold_9_vs_97 <- quantile(melted_Fstlocus$Fstvalues[melted_Fstlocus$comparison_pops=="9_vs_97"], probs = 0.99, na.rm = TRUE)
nrow(pooldata@snp.info[which(melted_Fstlocus$Fstvalues[melted_Fstlocus$comparison_pops=="9_vs_97"]>=OUTLIERS_threshold_9_vs_97),])
-> 88
```

```
OUTLIERS_threshold_O1_vs_O16 <- quantile(melted_Fstlocus$Fstvalues[melted_Fstlocus$comparison_pops=="O1_vs_O16"], probs = 0.99, na.rm = TRUE)
nrow(pooldata@snp.info[which(melted_Fstlocus$Fstvalues[melted_Fstlocus$comparison_pops=="O1_vs_O16"]>OUTLIERS_threshold_O1_vs_O16),])
-> 88
```

Same number of outliers for the two comparisons ! Why?

Because we **similarly select the top 1% of SNPs with the highest F_{ST} values**, corresponding to a given number of SNPs. By doing so, we implicitly assume that 1% of the genome is indeed under positive selection! And we have no idea of this proportion in reality! Could be more, could be less, could be even zero!

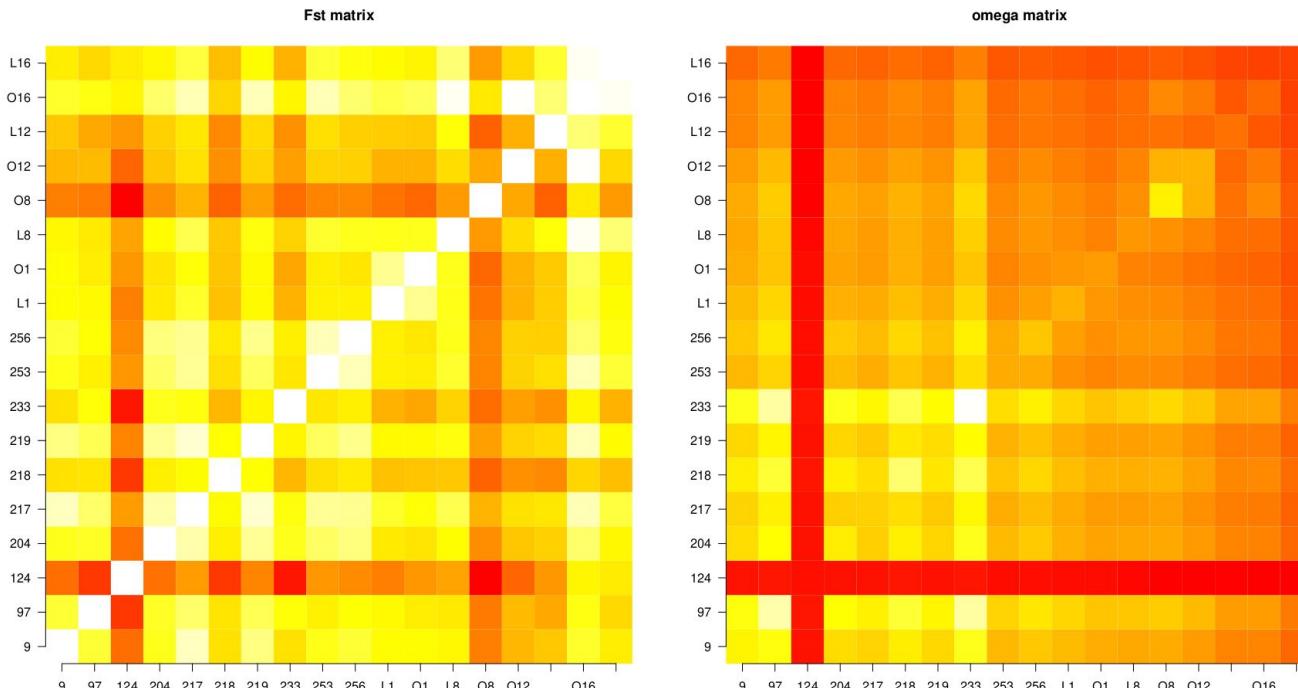
BayPass inference - omega matrix

OPTIONAL: try to generate a single plot containing the results of both the omega matrix and the F_{ST} matrix to check whether the results are consistent

-> Day2

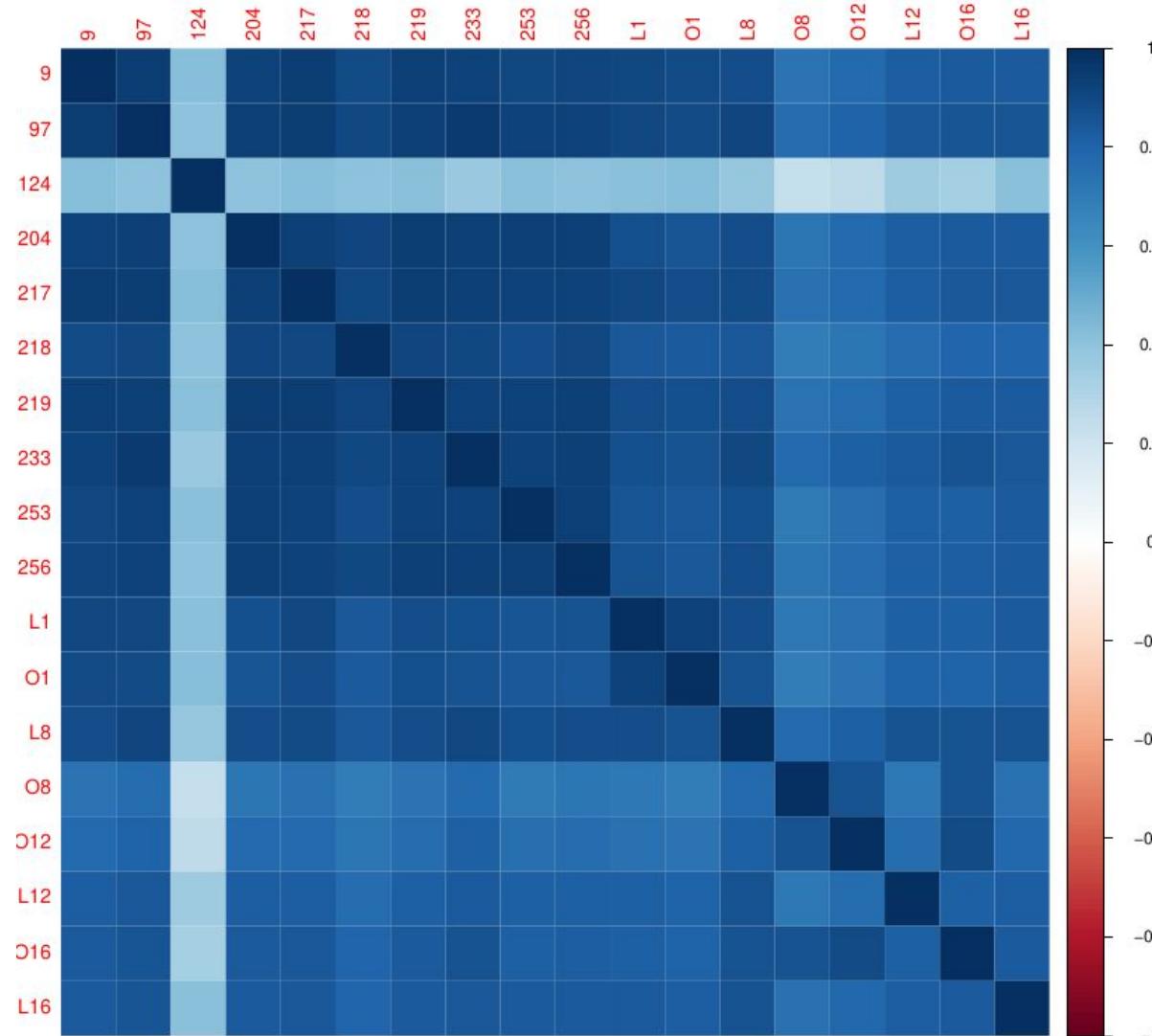
```
par(mfrow=c(2,1))
```

```
image(Fstmatrix, axes=FALSE,main="Fst matrix",col=rev(heat.colors(200)))
axis(1, at = seq(0, 1, length = nrow(Fstmatrix)), labels = rownames(Fstmatrix))
axis(2, at = seq(0, 1, length = ncol(Fstmatrix)), labels = colnames(Fstmatrix),las=2)
image(omega_matrix,axes=FALSE,main="omega matrix",col=heat.colors(200))
axis(1, at = seq(0, 1, length = nrow(Fstmatrix)), labels = rownames(Fstmatrix))
axis(2, at = seq(0, 1, length = ncol(Fstmatrix)), labels = colnames(Fstmatrix),las=2)
```

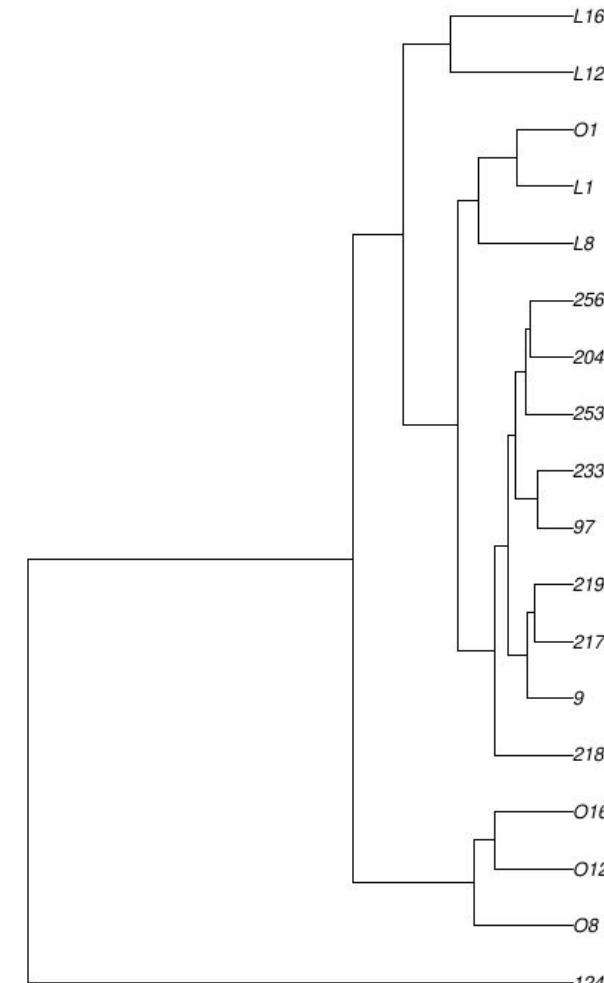


BayPass inference - omega matrix

Correlation matrix

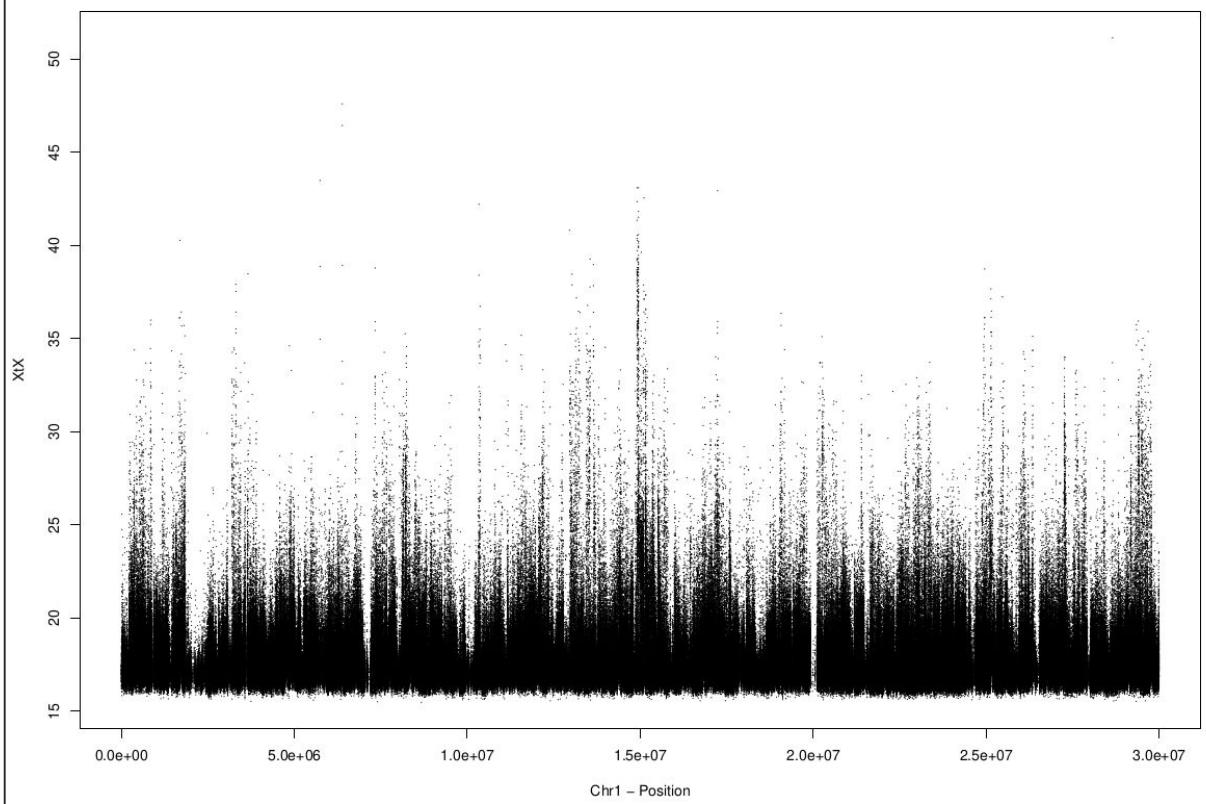
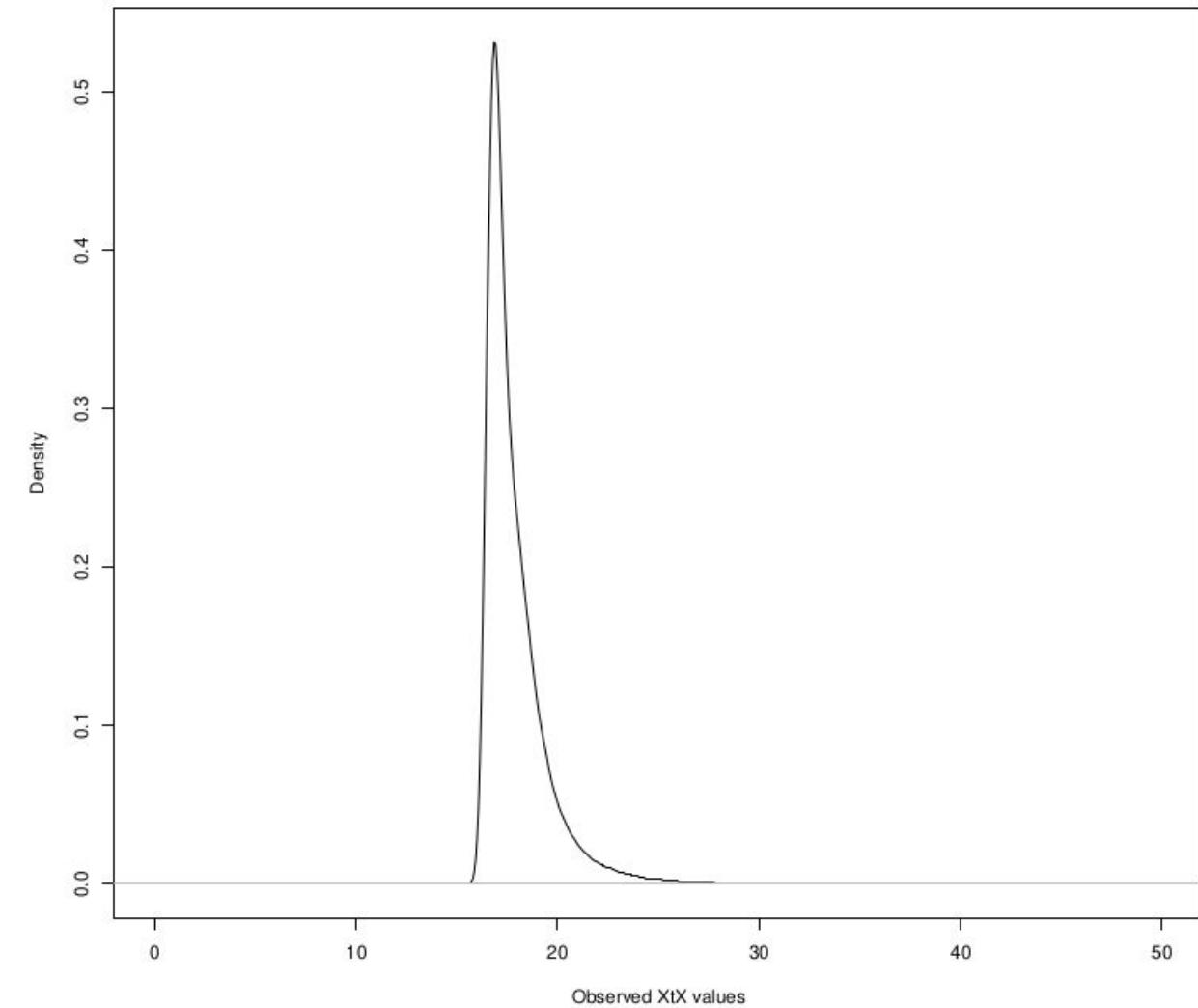


-> Day2



BayPass inference - XtX

-> Day4



BayPass - Neutral simulations & calibration

-> Day4

```
xtx.pods=read.table("simulated PODS_XtX.txt",h=T)
xtx.threshold=quantile(xtx.pods$M_XtX,probs=c(0.999,0.9999,1))
```

```
length(XtX_BF$XtX[XtX_BF$XtX>xtx.threshold[1]])
[1] 28150
```

```
length(XtX_BF$XtX[XtX_BF$XtX>xtx.threshold[2]])
[1] 7113
```

```
> length(XtX_BF$XtX[XtX_BF$XtX>xtx.threshold[3]])
[1] 3090
```

```
length(XtX_BF$XtX)
1349416
```

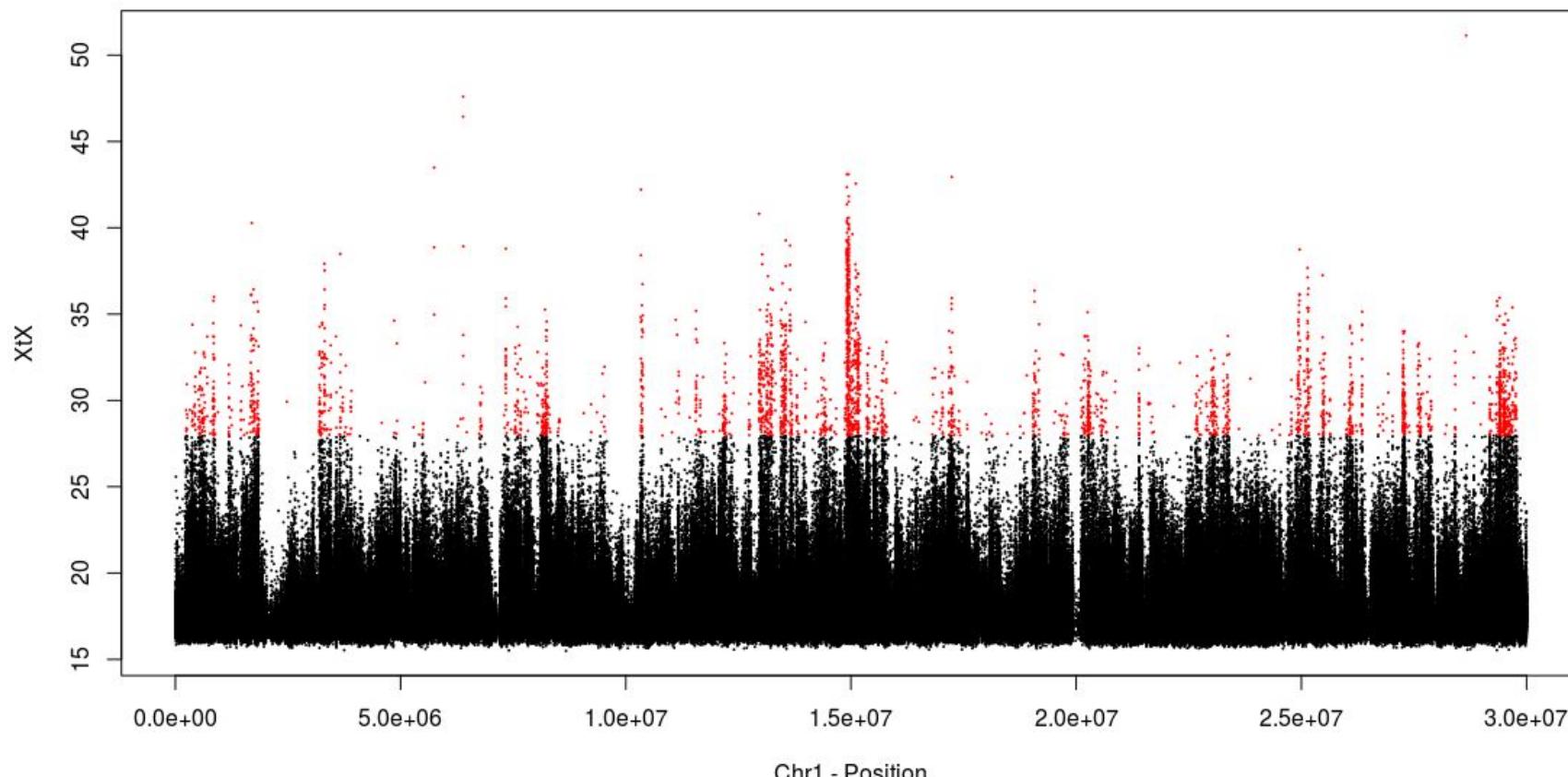
```
3090/1349416
=> 0.002289879 (0.23% of SNPs exhibit higher levels of XtX than observed under the neutral simulations)
```

————> Quite ≠ from the previously assumed 1% !

BayPass - XtX scan (outliers, after calibration)

-> Day4

```
plot(XtX_BF$XtX~XtX_BF$position,cex=0.05,pch=20,xlab="Chr1 - Position",ylab="XtX",col=ifelse(XtX_BF$XtX>=xtx.threshold[3],"red","black"))
```

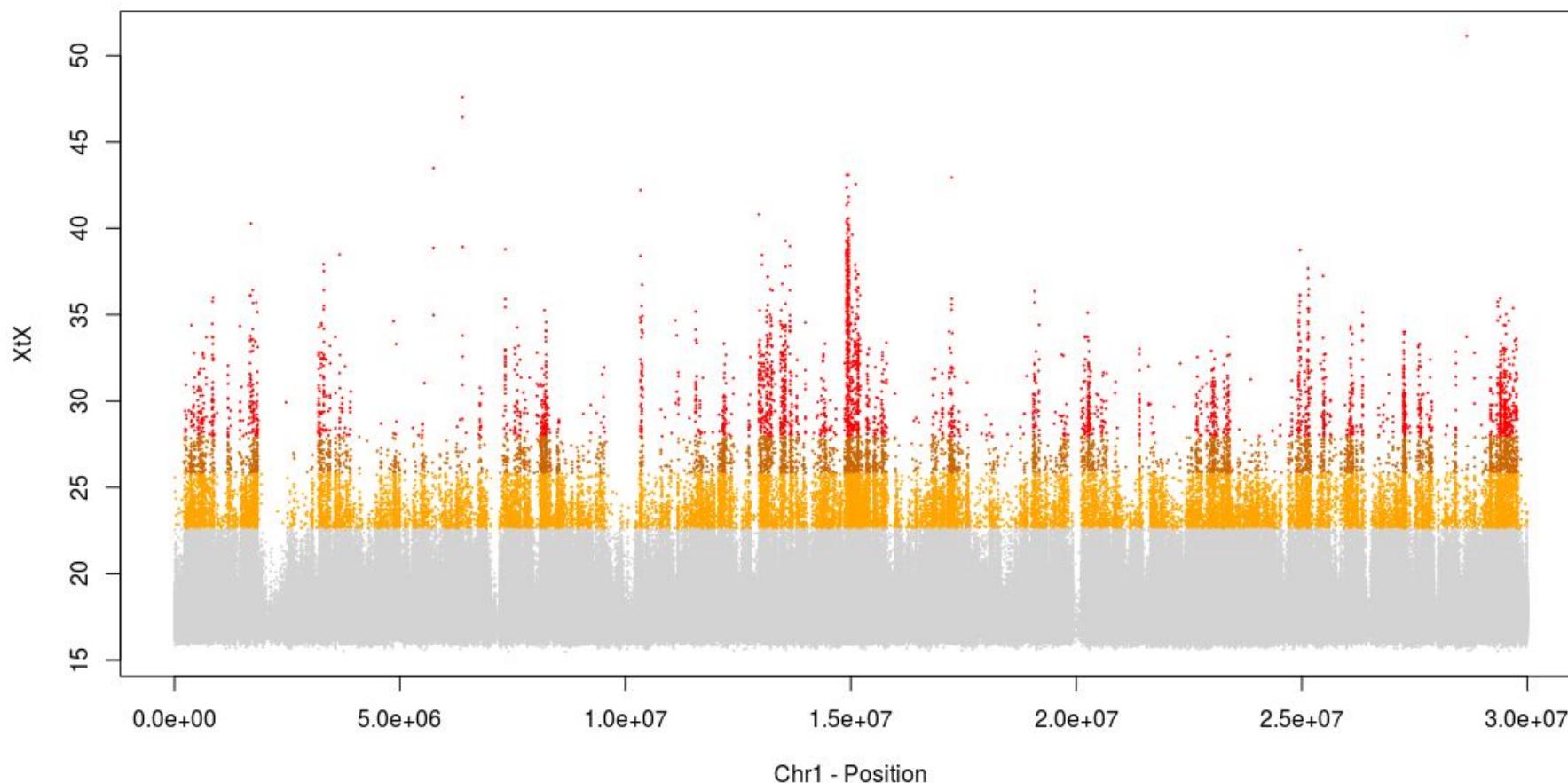


BayPass - XtX scan (outliers, after calibration)

Step5, option 1

-> Day4

```
plot(XtX_BF$XtX~XtX_BF$position,cex=0.05,pch=20,xlab="Chr1 -  
Position",ylab="XtX",col=ifelse(XtX_BF$XtX>=xtx.threshold[3],"red",ifelse(XtX_BF$XtX>=xtx.threshold[2],"darkorange3",ifelse(XtX_BF$XtX>=  
xtx.threshold[1],"orange","lightgrey")))
```



BayPass - XtX - Sliding window approach

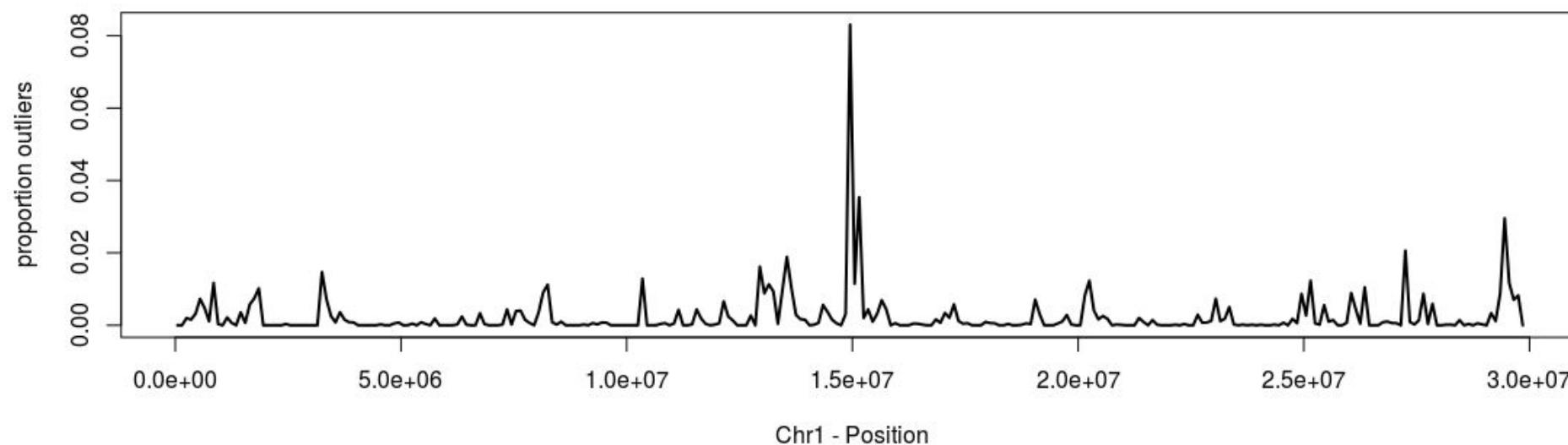
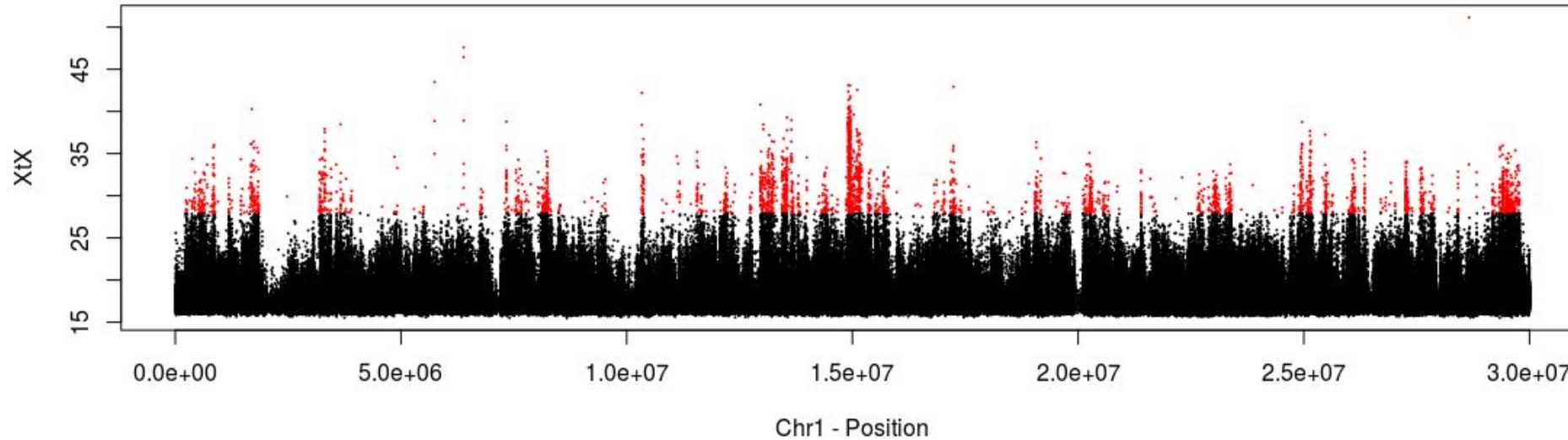
Step5, option 2

```
TablePropOutlier=NULL
# Sliding windows:
lengthsequence=30000000
starts_win <- seq(1, lengthsequence-100000, by = 100000)
n <- length(starts_win)
for (i in 1:n) {
  minregion=starts_win[i]
  maxregion=starts_win[i]+99999
  positionmedian=minregion+(maxregion-minregion)/2
  xtxregion <- subset(XtX_BF, (position >= minregion & position <= maxregion) )
  nb_SNP=nrow(xtxregion)
  outlierthresholdmax<- length(xtxregion$XtX[xtxregion$XtX>=xtx.threshold[3]])
  prop_outlierthresholdmax=outlierthresholdmax/ nb_SNP

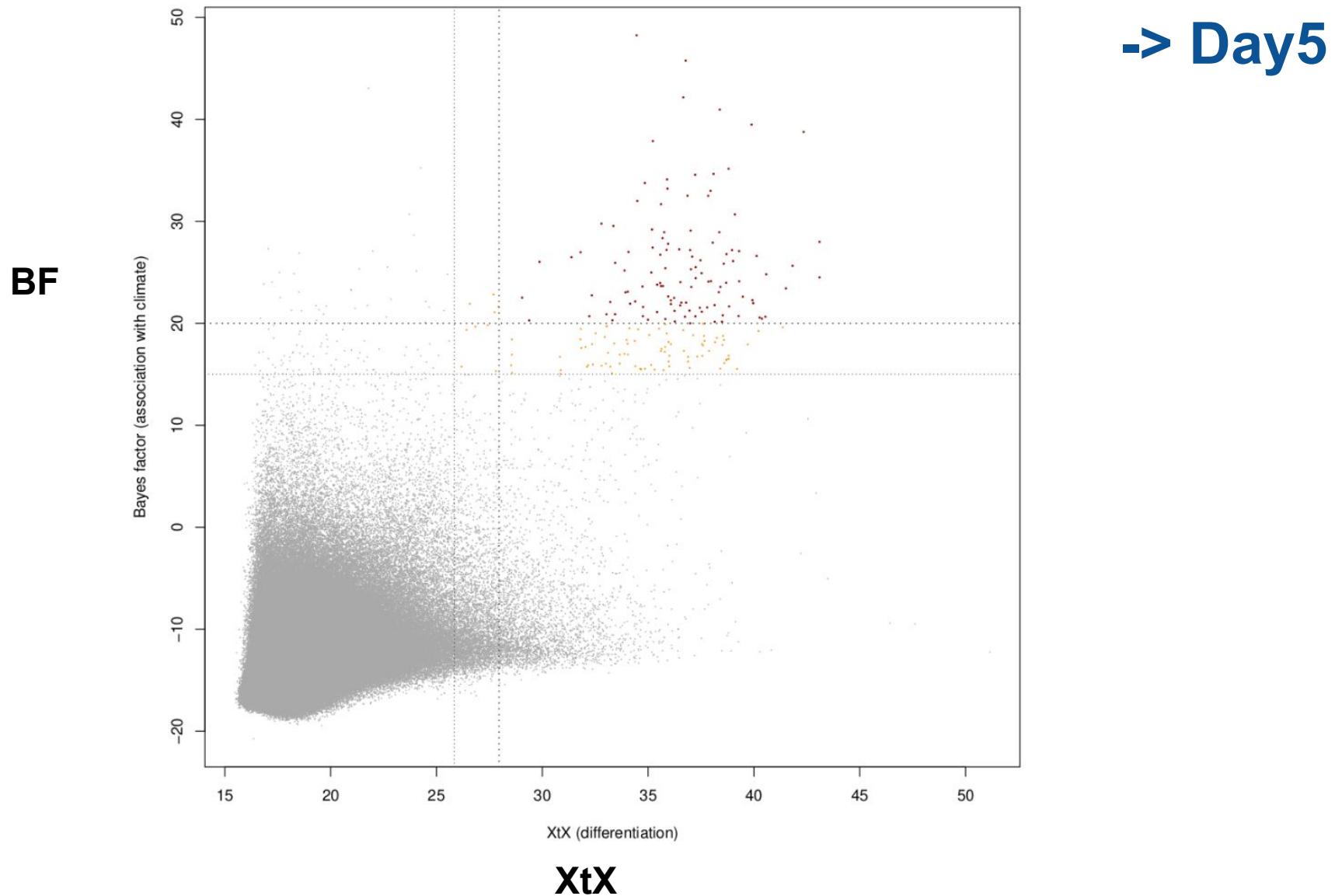
  TablePropOutlier=rbind(TablePropOutlier,cbind(positionmedian,starts_win[i],starts_win[i]+99999,outlierthresholdmax,nb_SNP,prop_outlierthresholdmax))
}
colnames(TablePropOutlier)<-
c("position_median","start_windows","end_windows","outliers_thresholdmax","nb_SNPs_windows","proportion_outliers")
TablePropOutlier<- as.data.frame(TablePropOutlier)
```

BayPass - XtX - Sliding window approach

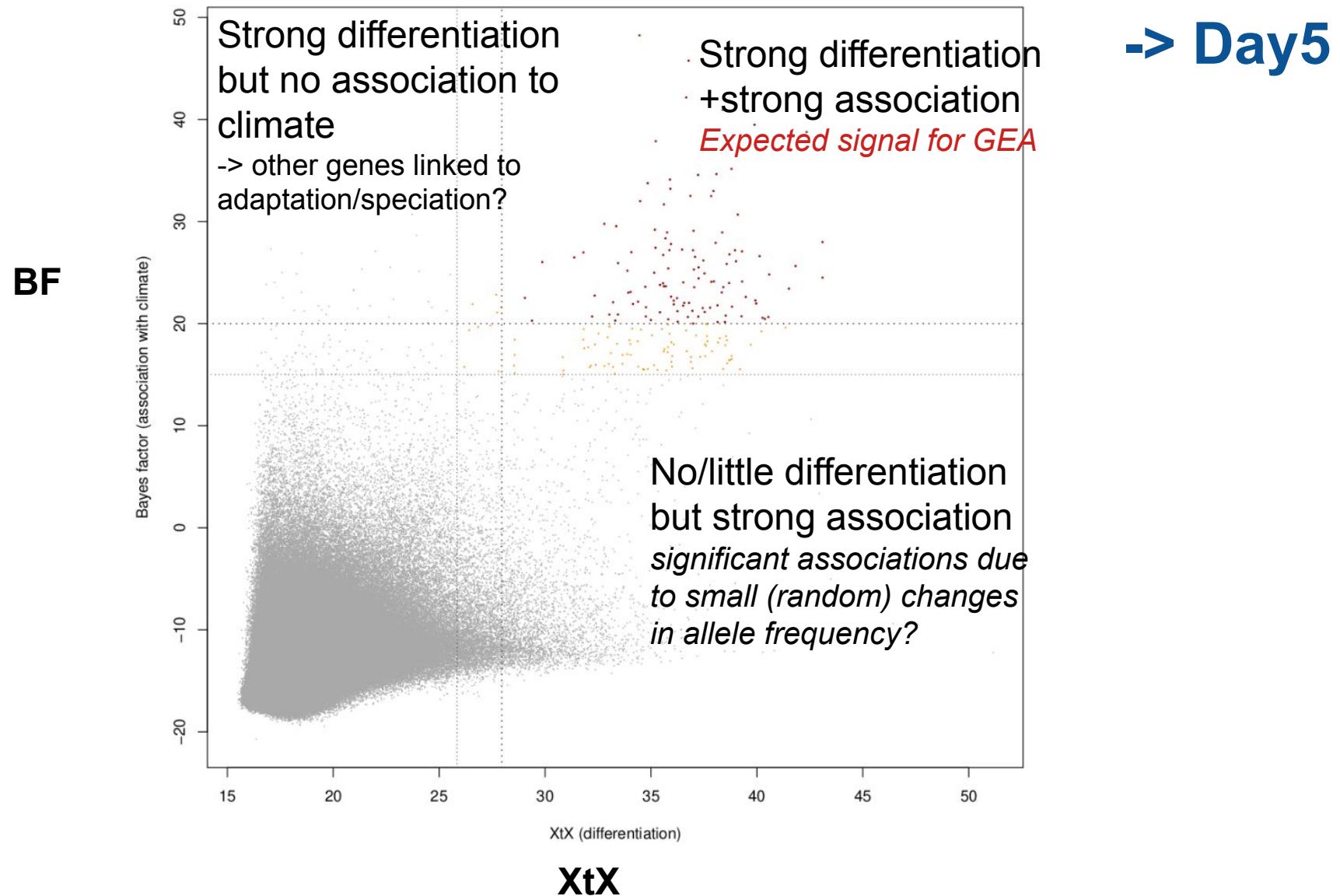
Step5, option 2



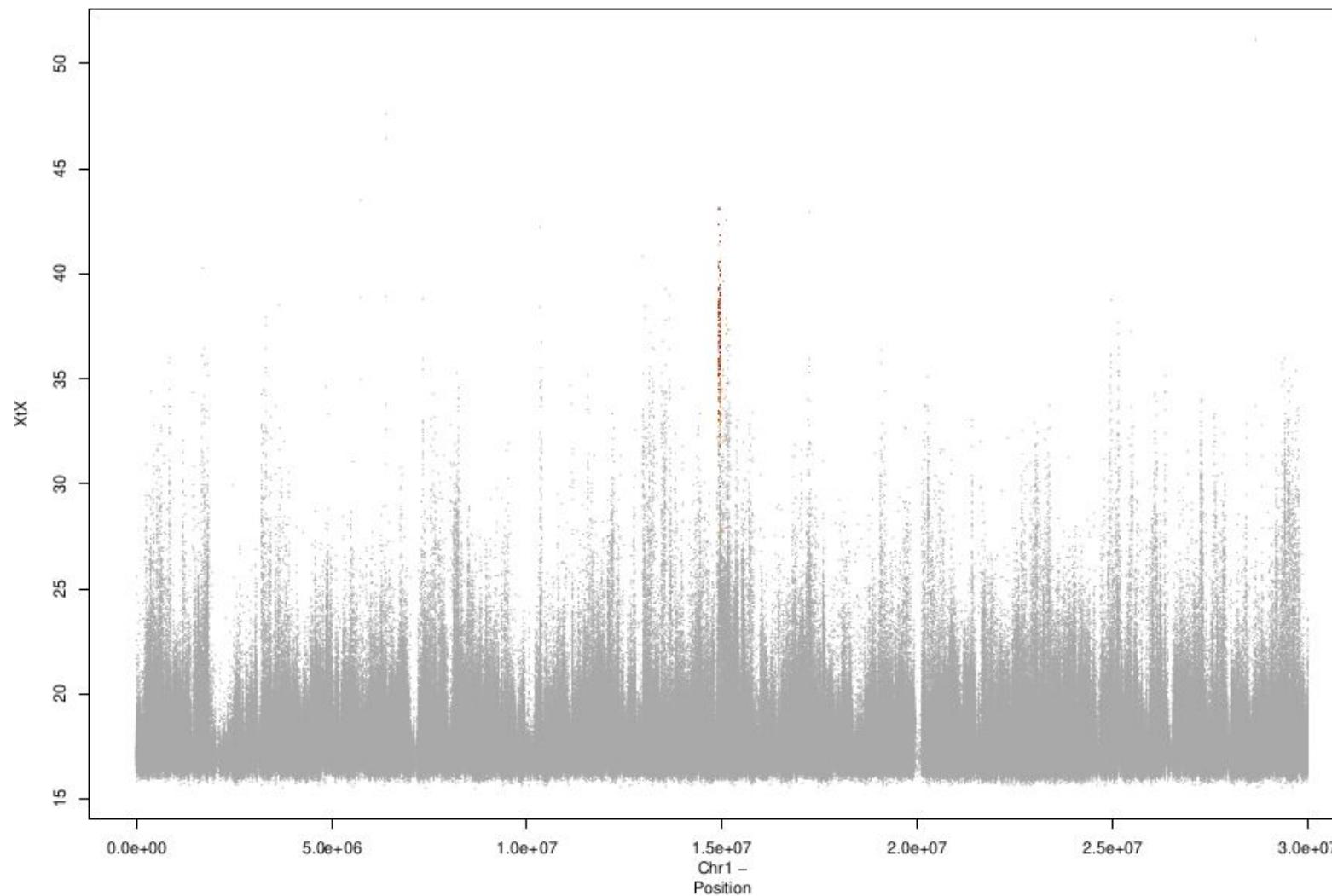
BayPass - combining XtX and GEA !



BayPass - combining XtX and GEA !

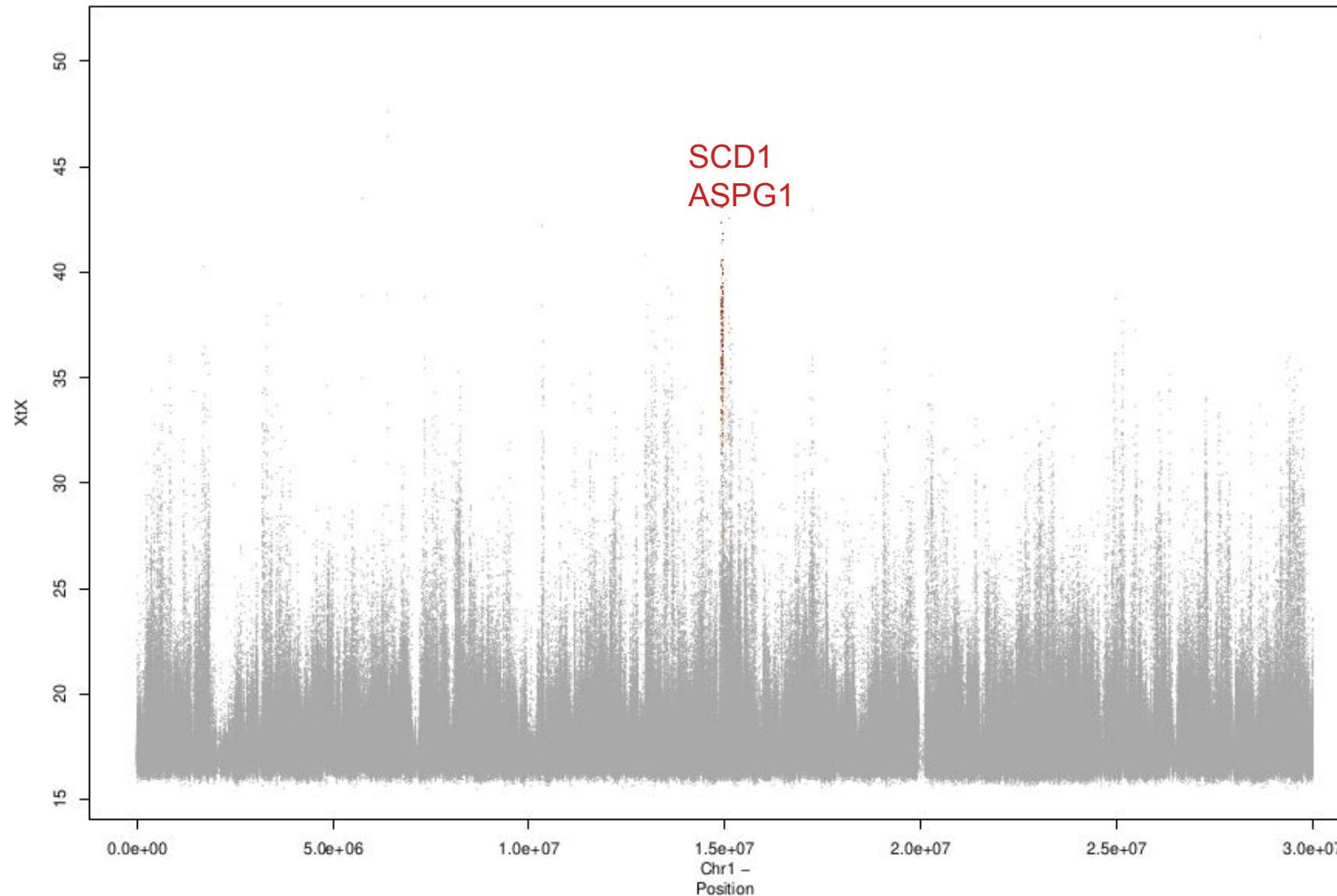


BayPass - XtX + GEA to reveal local adaptation!



-> Day5

BayPass - XtX + GEA to reveal local adaptation!



-> Day5

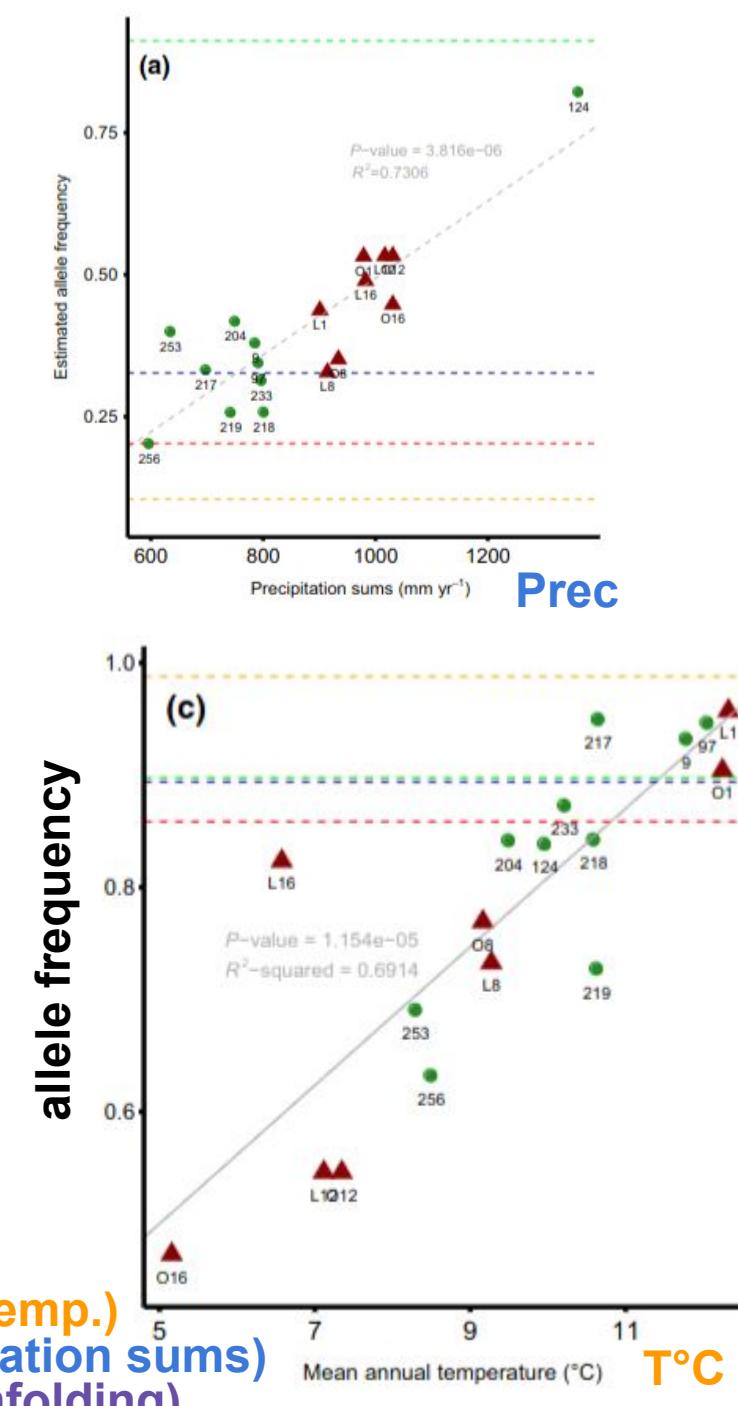
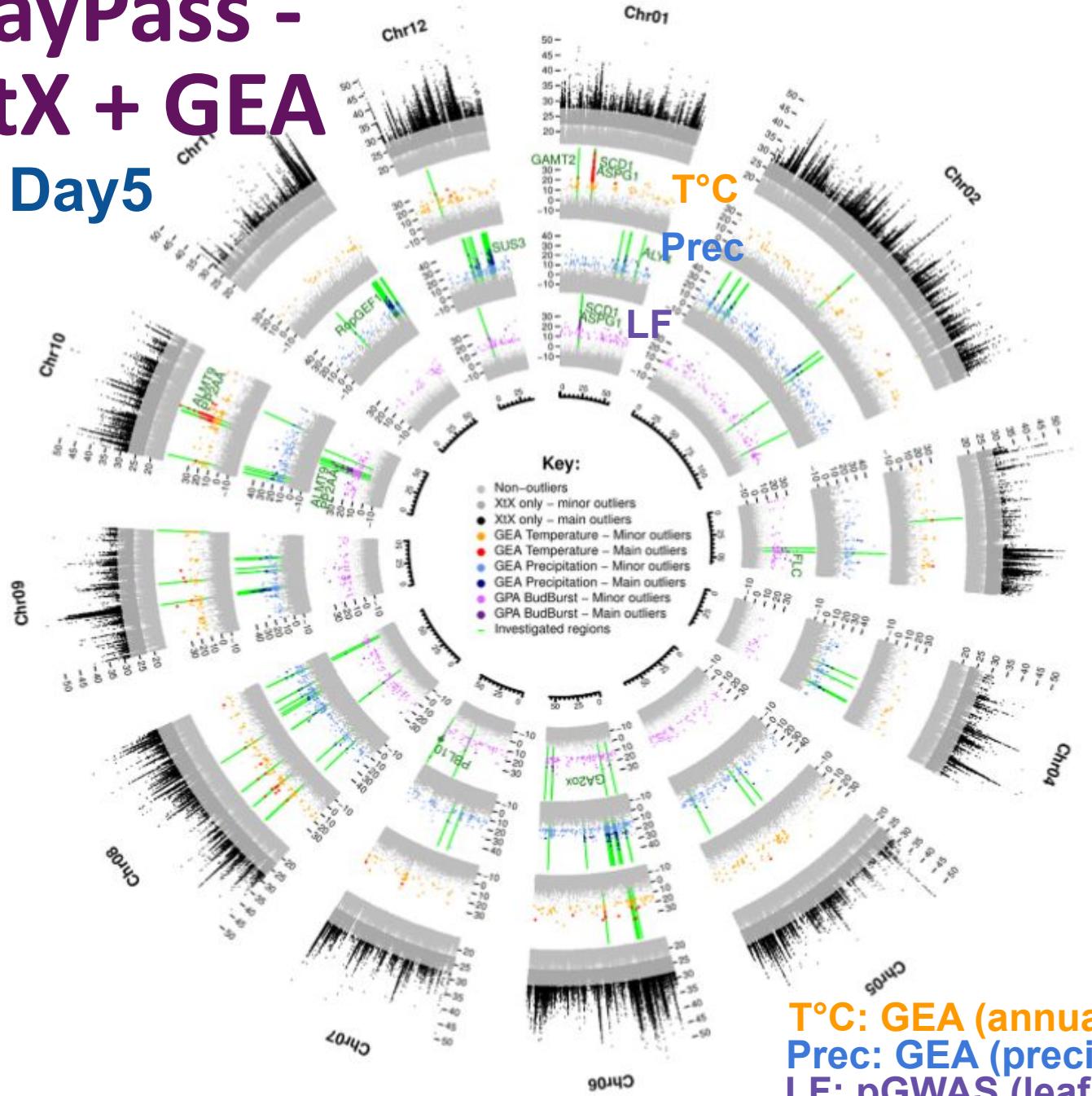
2 candidate genes:

SCD1 is known to be involved in stomatal responses to water stress, making it a plausible candidate for temperature-related associations.

ASPG1 plays a key role in the production of gibberellins, which are crucial for various plant developmental processes, including seed dormancy.

-> determining the optimal timing of seed germination and, consequently, plant survival.

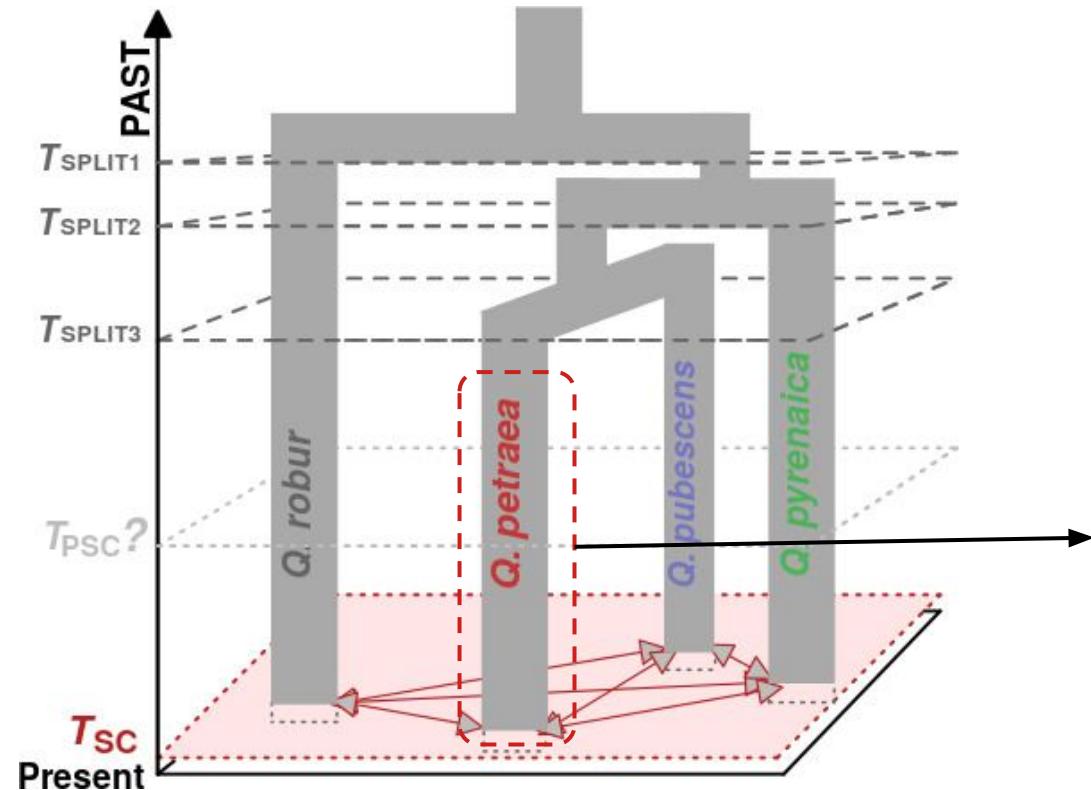
BayPass - XtX + GEA -> Day5



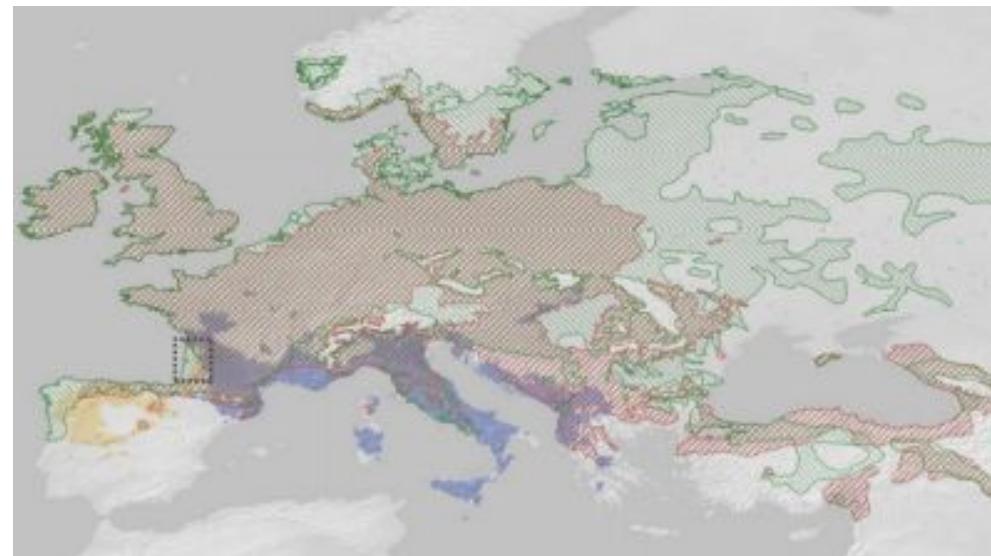
Back to the context: European white oaks

Demographic history

-> Day3



Distribution range of *Quercus petraea* (red) covers very different latitudes



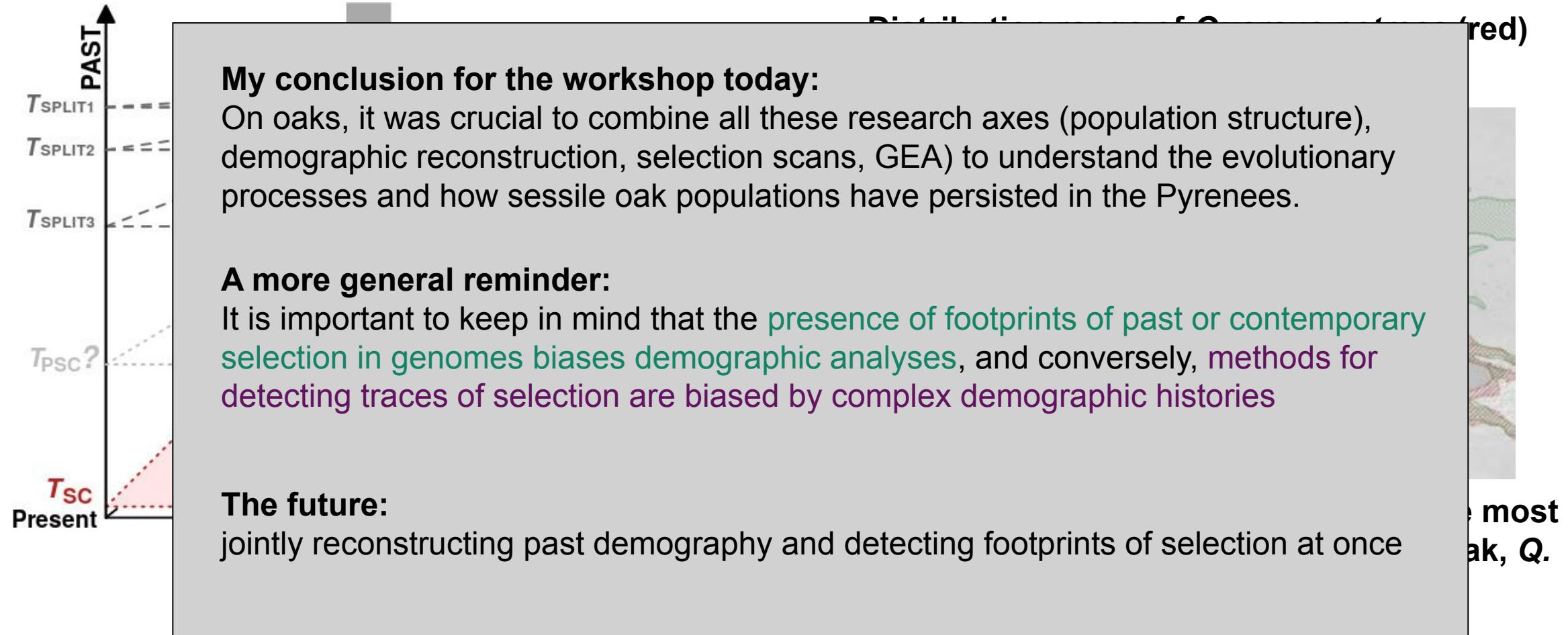
-> adaptive allele transferred from the most cold-adapted species (pedunculate oak, *Q. robur*) to sessile oak (*Q. petraea*)

i.e. adaptive introgression

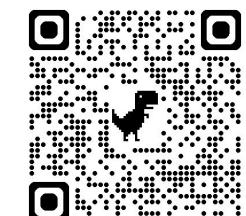
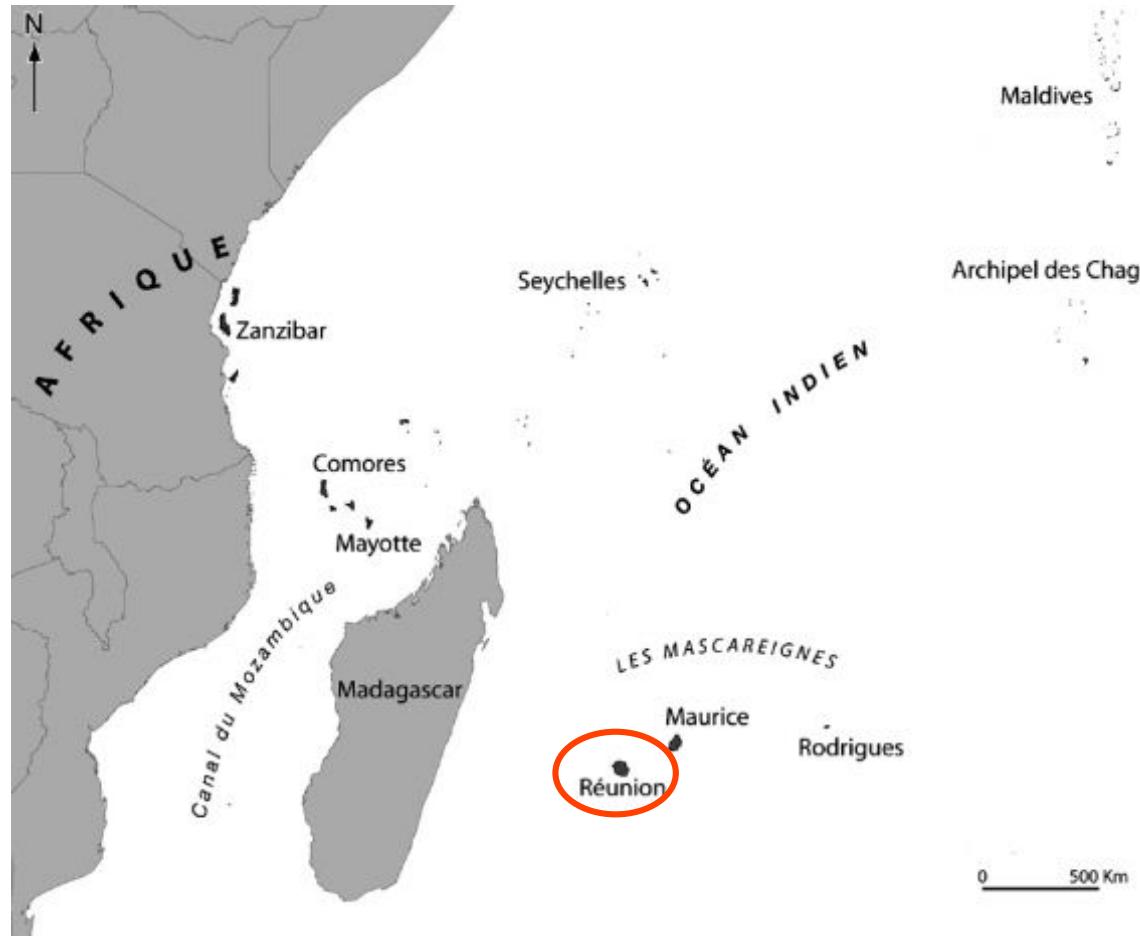
Back to the context: European white oaks

Demographic history

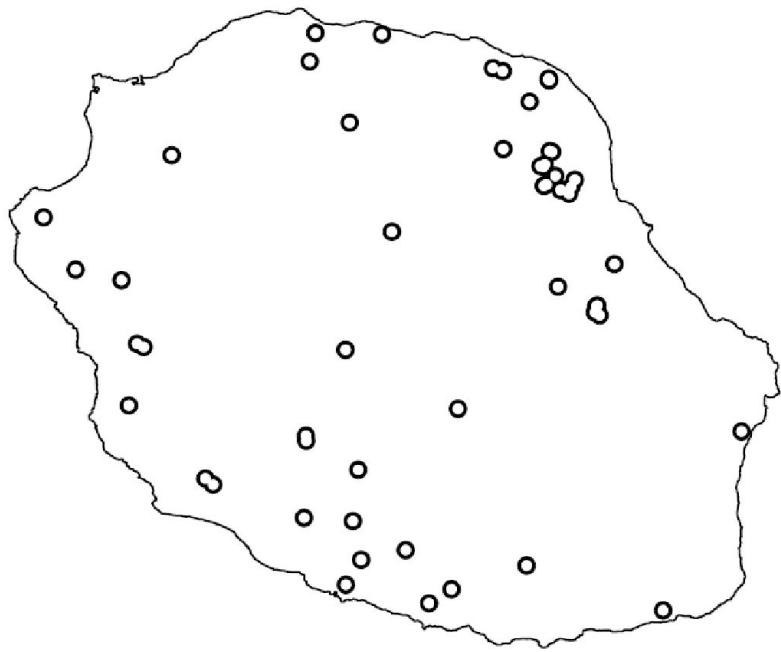
-> Day3



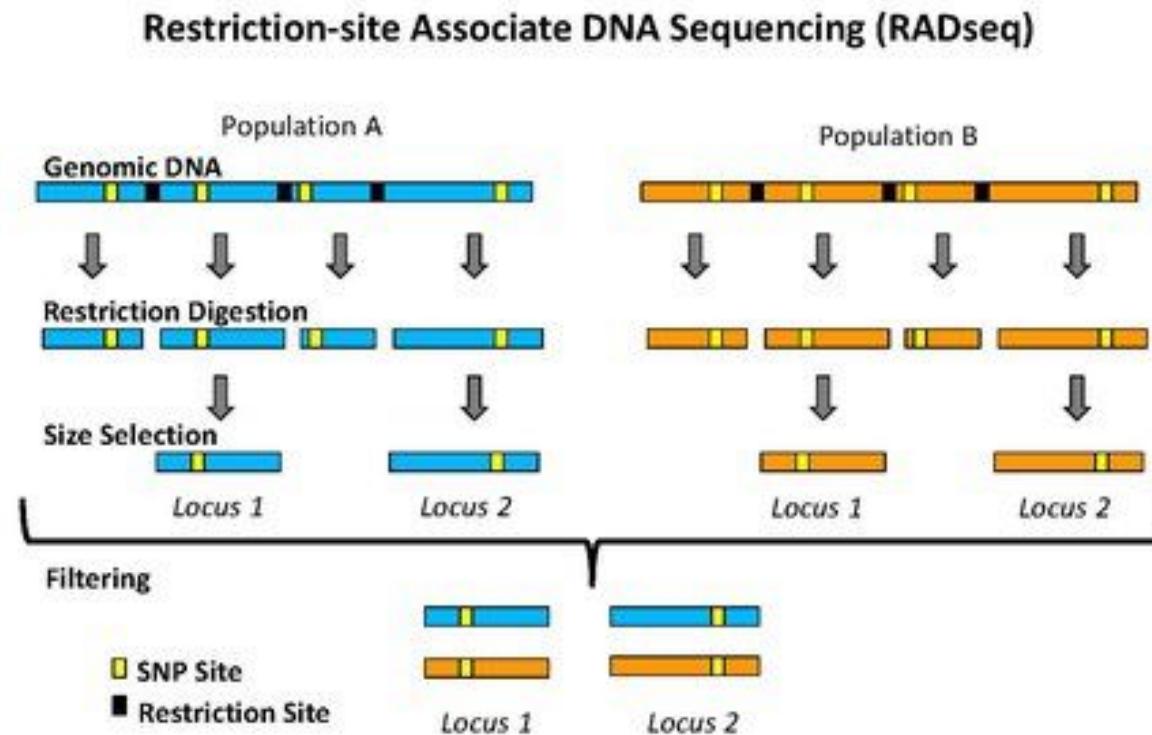
An application: The Réunion harrier (*Circus maillardi*)



Sampling and sequencing techniques

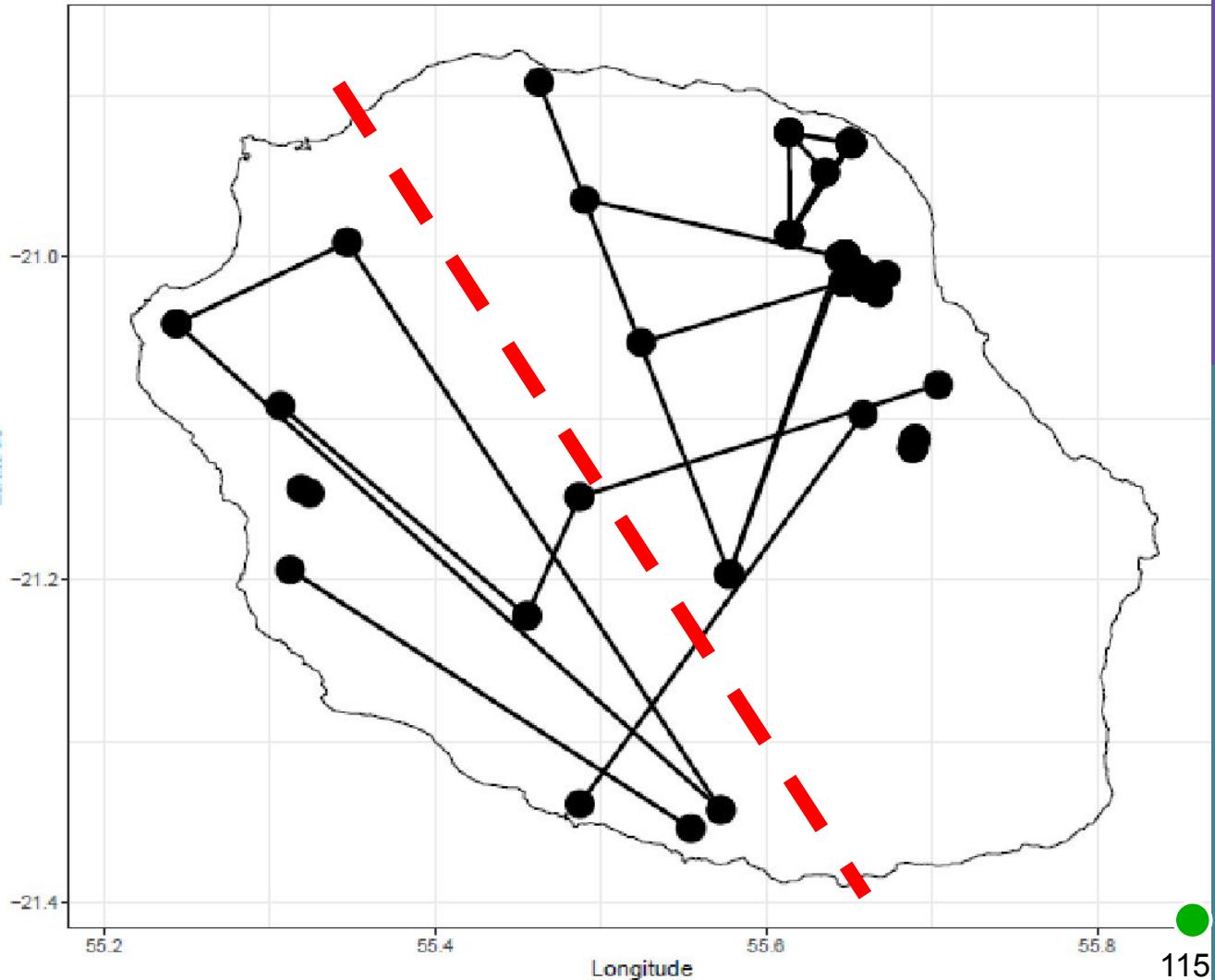


- GBS (*Genotyping by Sequencing*)
- Sequences loci around restriction sites > less sequencing effort
- About 6000 Single Nucleotide Polymorphisms.

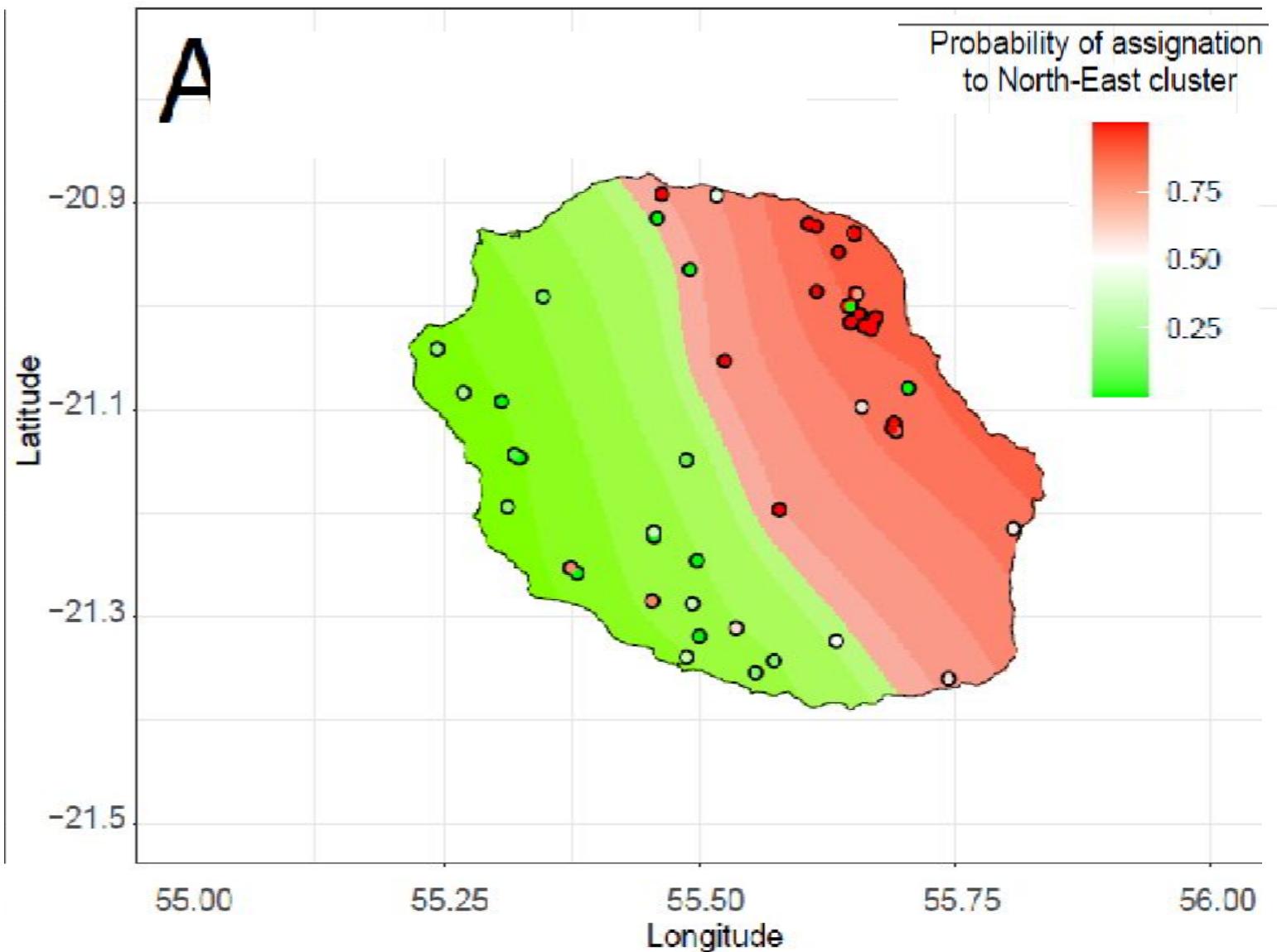


Relatedness and recent dispersal

- Pairs of related individuals (first et second degree, e.g. parent-offspring, half-siblings) follow a NW-SE axis
- Geographical barriers to migration are still extant

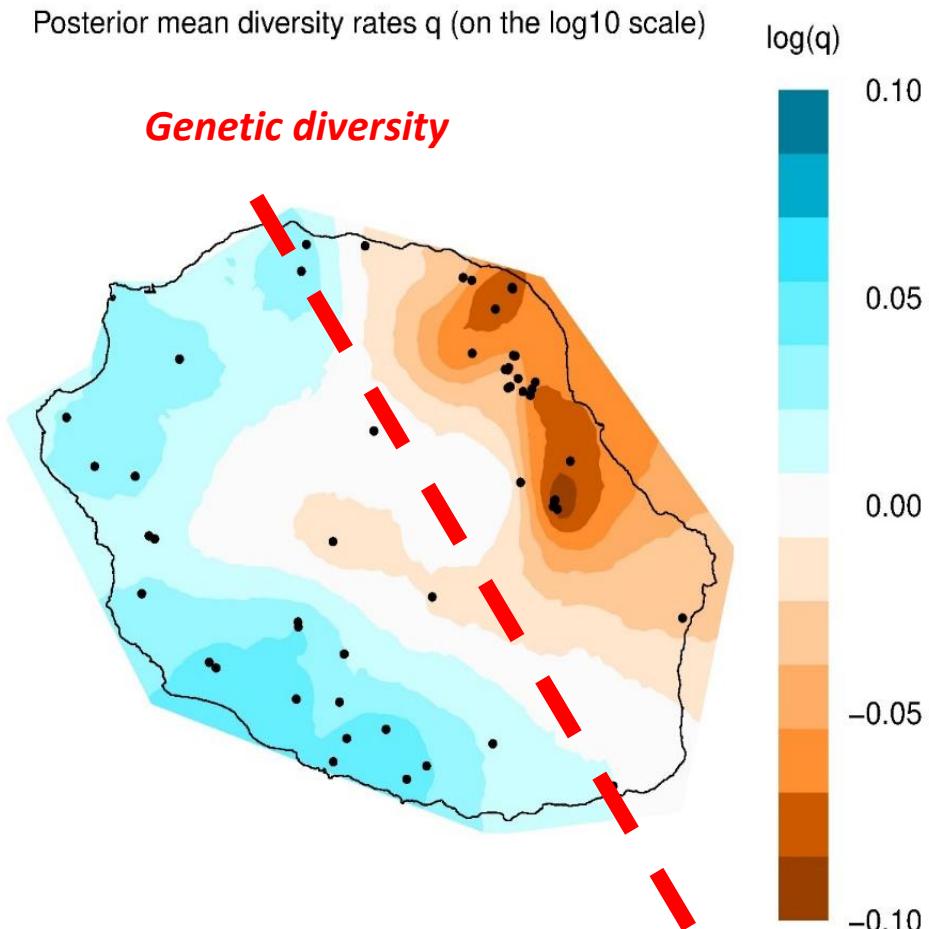
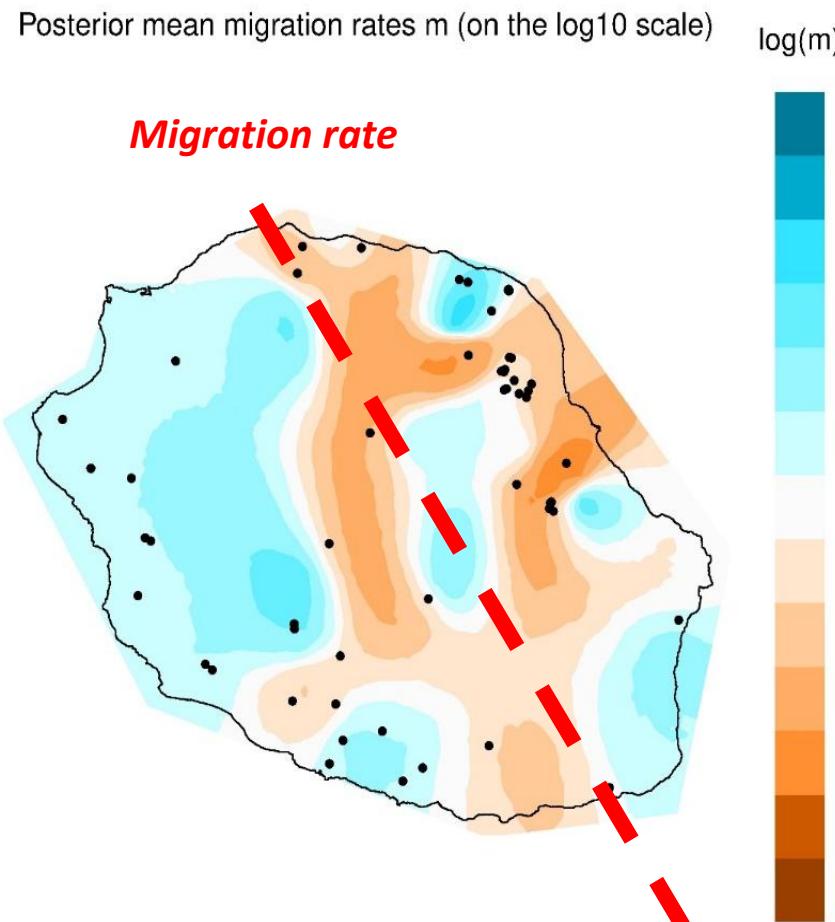


Population structure

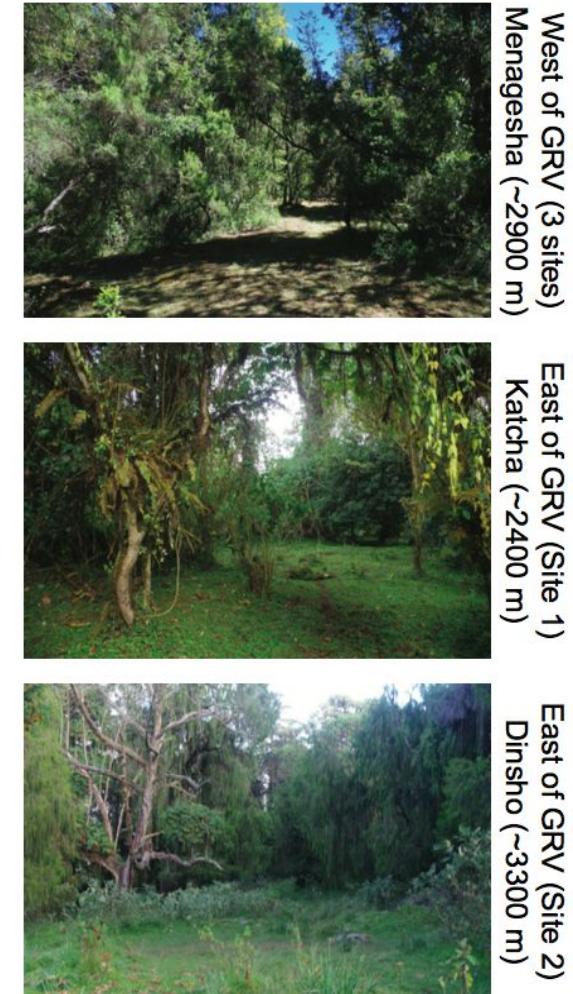
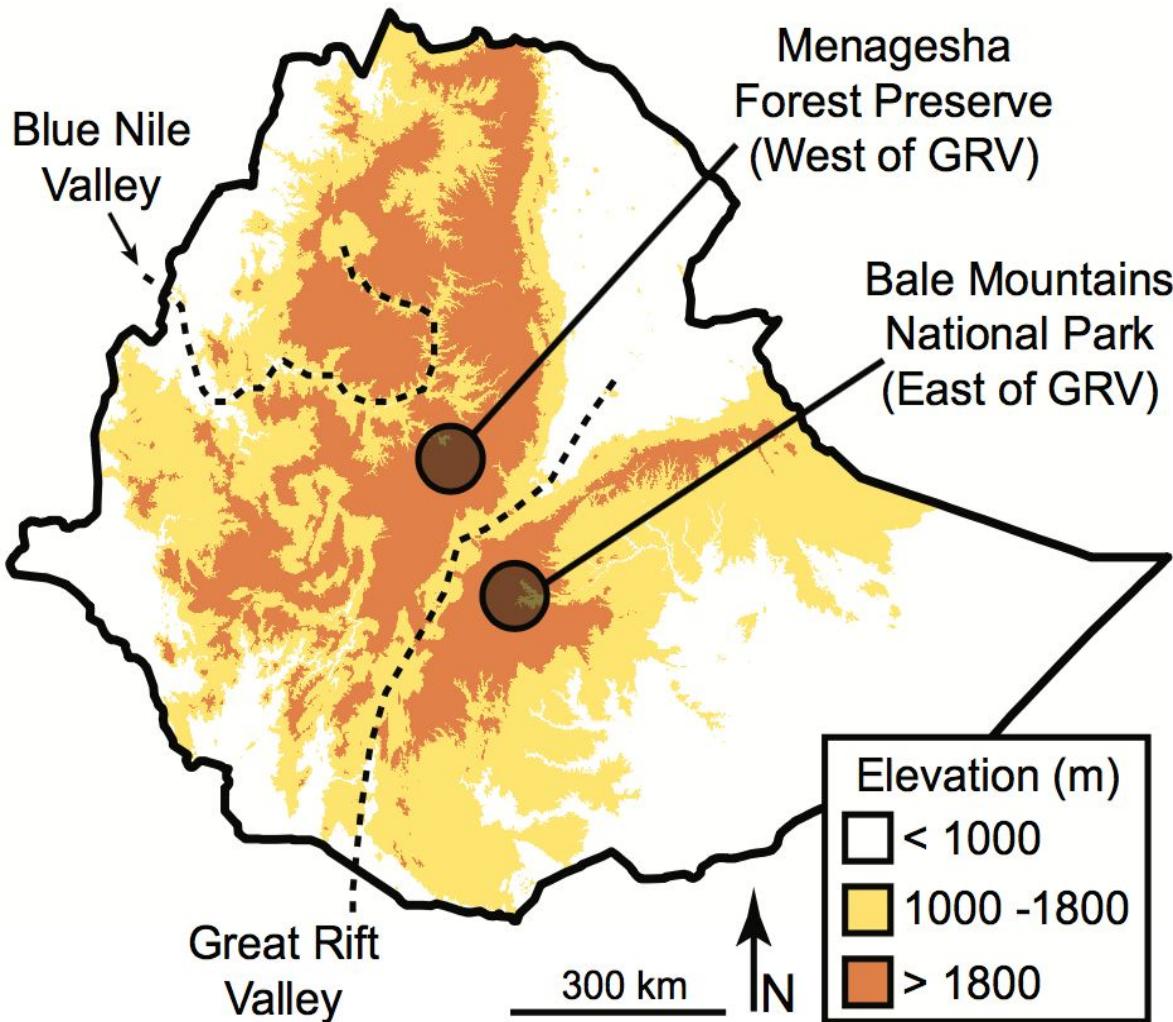


Population structure

This is useful when studying waves of expansion or demography in space



Spatializing structure analyses



Spatializing structure analyses

