Physalia

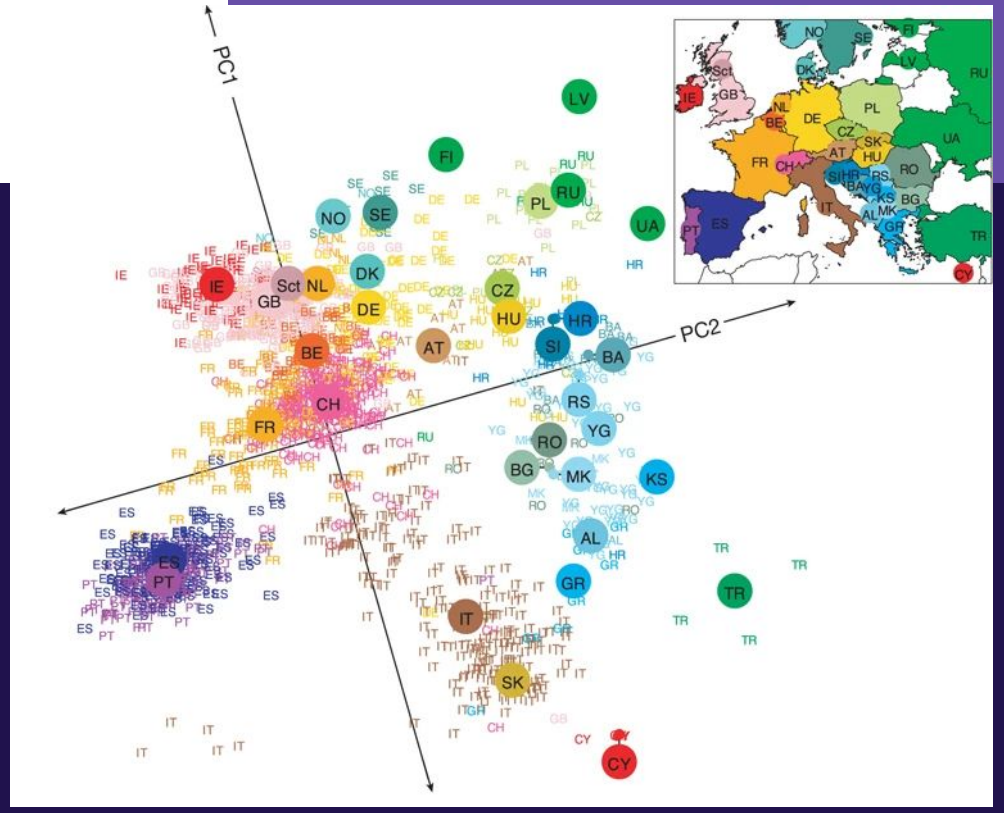
Courses

Landscape genomics

29/11/2024

Physalia course

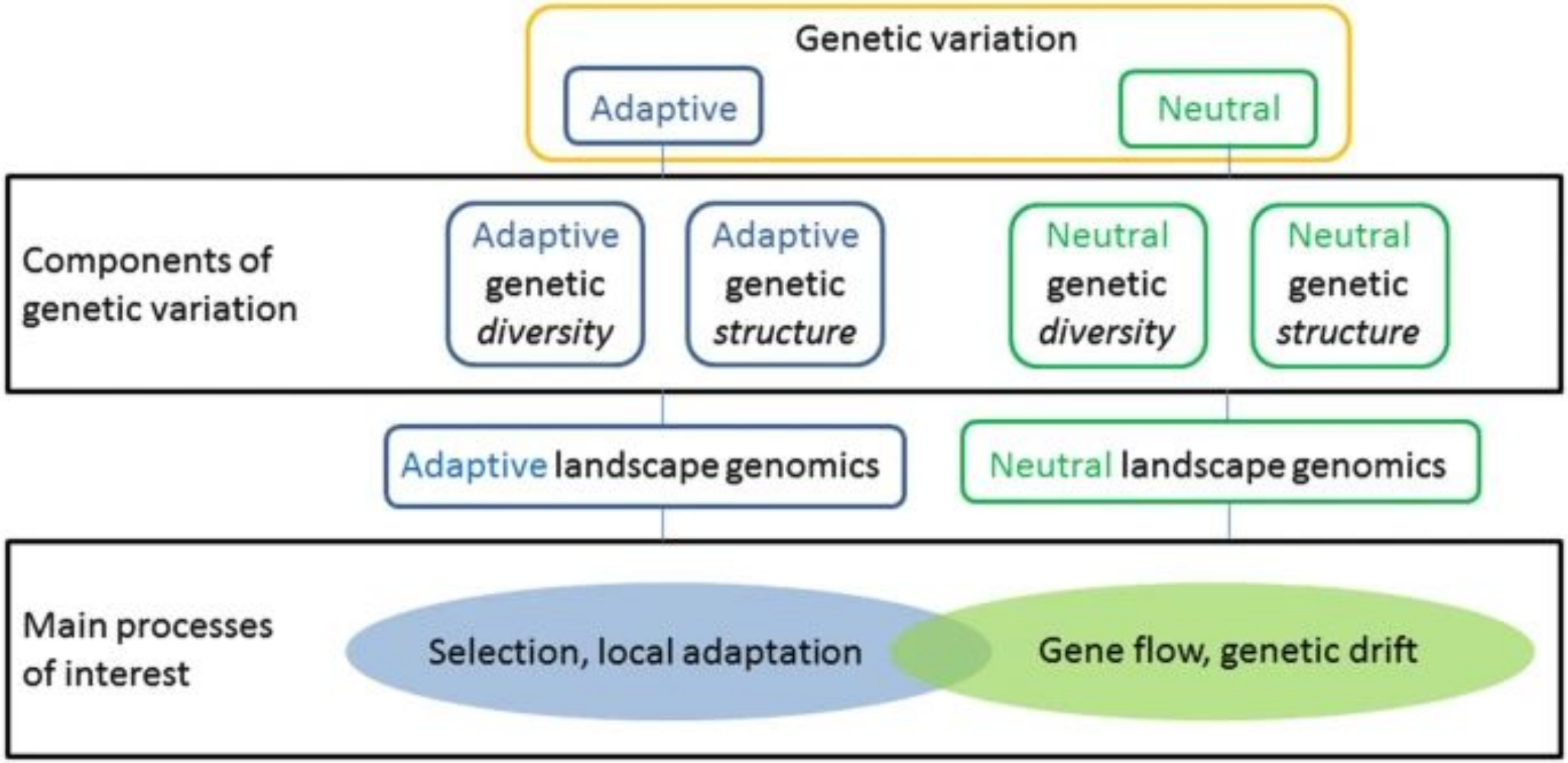
Yann Bourgeois, Thibault Leroy



Goals for today's lecture

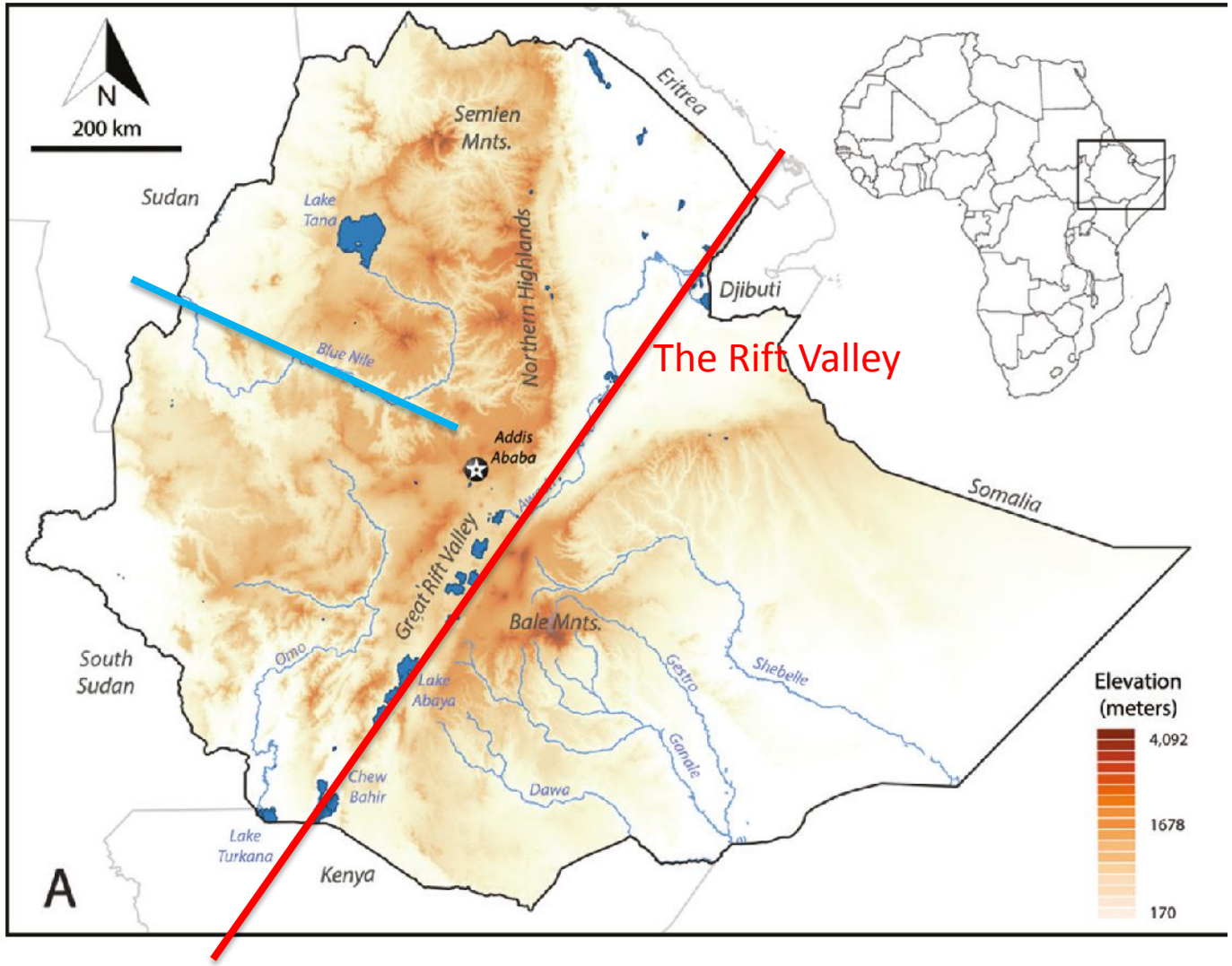
- How can we incorporate spatial information in popgen?
- How can we detect barriers to dispersal?
- How can we test hypotheses about the role of landscape features?
- How can we identify signatures of local adaptation?
- And throughout, as usual: a few limitations to keep in mind.

General framework



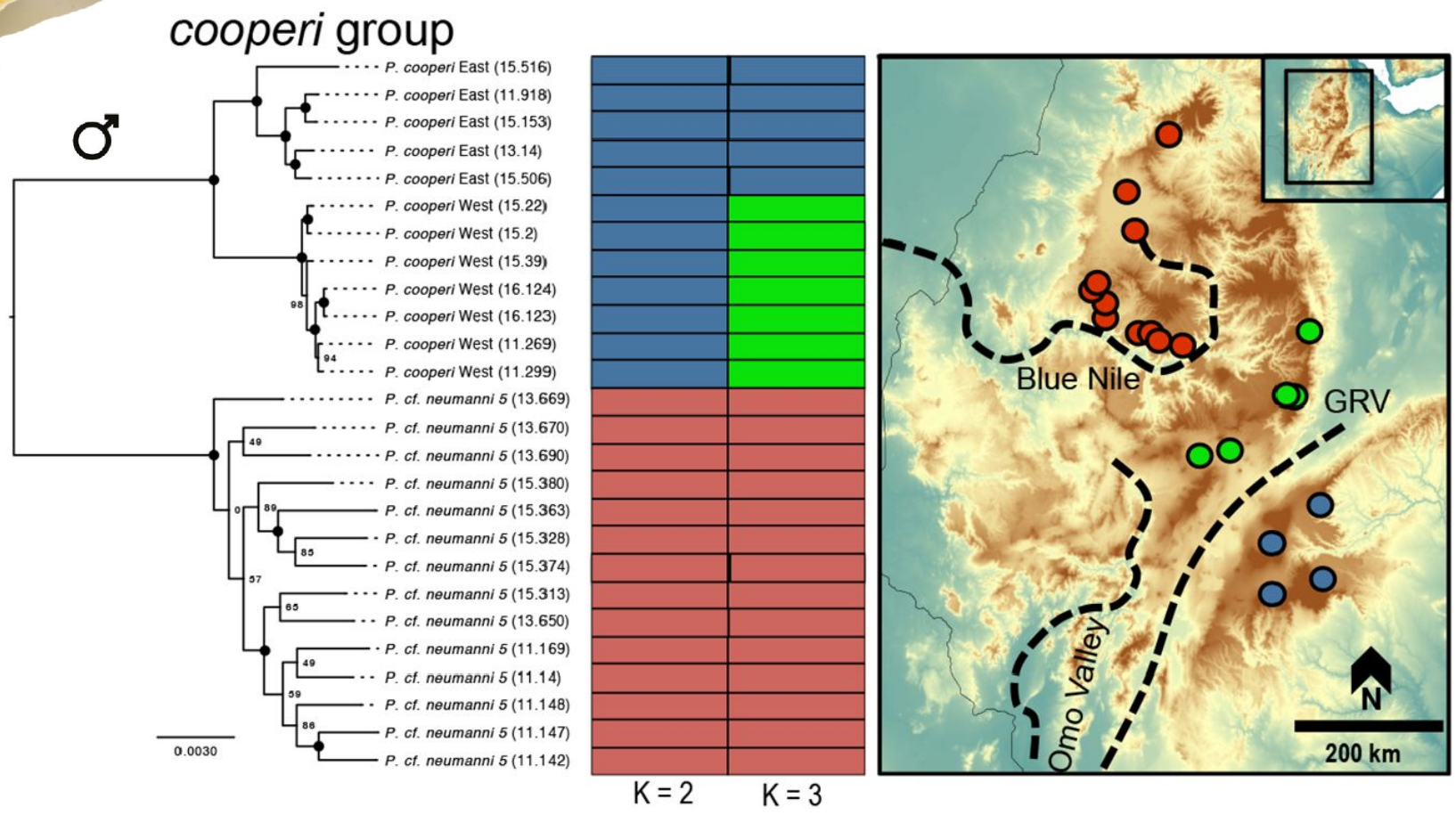
Spatializing structure analyses (Tuesday lecture, in space!)

The Blue Nile

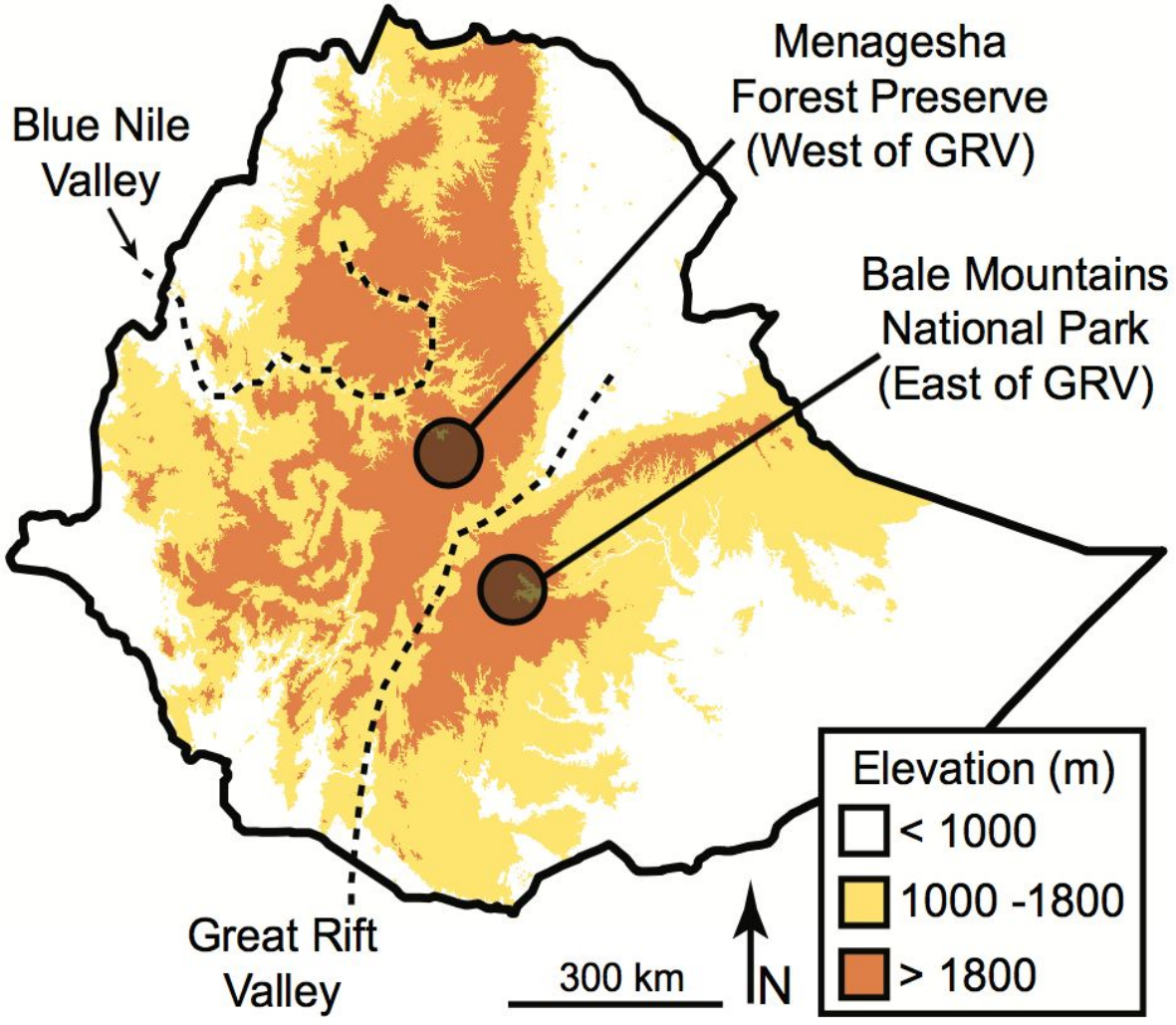


Spatializing structure analyses

Ptycha dena cooperi



Spatializing structure analyses

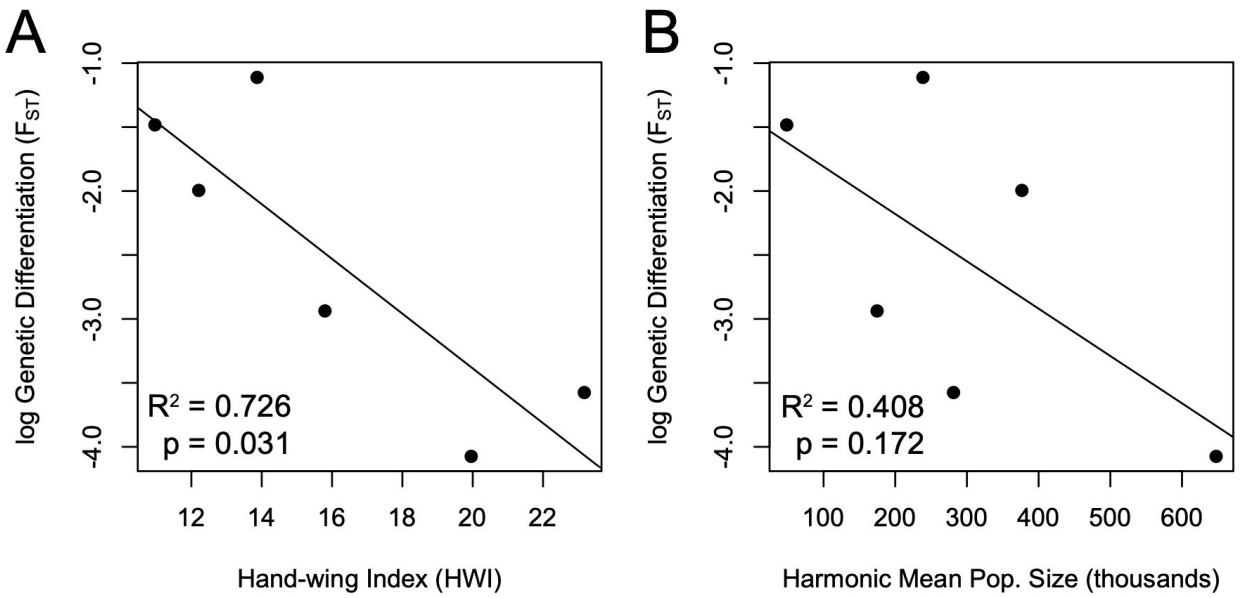


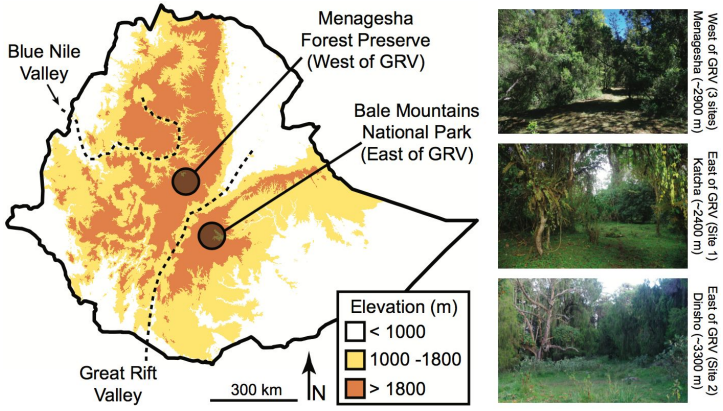
West of GRV (3 sites)
Menagesha (~2900 m)

East of GRV (Site 1)
Katcha (~2400 m)

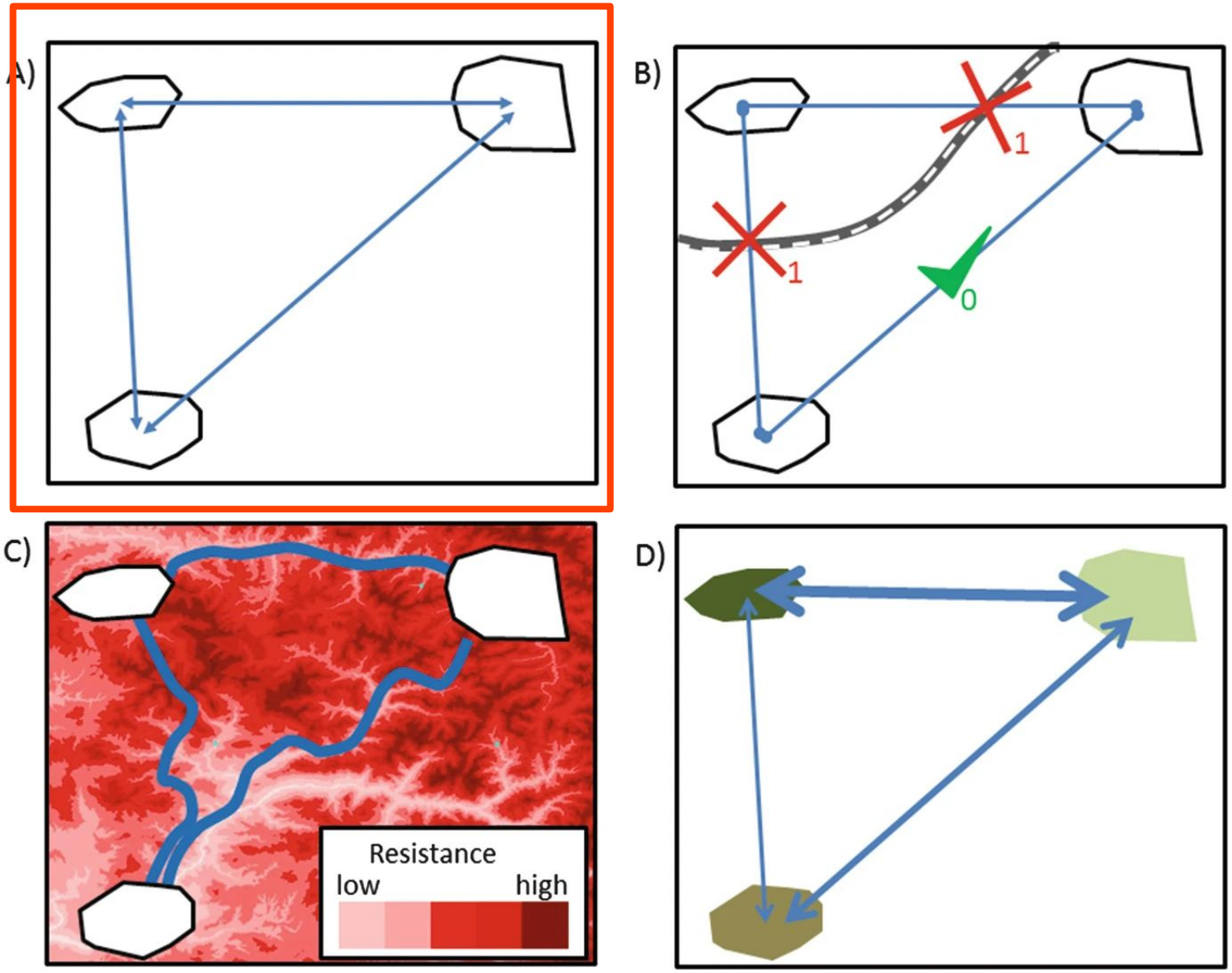
East of GRV (Site 2)
Dinsho (~3300 m)

Spatializing structure analyses

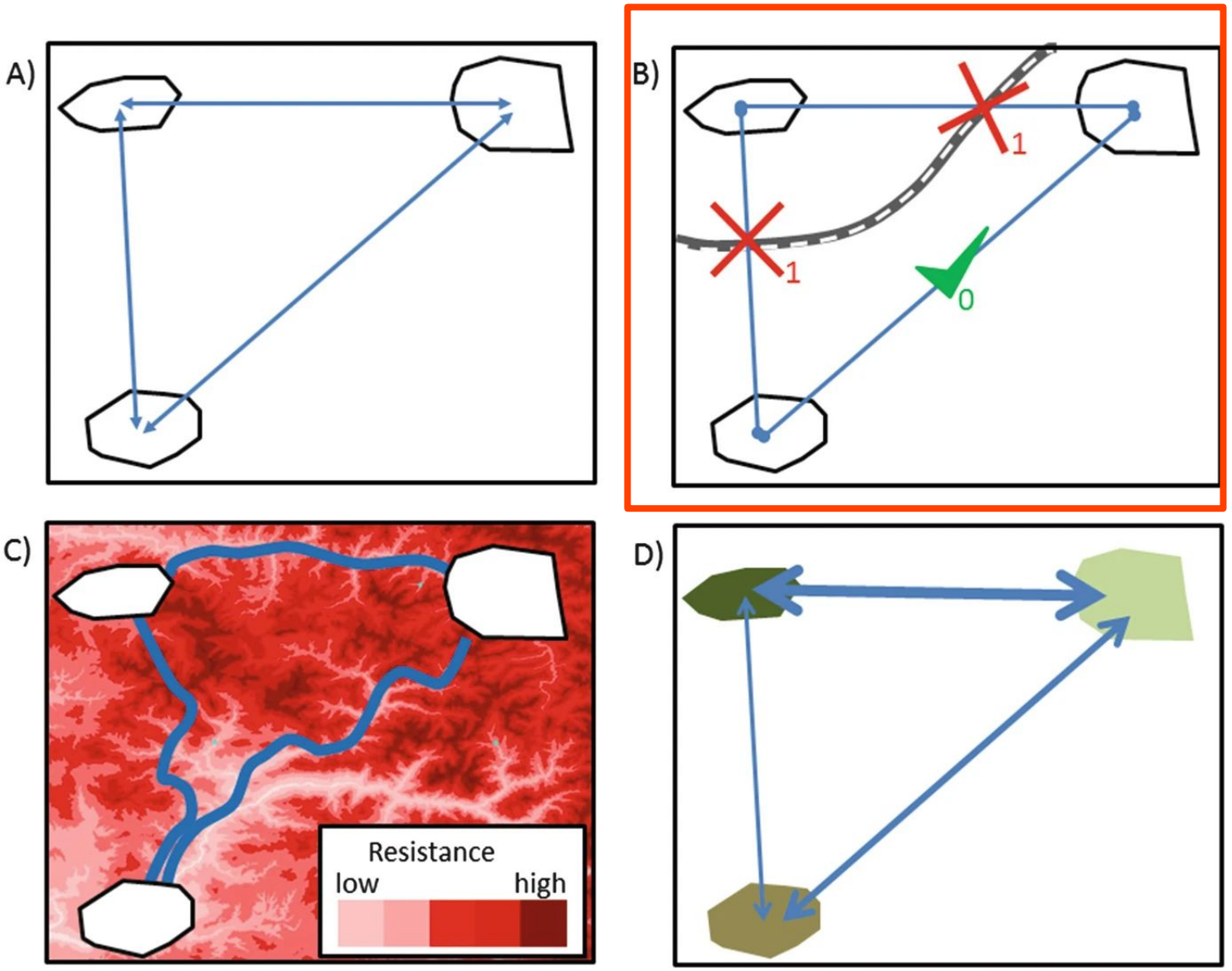




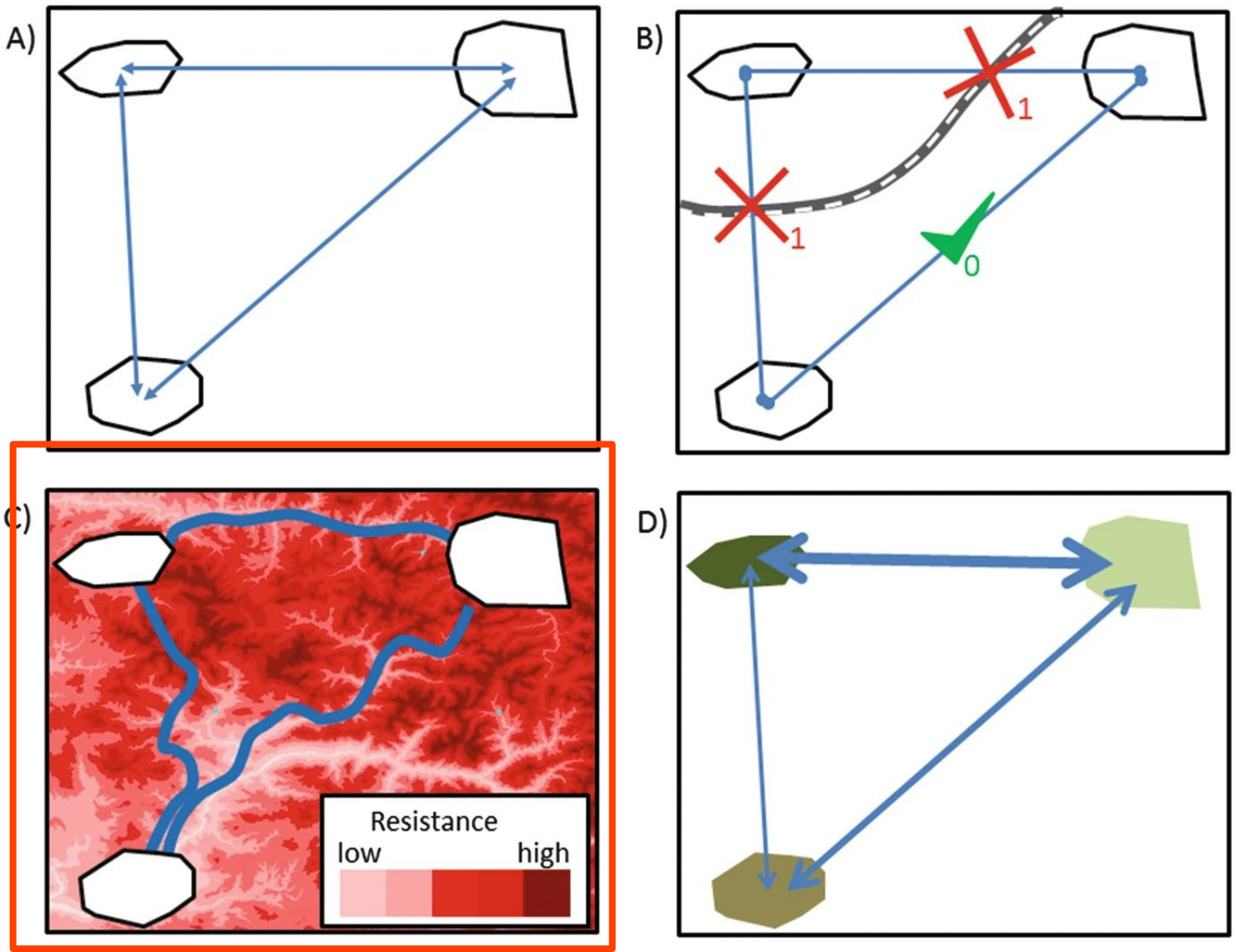
Going a bit further: Isolation-by-Distance



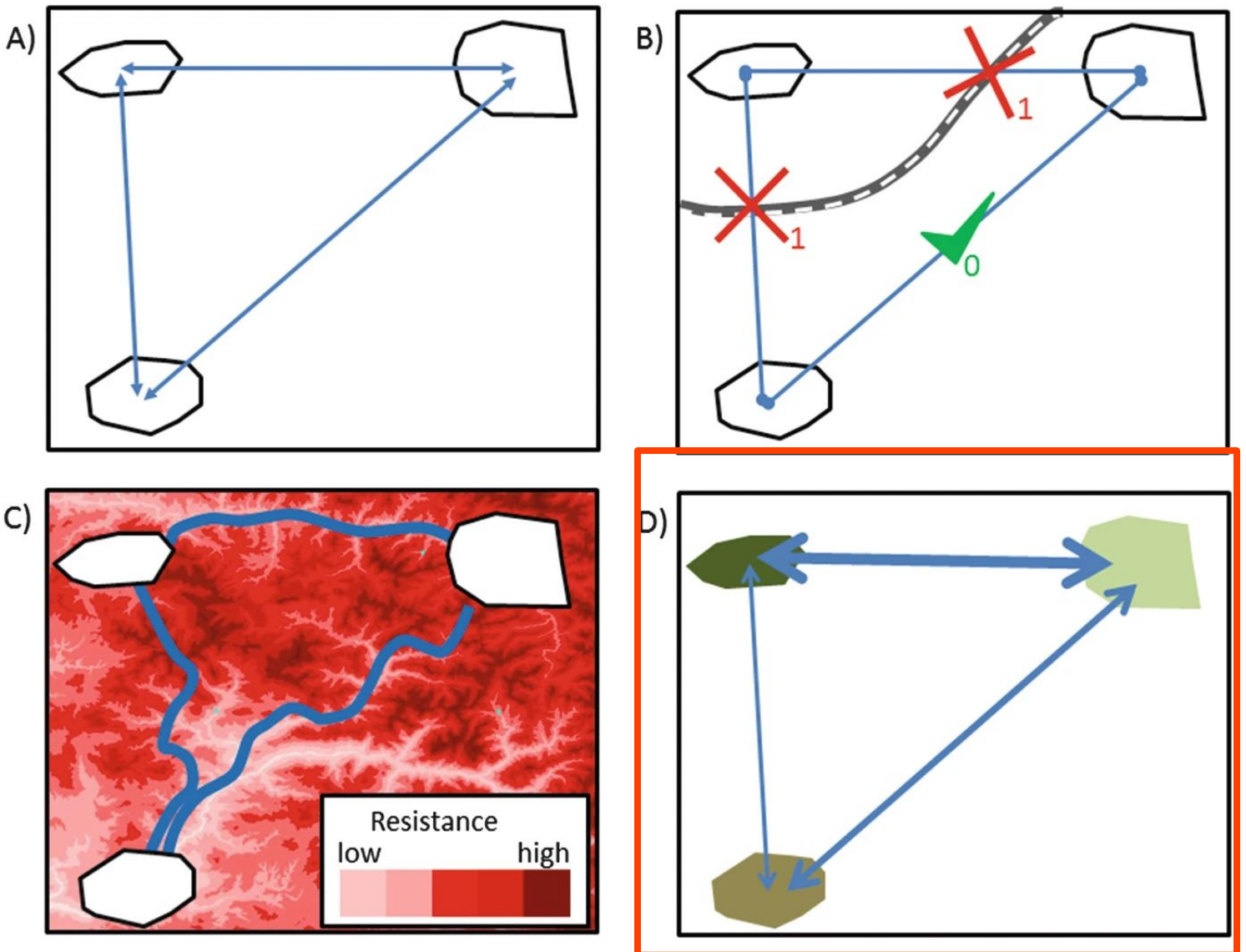
Going a bit further: Isolation-by-Barrier



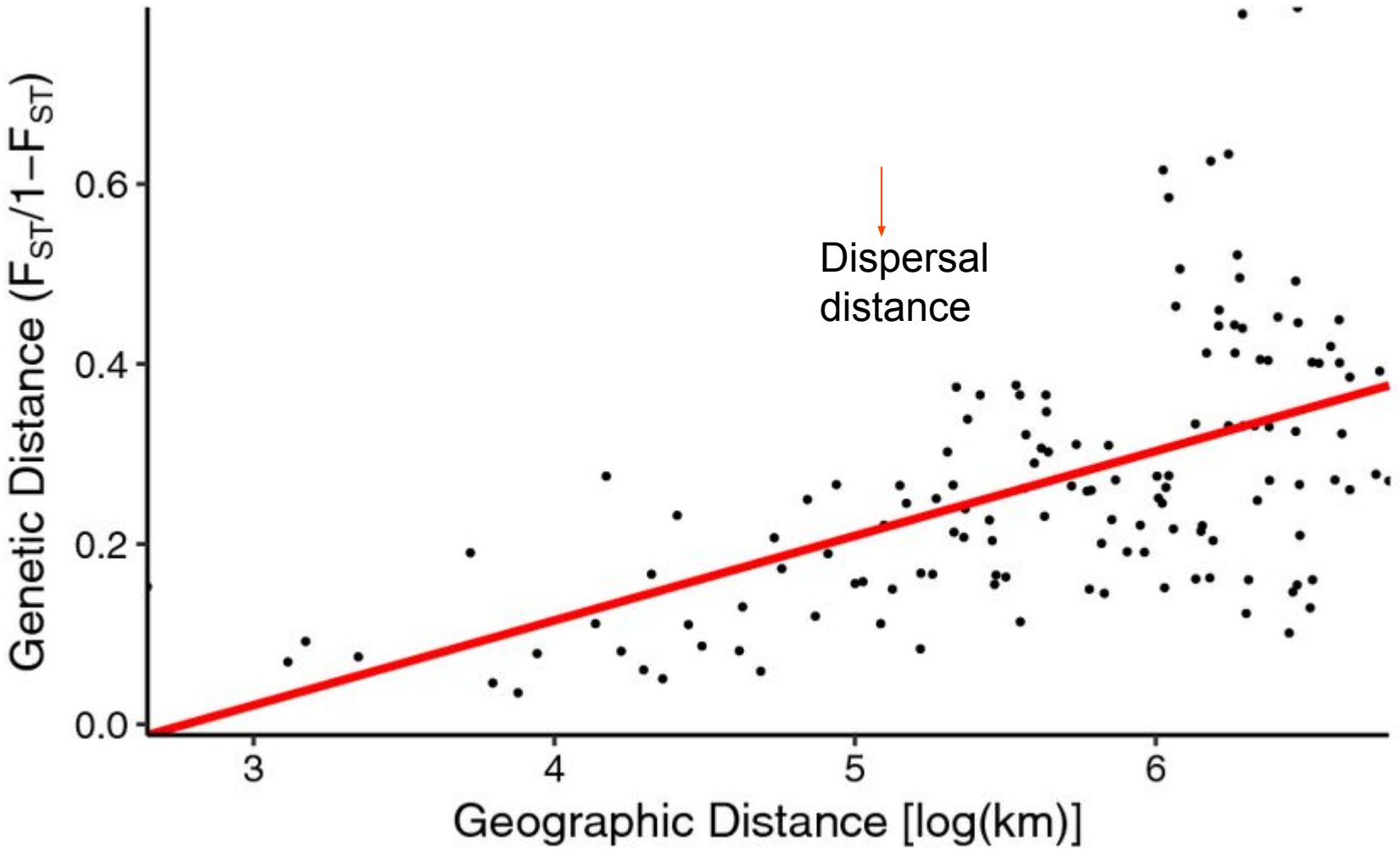
Going a bit further: Isolation-by-resistance



Going a bit further: Isolation-by-Environme



Isolation by distance



Rousset 1997

Mantel tests.

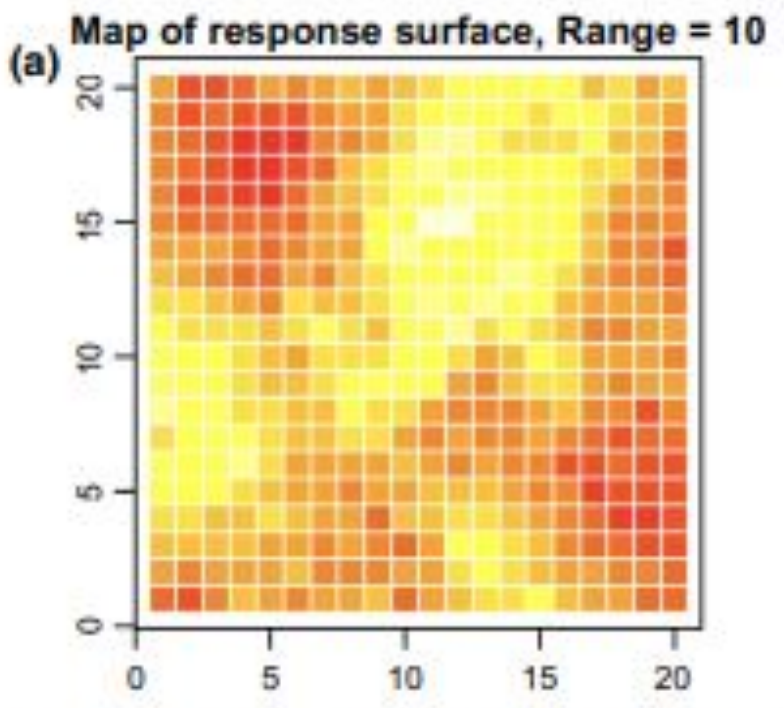
- Compares dissimilarity matrices (genetic distance G v. geographical distance S)

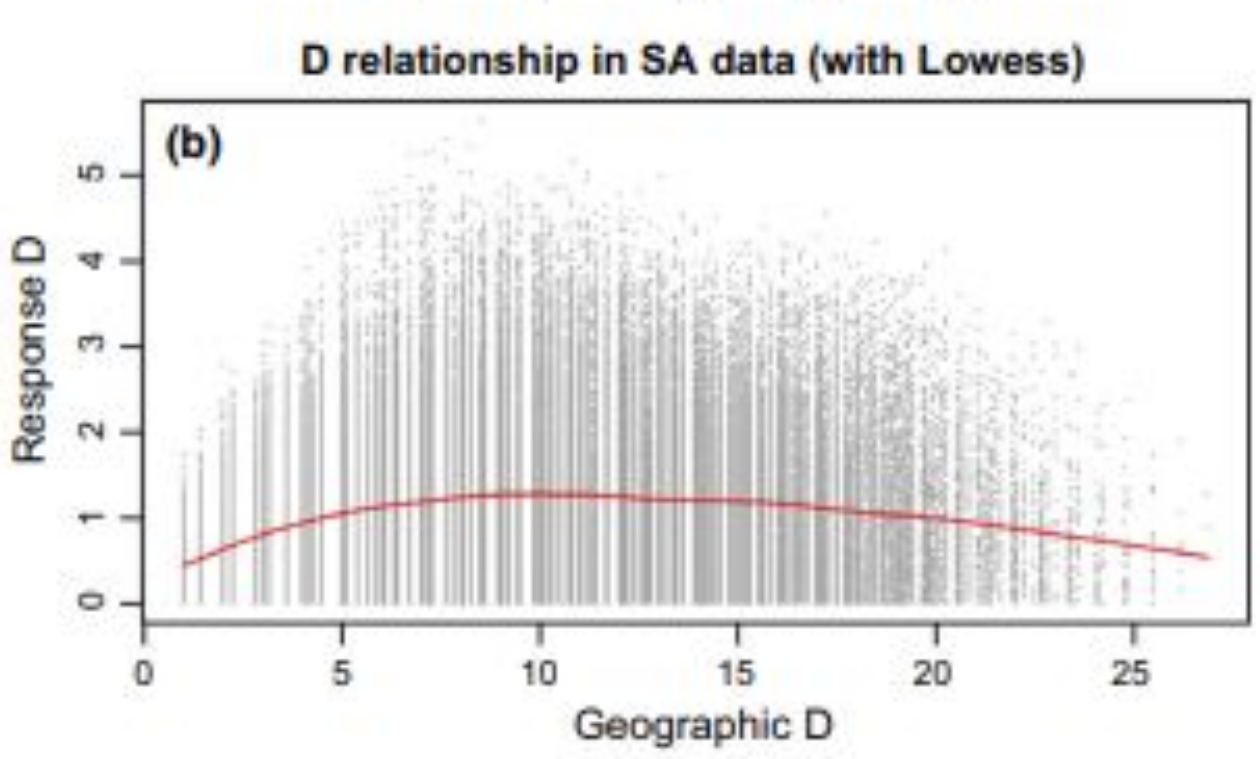
- Partial Mantel tests can be used to test for both environment E and geography S: G ~ S + E

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.

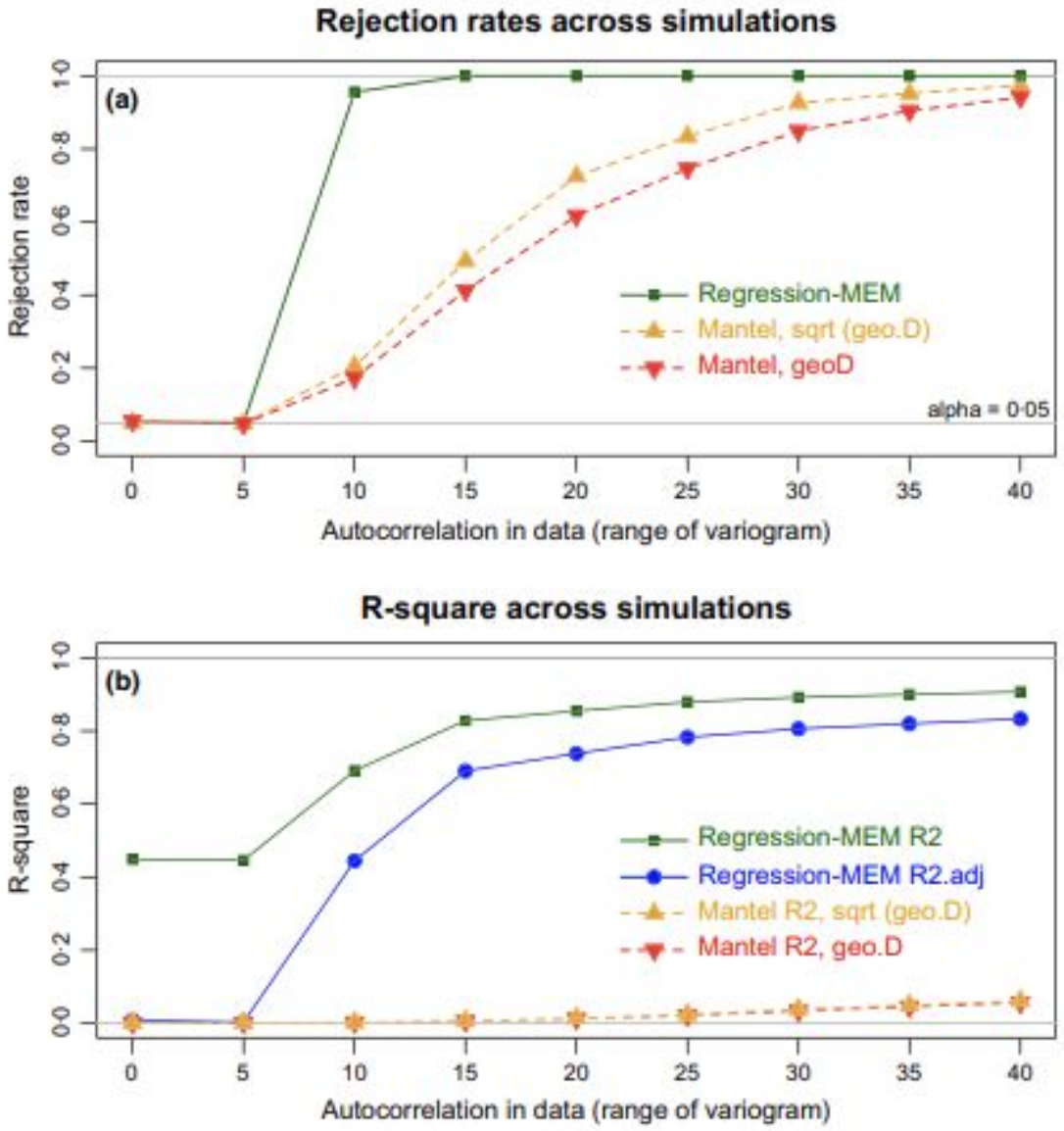
Mantel tests: Limitations.





Legendre et al. MEE, 2015

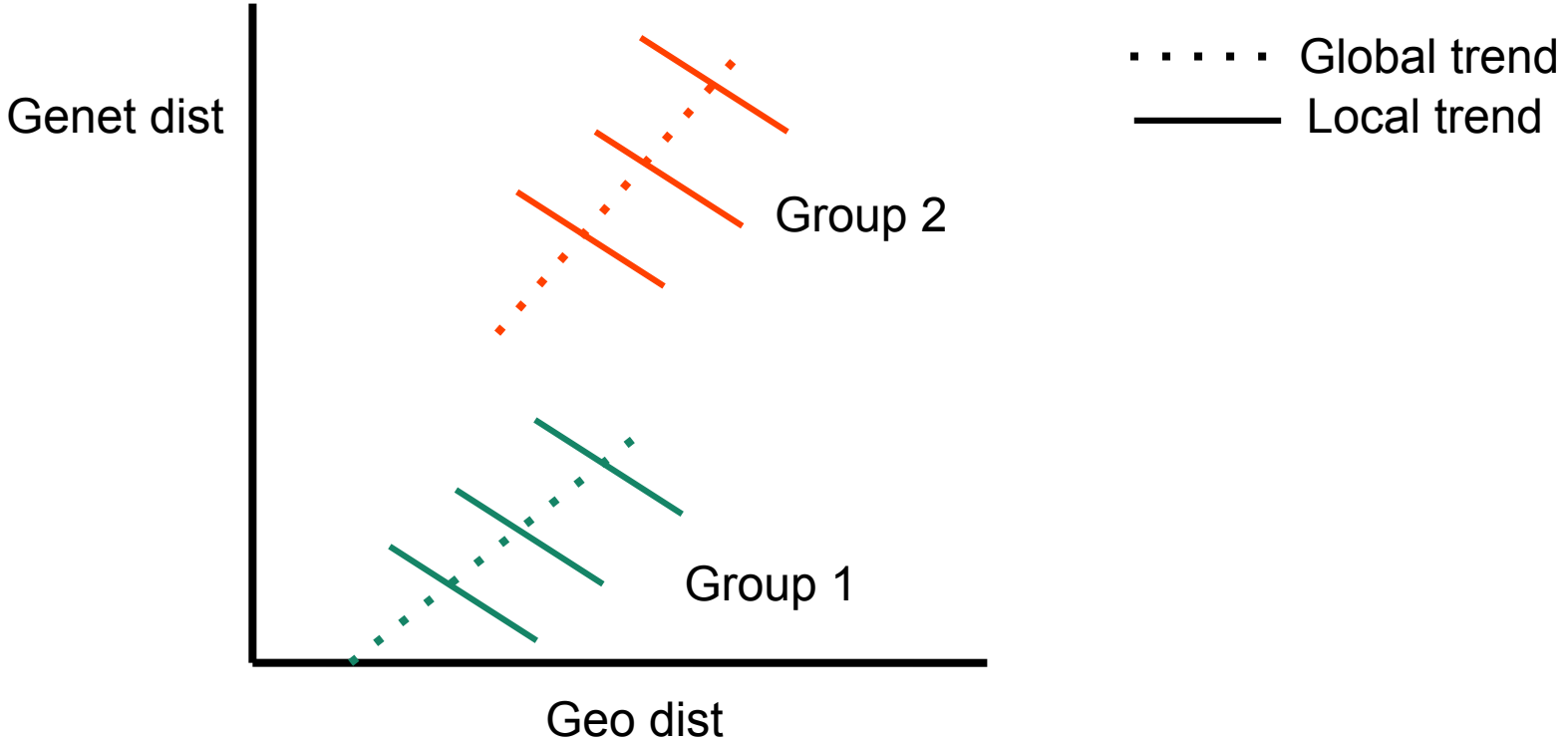
Mantel tests: Limitations.



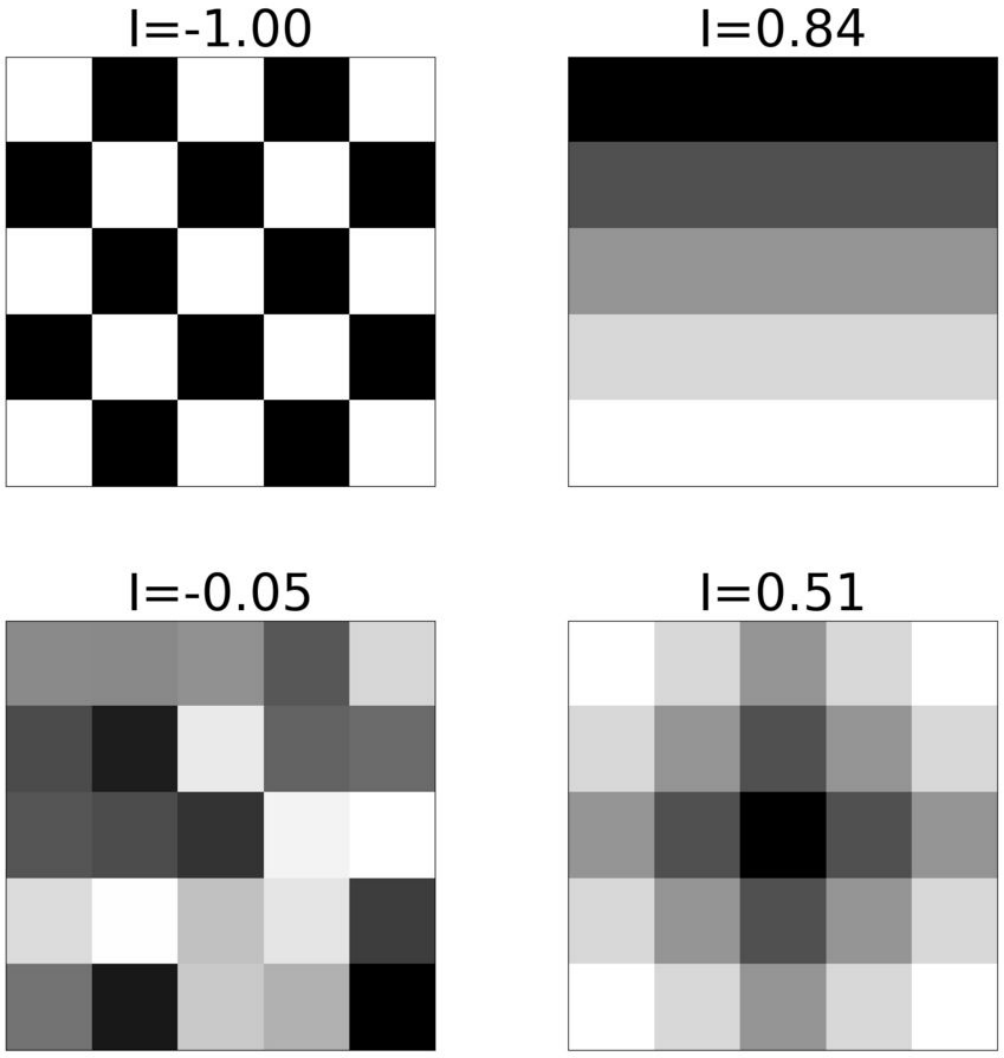
Legendre et al. MEE, 2015

Limitations to Mantel test

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



Describe spatial autocorrelation. Moran's I.

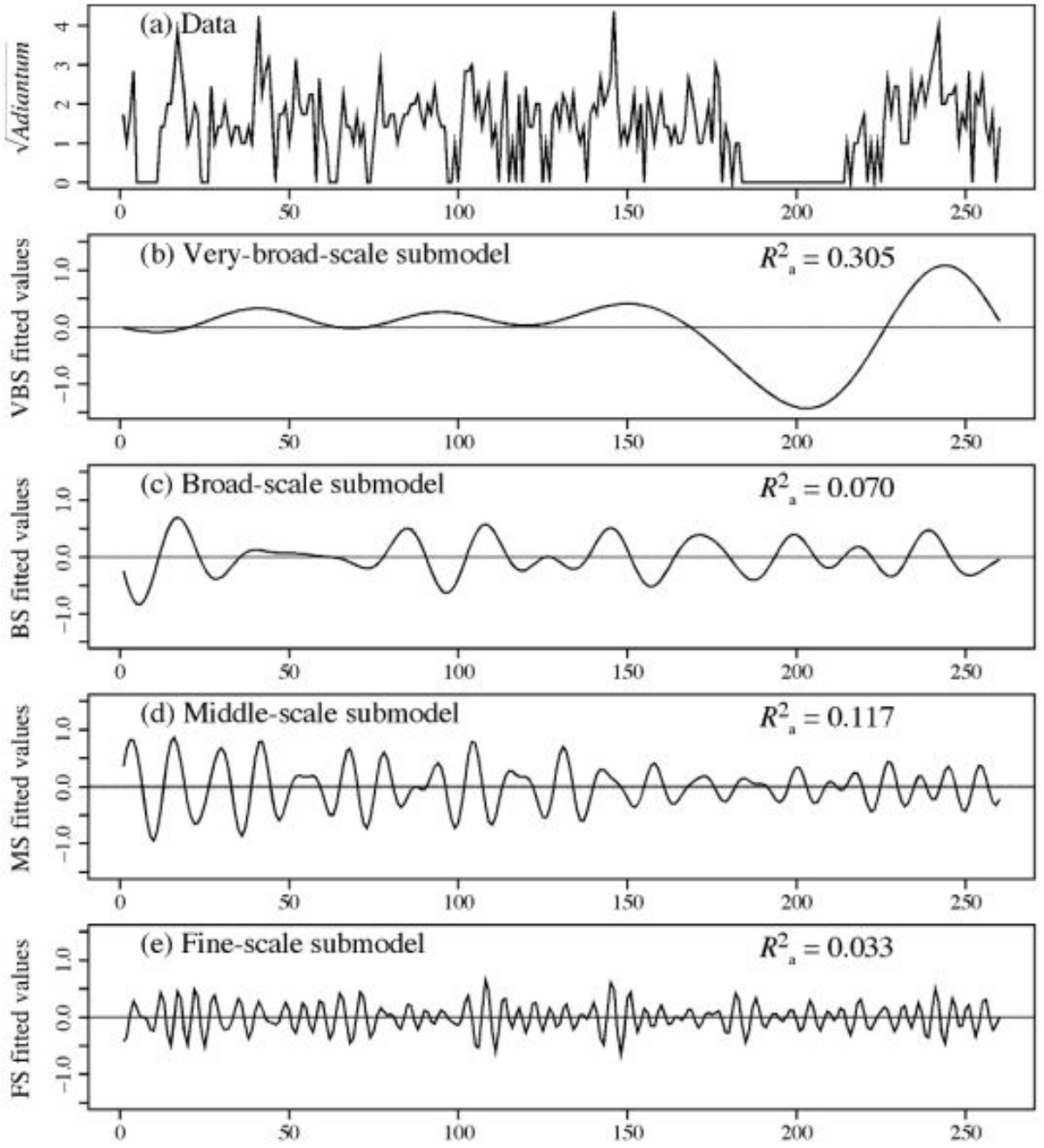


dbMEM.

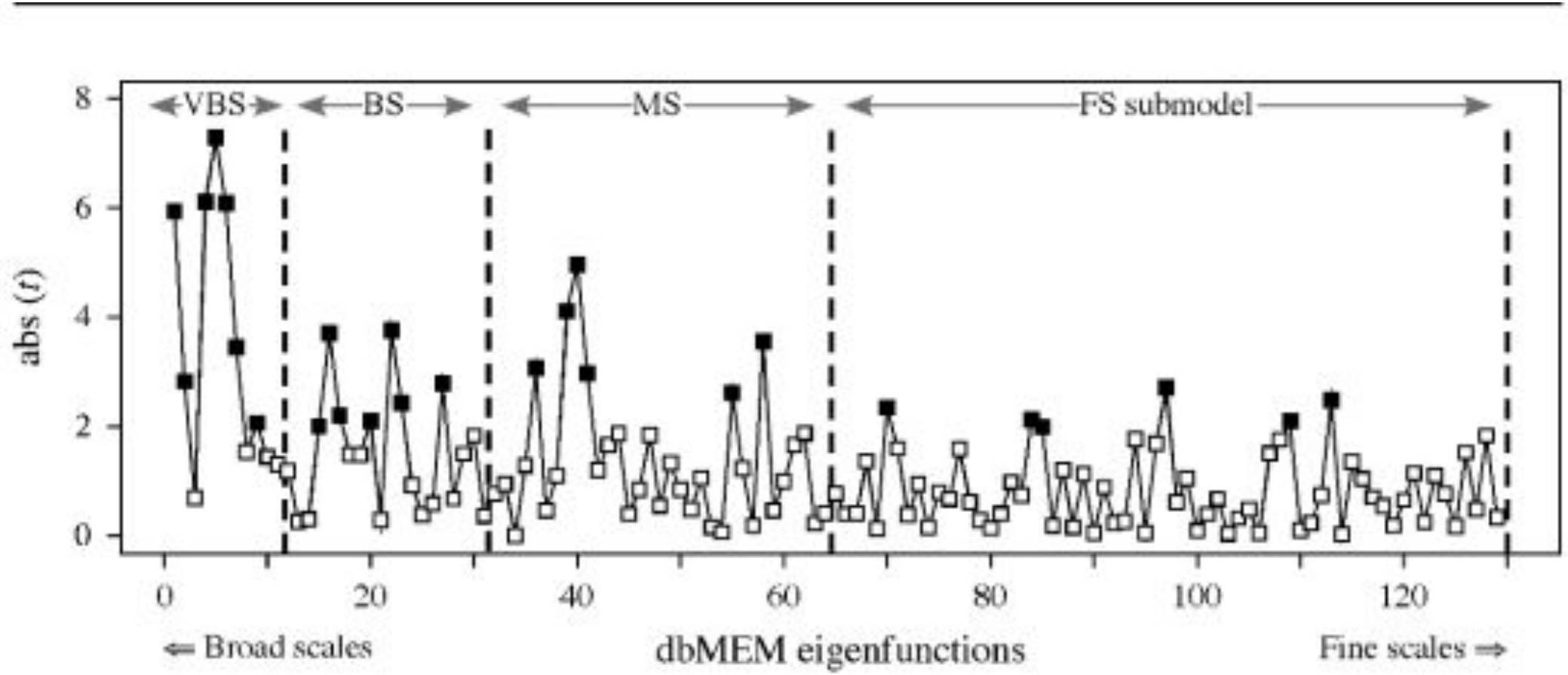
- We can decompose spatial variation along a gradient (below, 1D) into a combination of independent functions (eigenvectors)

- Some will reflect positive correlations between distances (positive Moran's I), some are negative.

- Approaches based on so-called 'distance-based Moran's eigenvector [illegible]'dbMEM

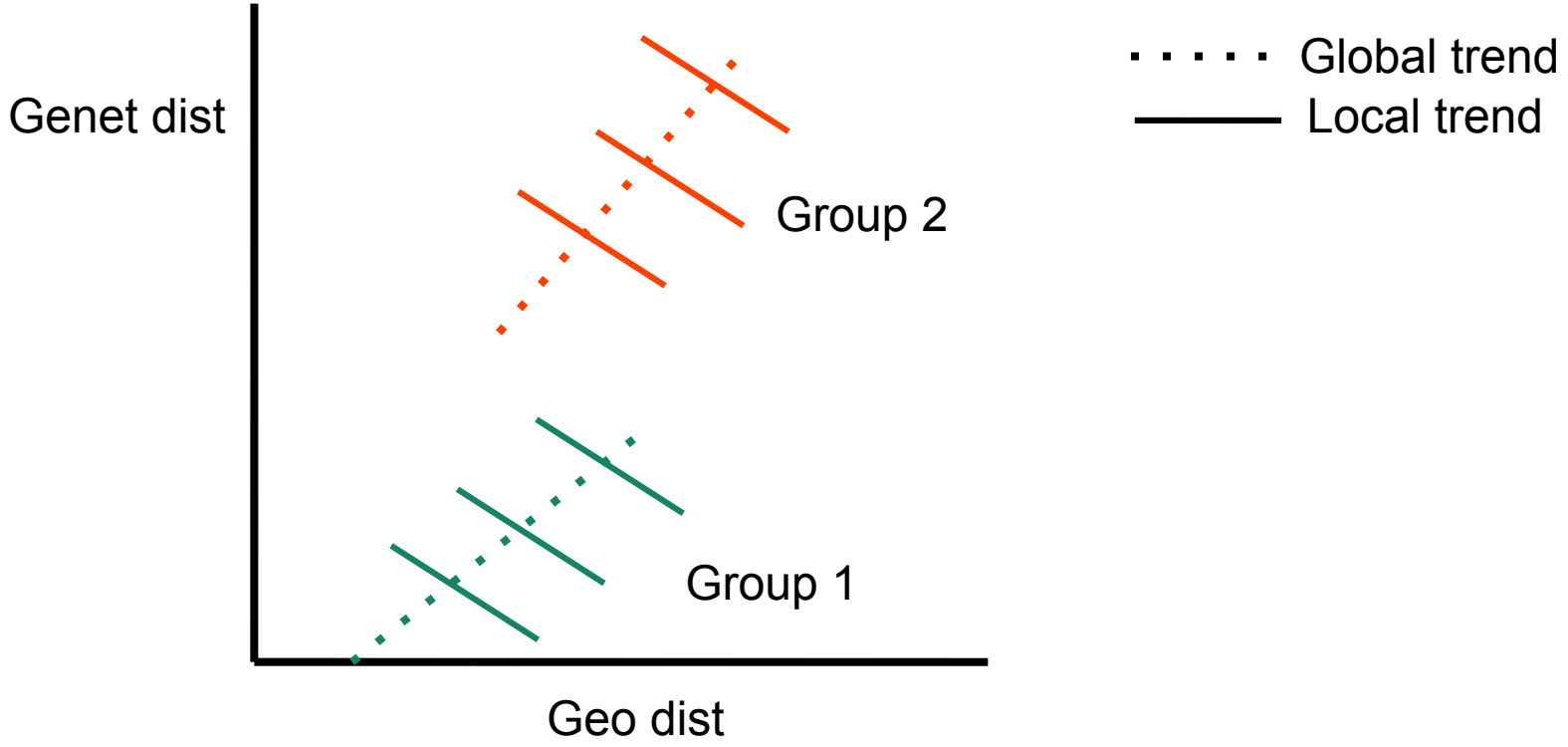


dbMEM.


VBS
BS
MS
FS submodel
abs (t)
8
6
4
2
0
0
20
40
60
80
100
120
⇐ Broad scales
dbMEM eigenfunctions
Fine scales ⇒

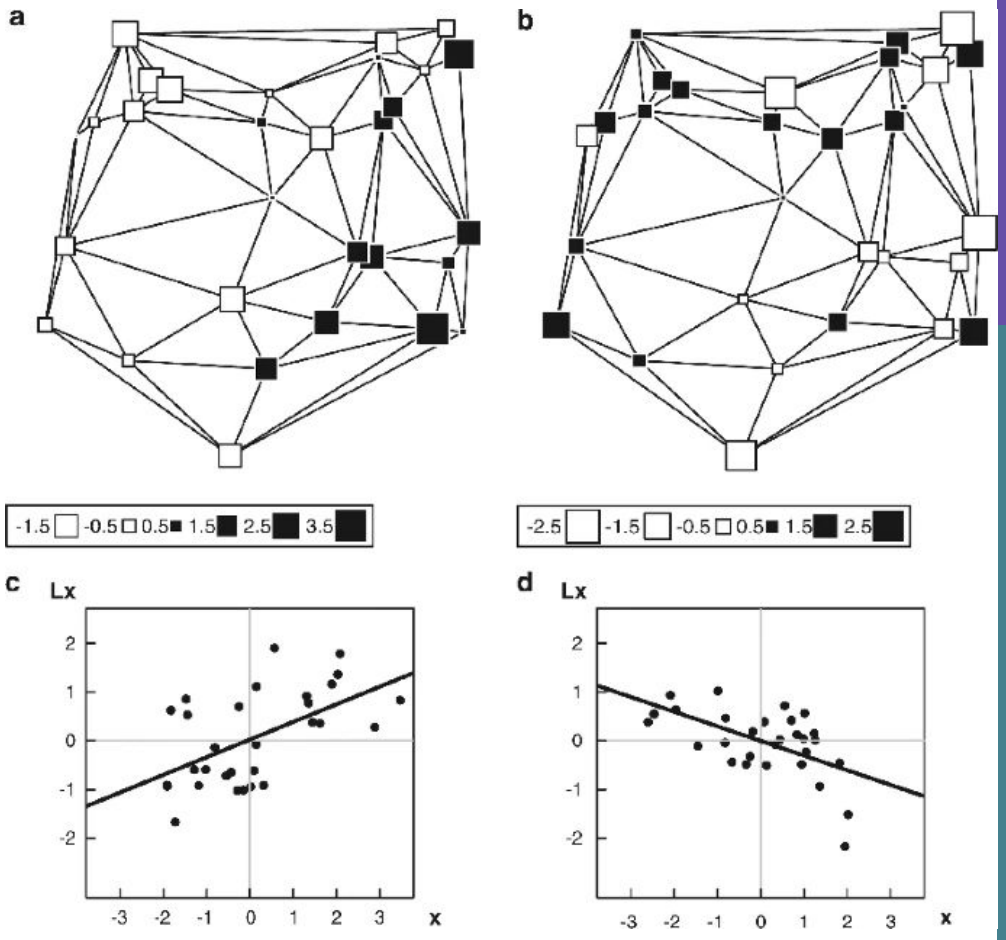
Spatial PCA (sPCA).

- 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)



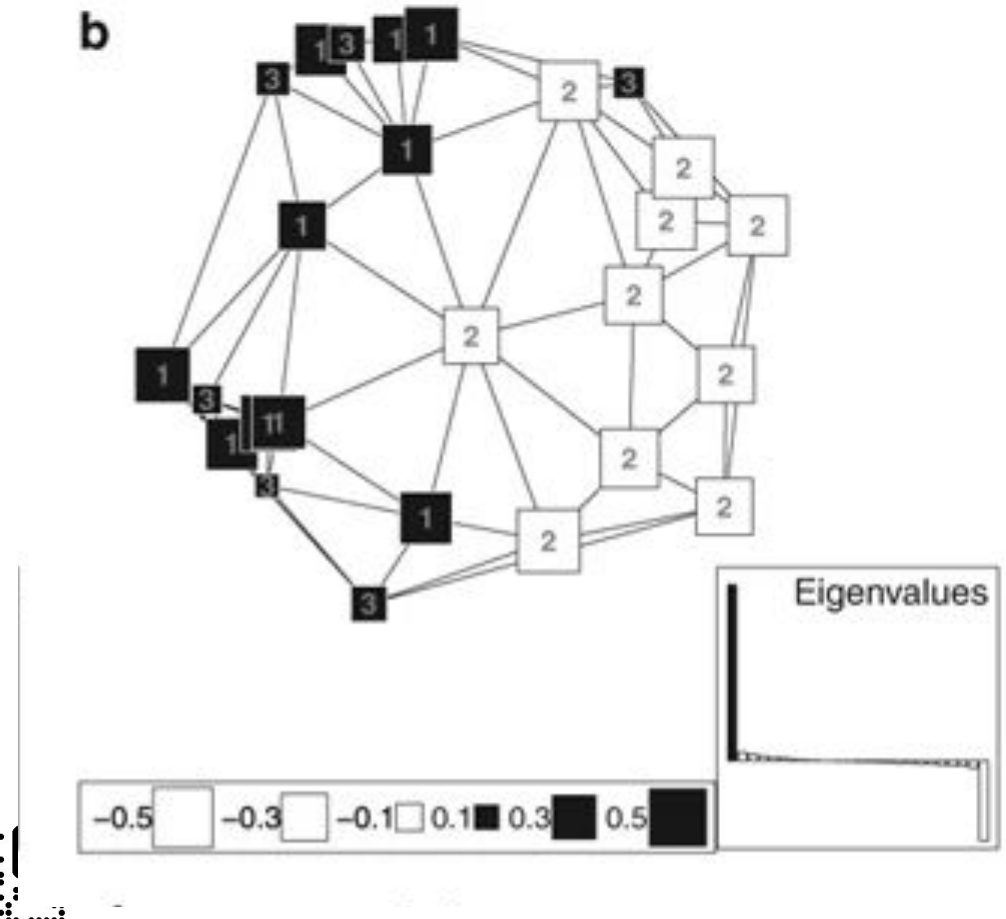
Spatial PCA (sPCA).

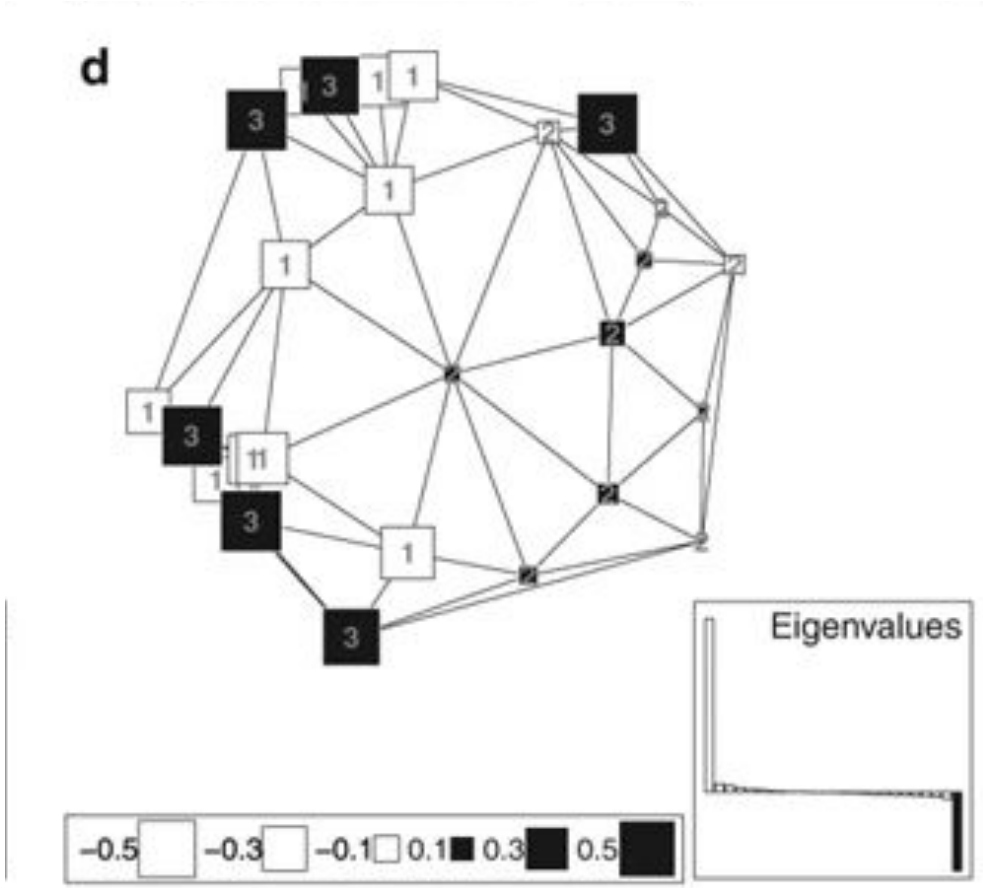
Illustration of global and local patterns of an allelic frequency for 20 fictitious populations overlying their sampling area. Each square represents the frequency of a population. Edges correspond to the connection network (Gabriel's graph). (**a**) Example of global structure, corresponding to $I(\mathbf{x})>I_0$. (**b**) Example of local structure, corresponding to $I(\mathbf{x})<I_0$. (**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.

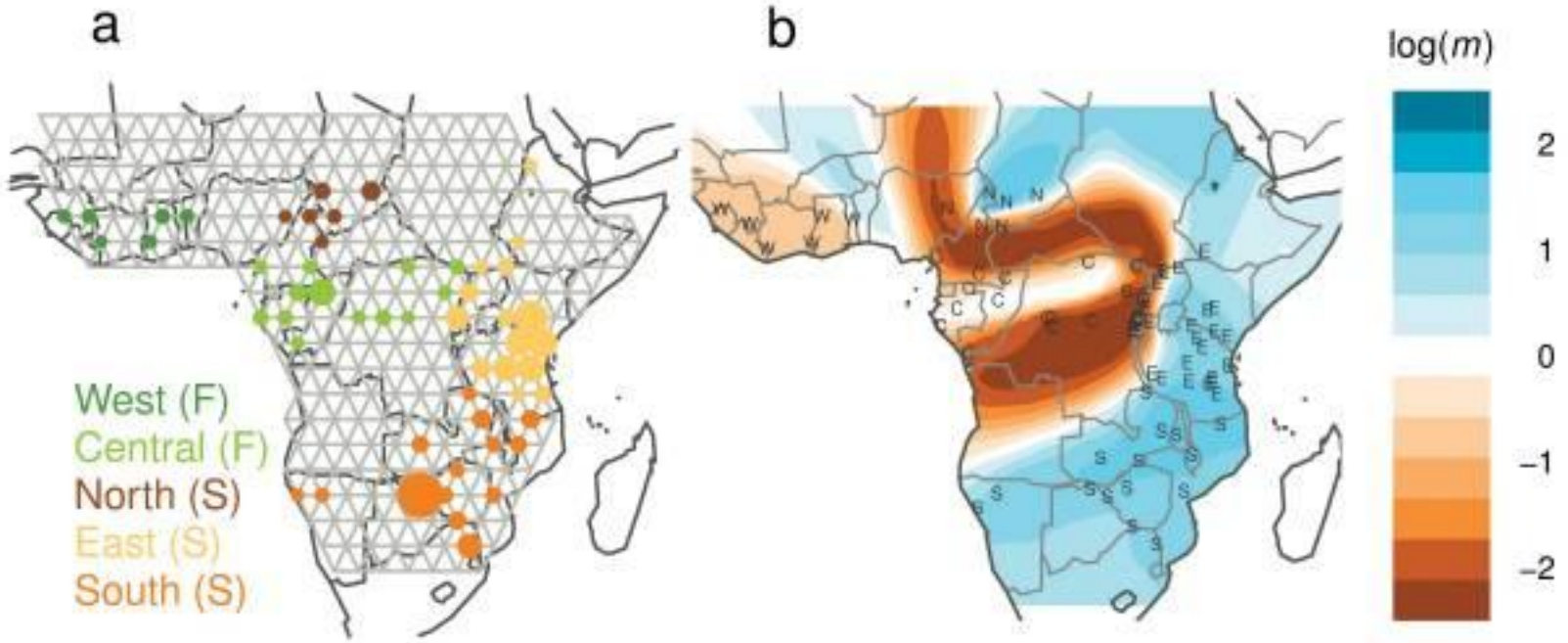
Barriers and corridors to gene flow

EEMS: isolation by resistance
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



Petkova et al. 2016, Nat. Genet.

Barriers and corridors to gene flow

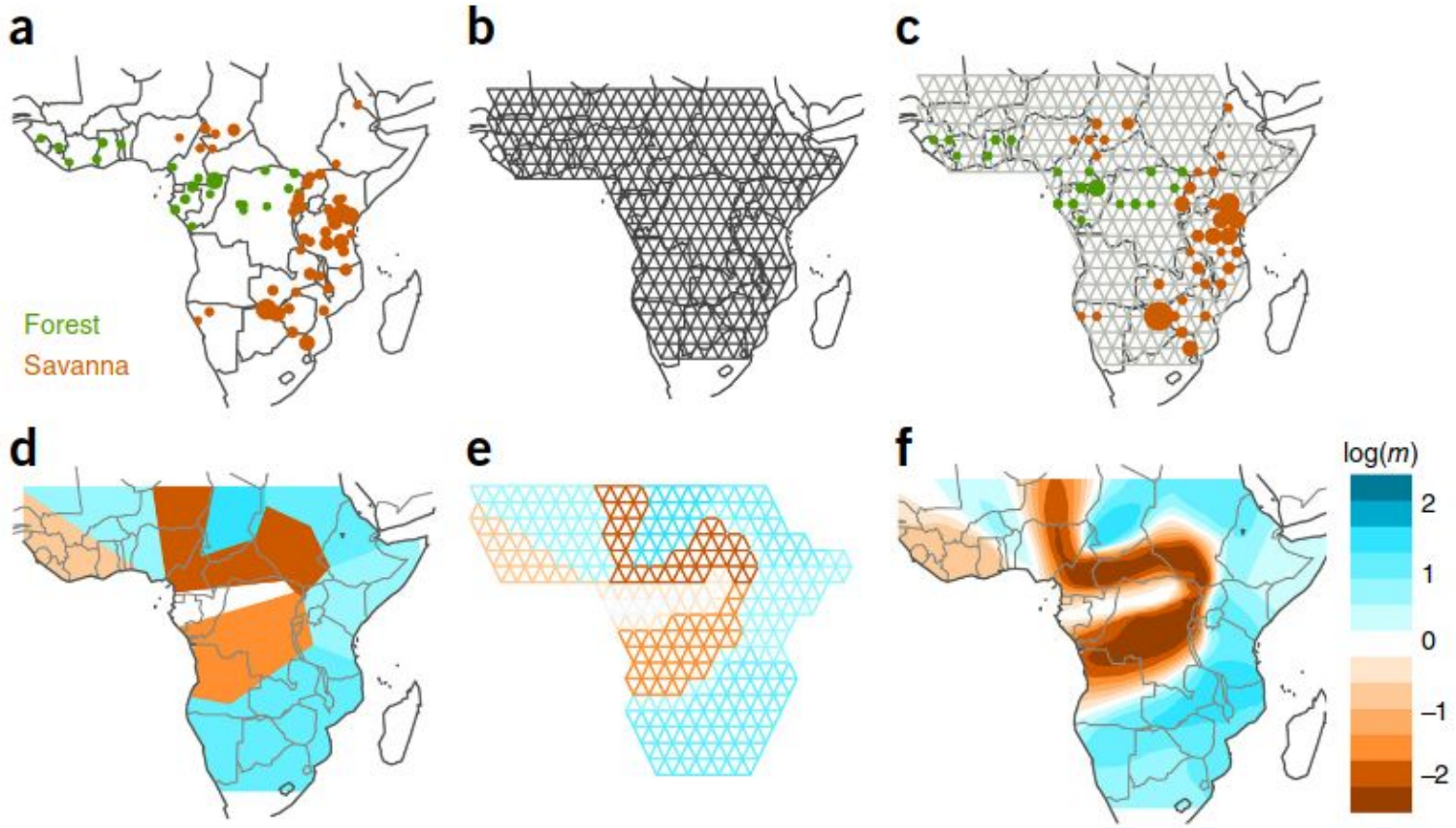
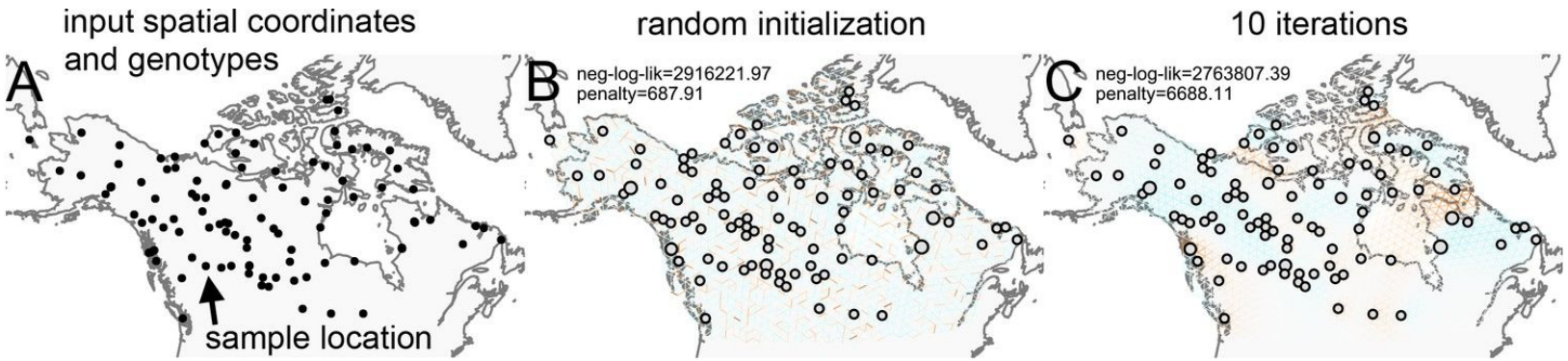
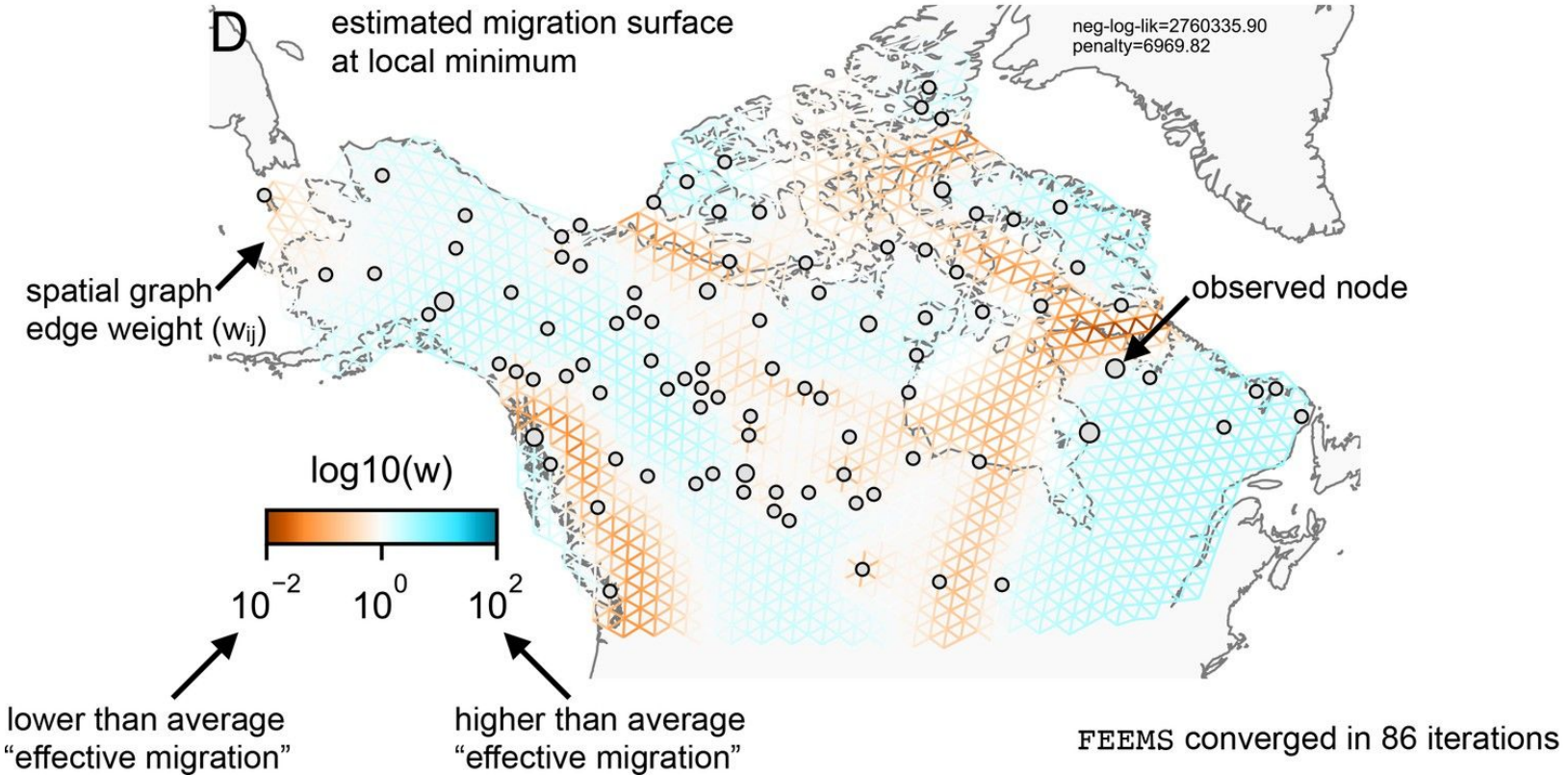


Figure 1 A schematic overview of EEMS (Estimated Effective Migration Surfaces), using African elephant data for illustration. (**a**–**c**) Setting up the population grid. (**a**) Samples are collected at known locations across a two-dimensional habitat; green and orange represent two species of African 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.

Petkova et al. 2016, Nat. Genet.

Barriers and corridors to gene flow



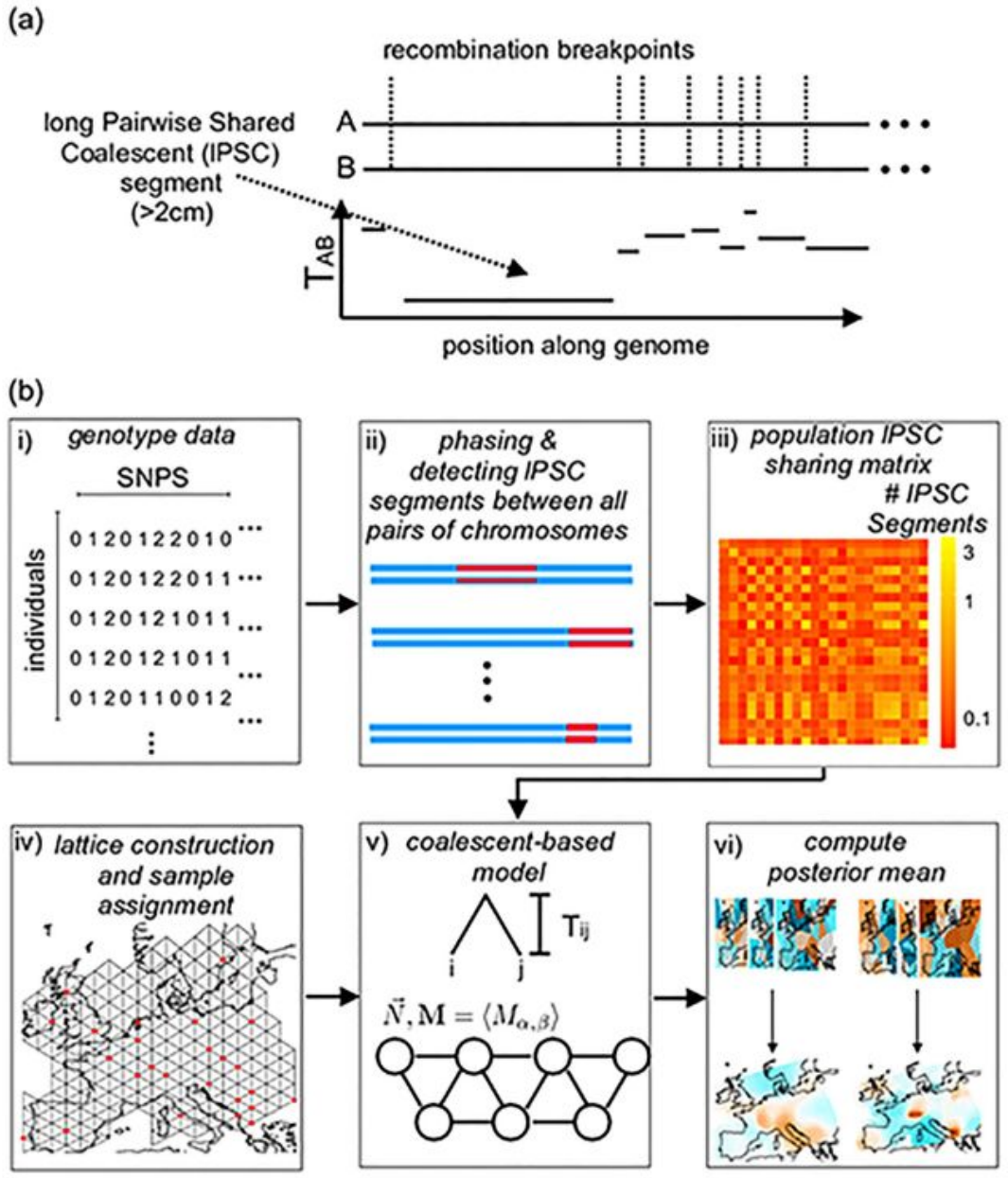


FEEMS converged in 86 iterations

FEEMS: faster than EEMS.

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.

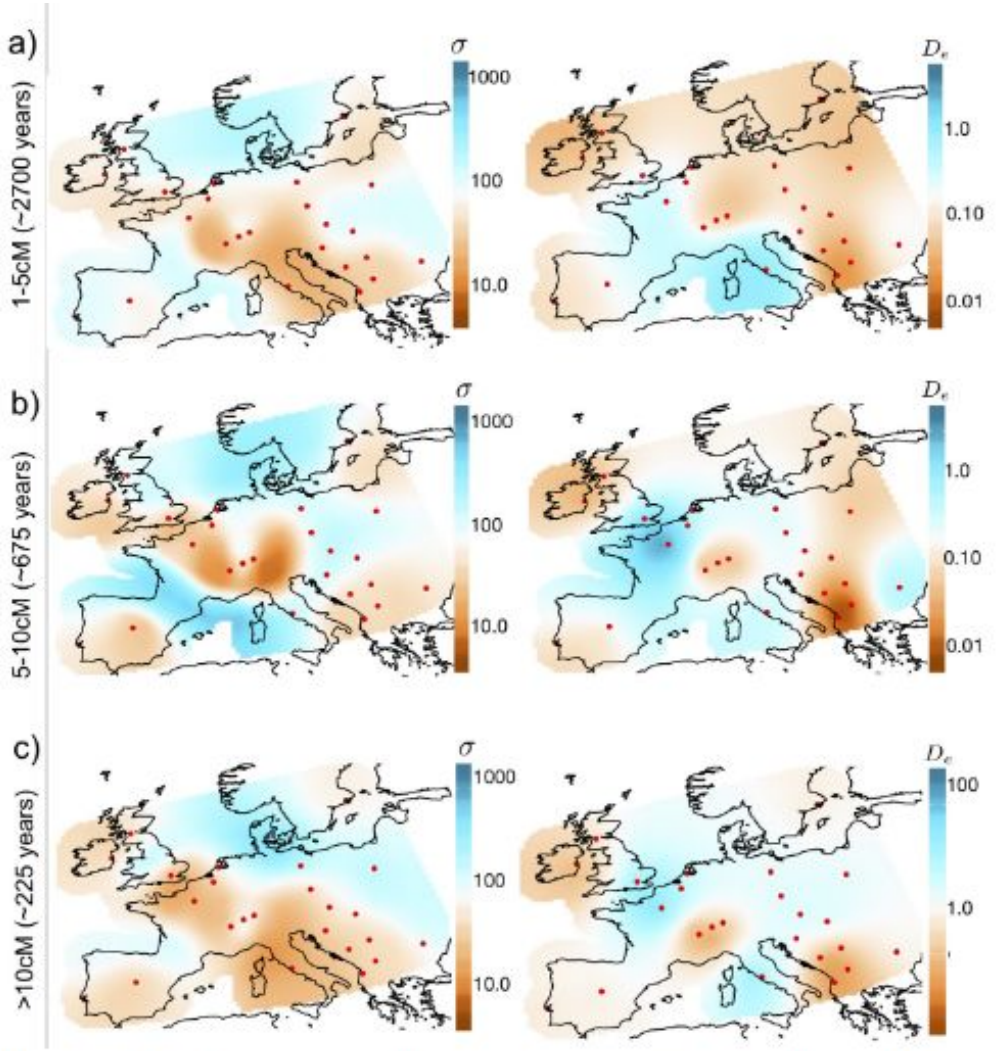
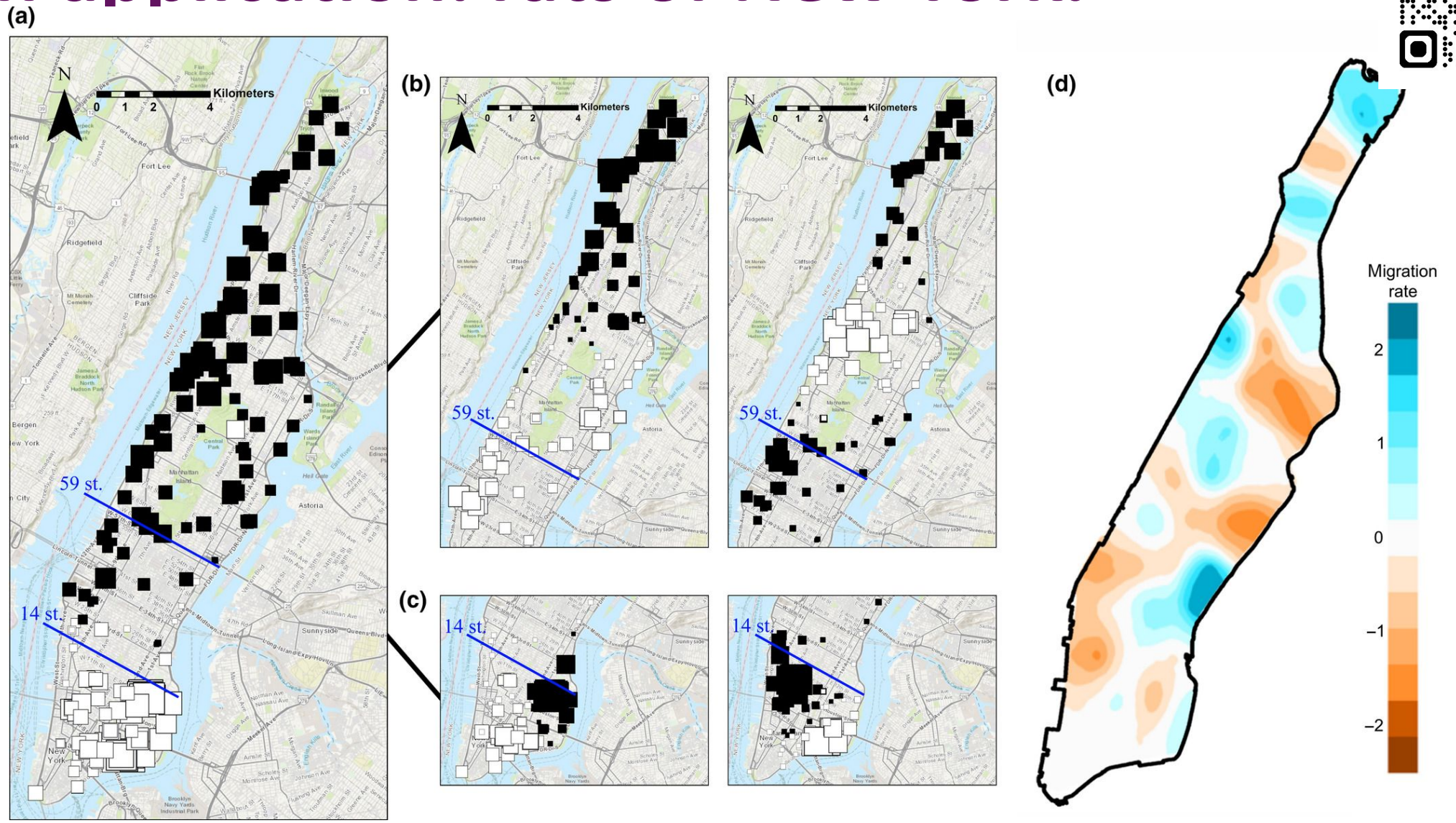


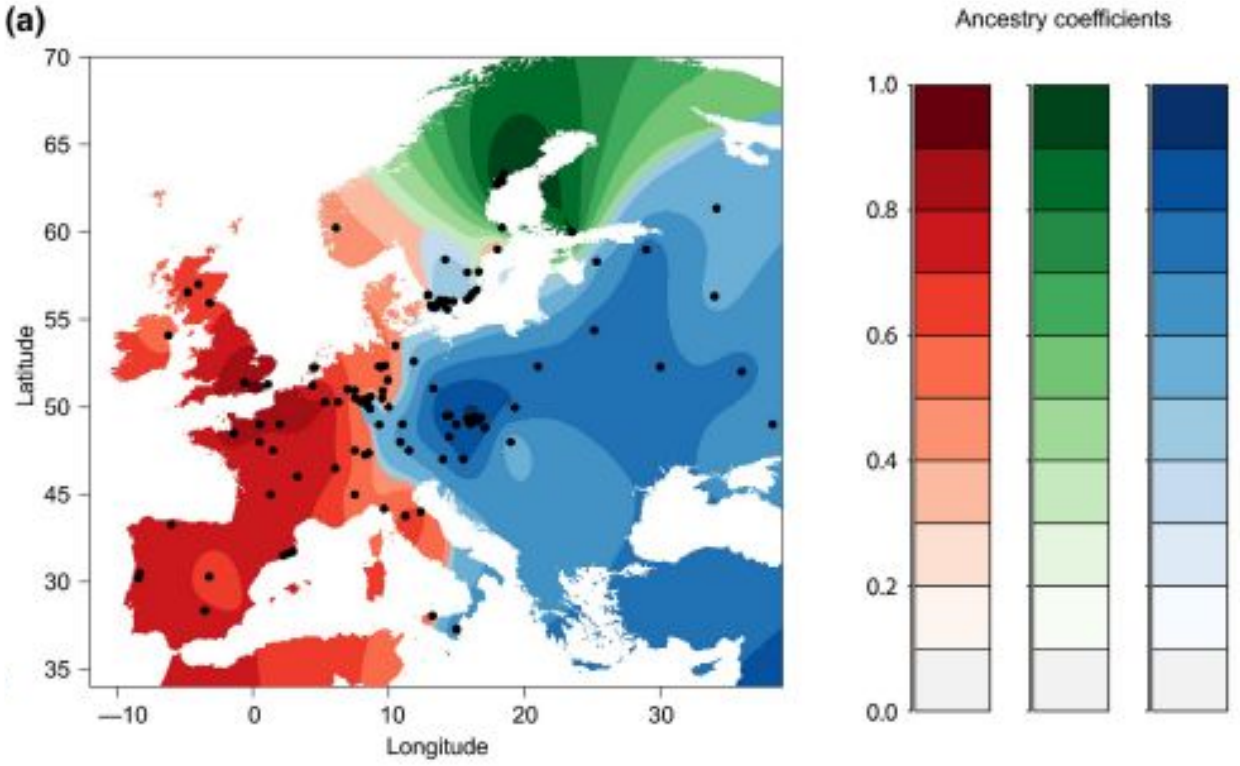
Fig 4. Inferred dispersal surfaces and population density surfaces over time for Europe. We apply MAPS to a European subset of POPRES [25] with 2,234 individuals and plot the inferred dispersal $\sigma(\vec{x})$ and population density $D_e(\vec{x})$ surfaces for PSC length bins (a) >1cM (b) 5-10cM and (c) >10cM. We transform estimates of $\vec{N}$ and **M** to estimates of $\sigma(\vec{x})$ and $D_e(\vec{x})$ 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 (>10cM). Note the wider plotting limits in inferred densities in the most recent time scale.

An application: rats of New York.



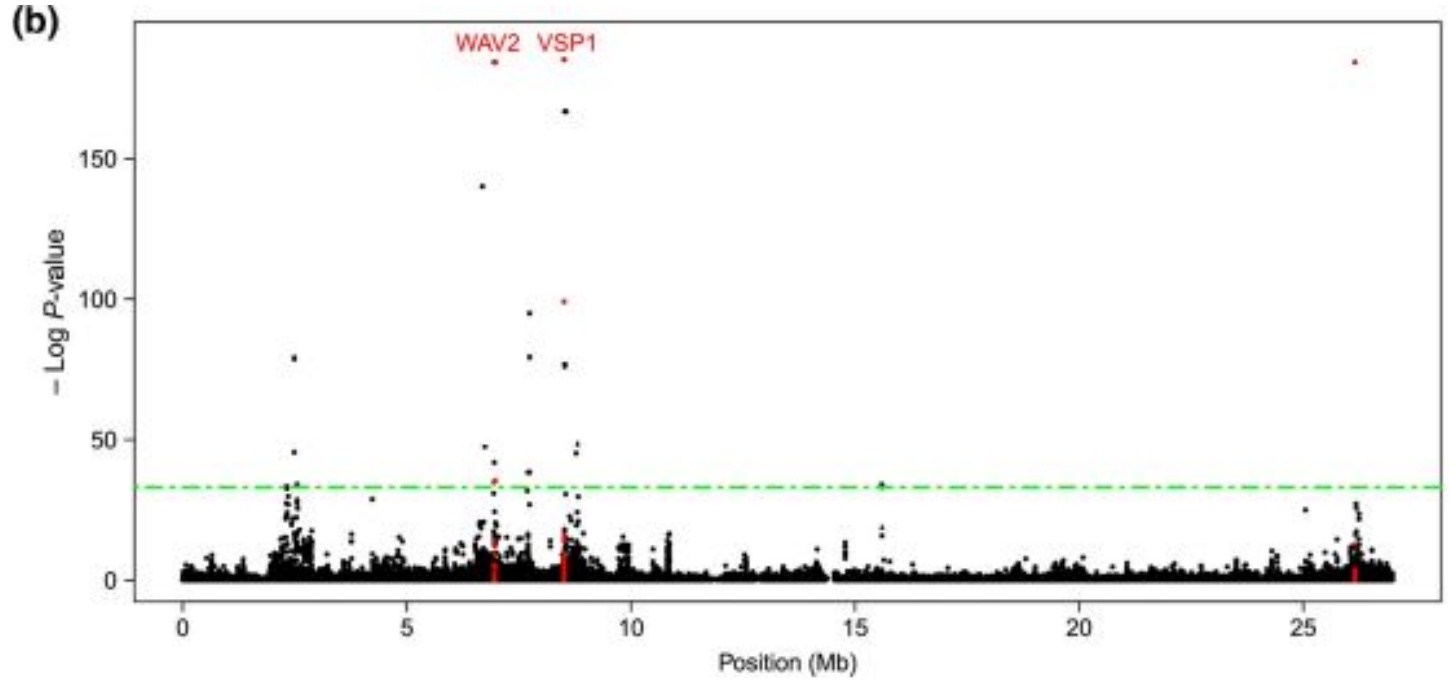
TESS3

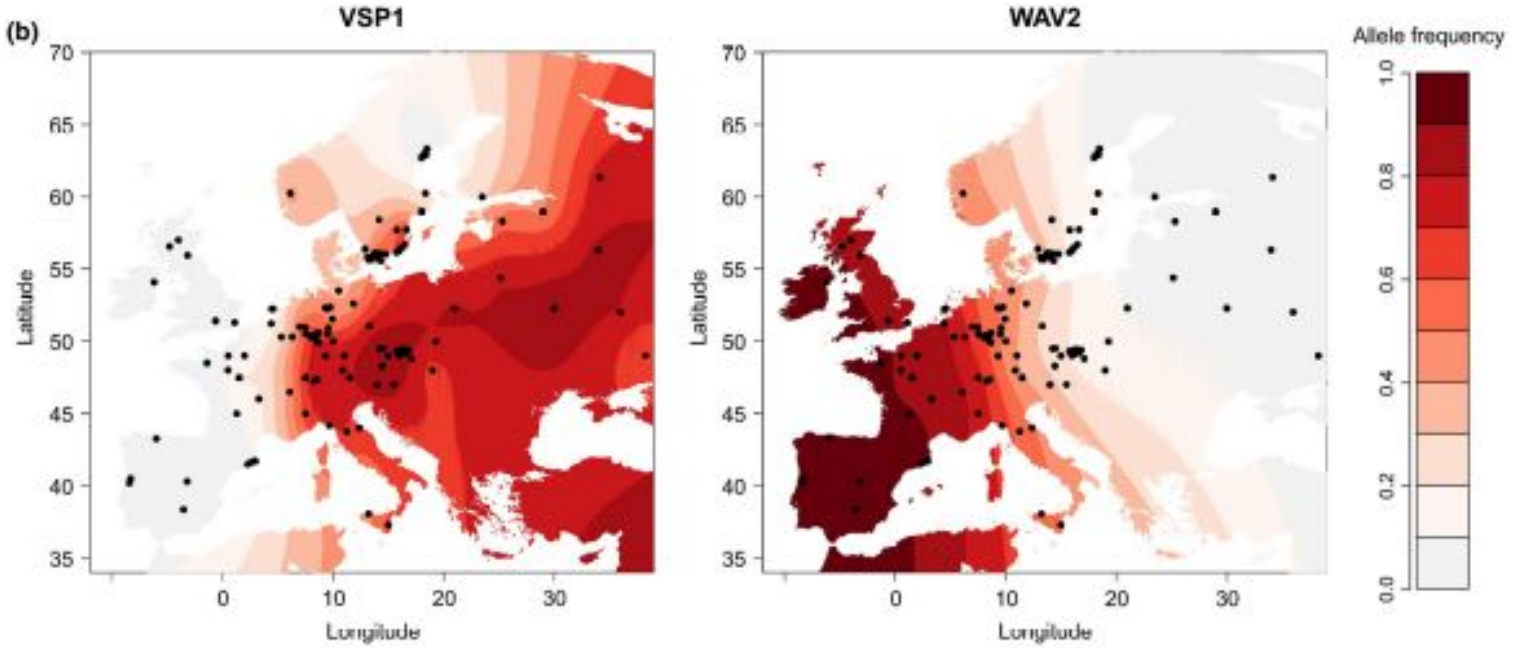
- Spatialized clustering analysis
- ADMIXTURE in space, but model-free
- Spatially close individuals have a higher prior probability for being related.



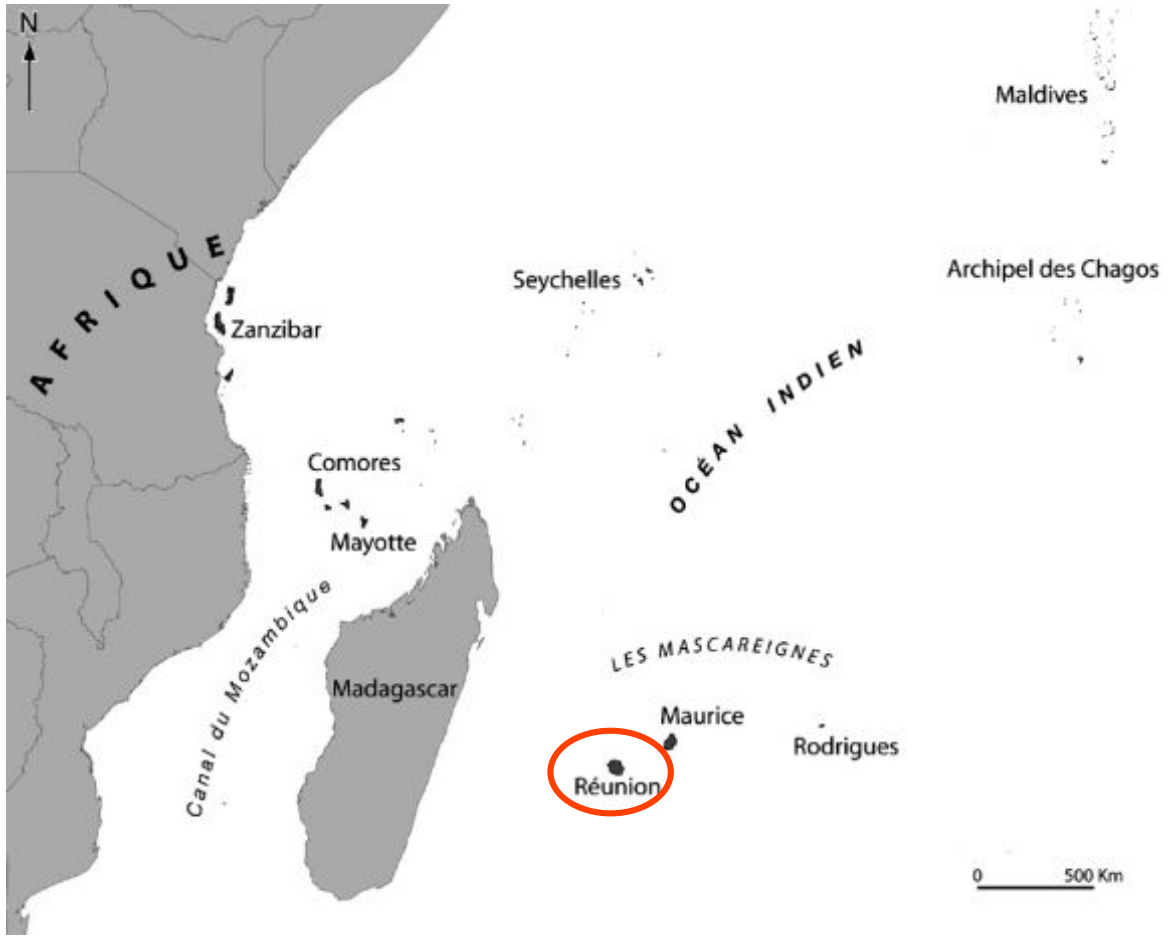
TESS3

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

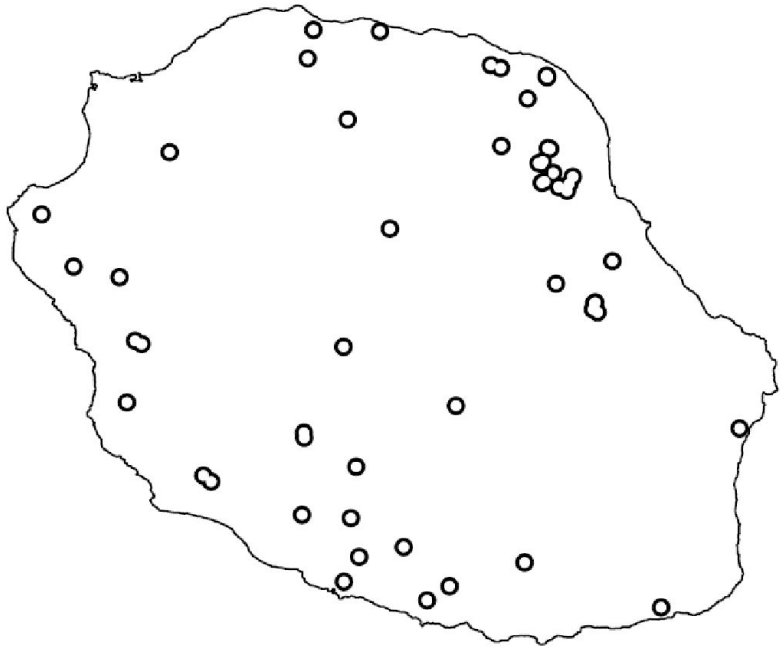


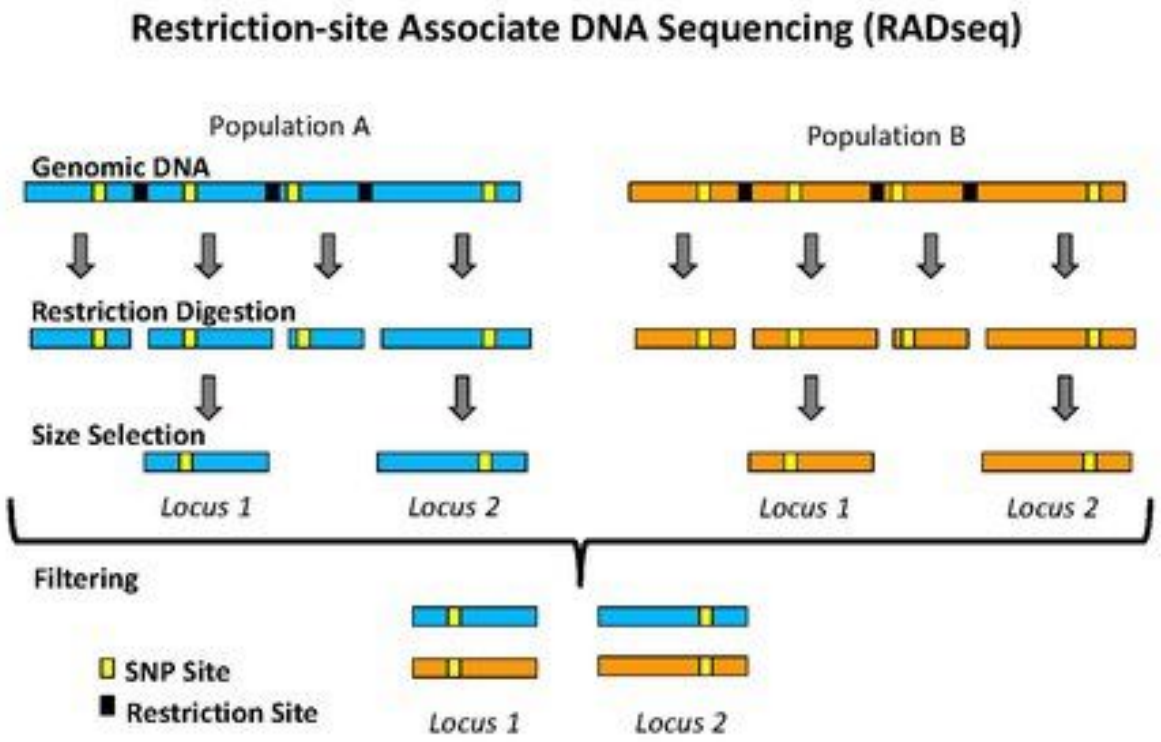


An application: The Réunion harrier (*Circus maillardi*)



Sampling and sequencing techniques



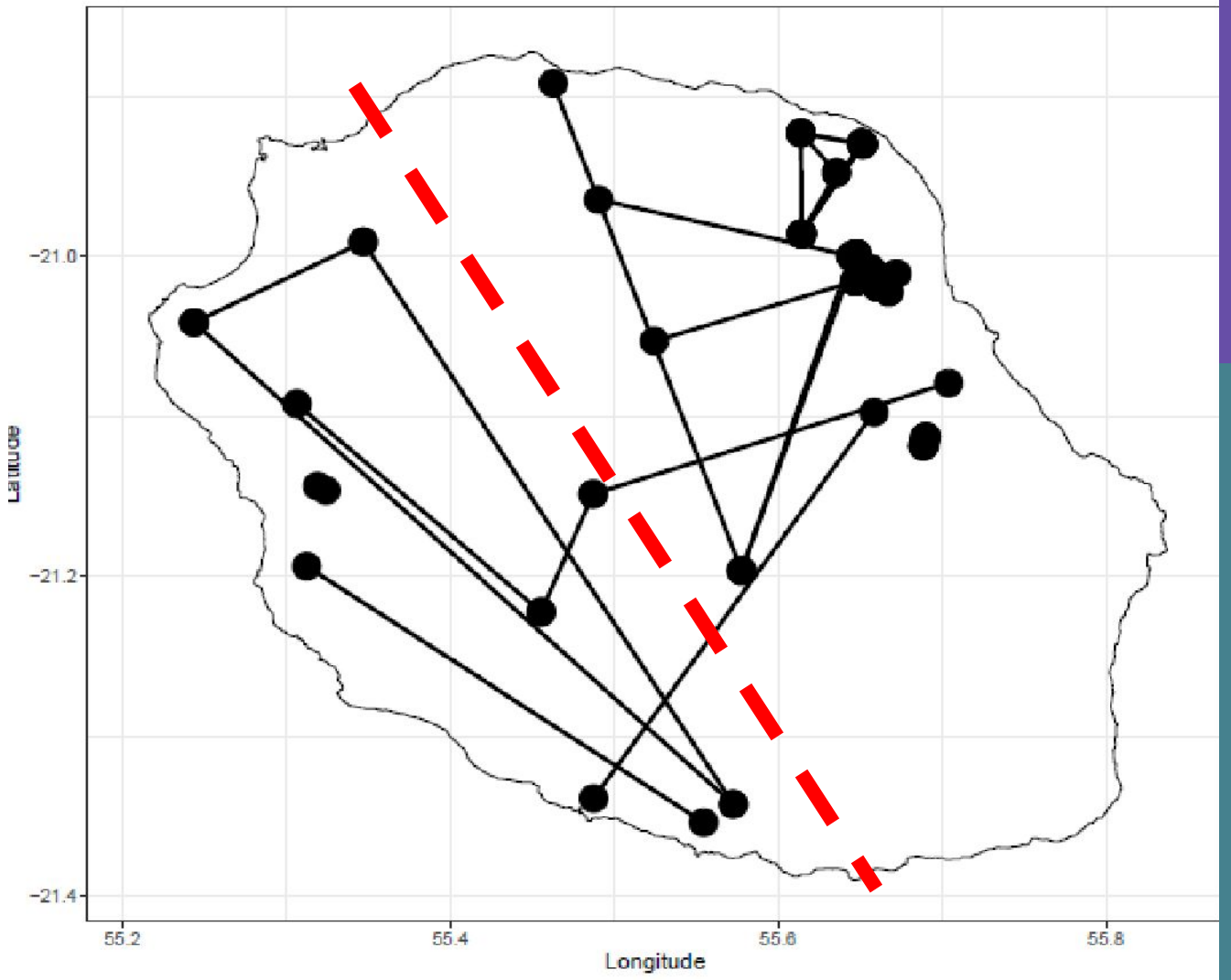


- GBS (*Genotyping by Sequencing*)
- Sequences loci around restriction sites > less sequencing effort
- About 6000 Single Nucleotide Polymorphisms.

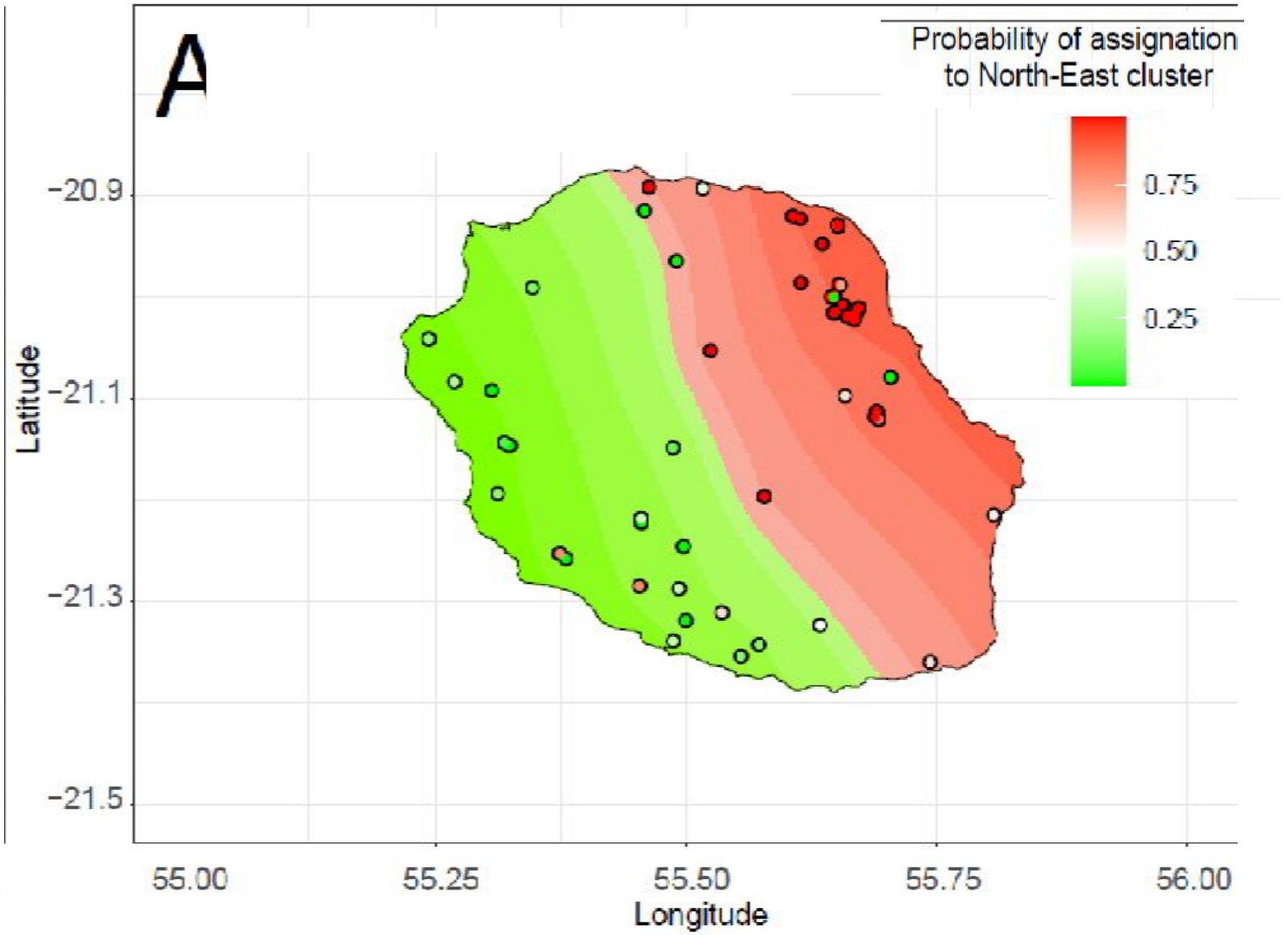
Relatedness and recent dispersal

- 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

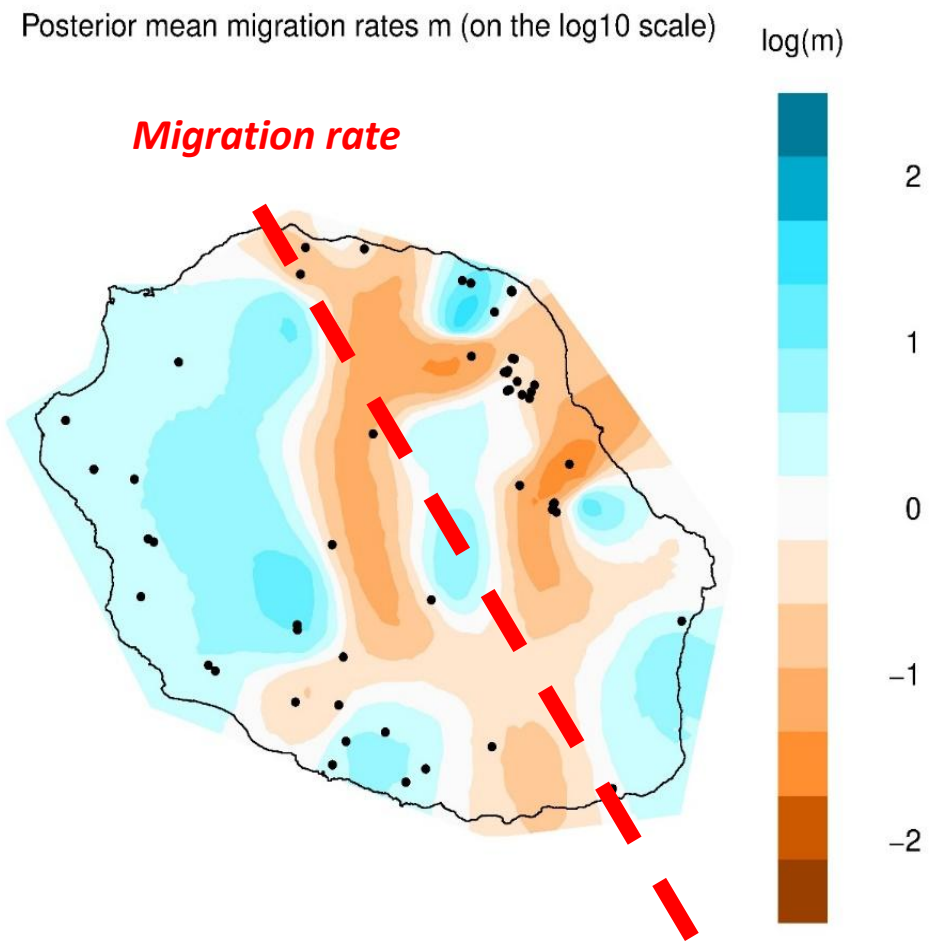


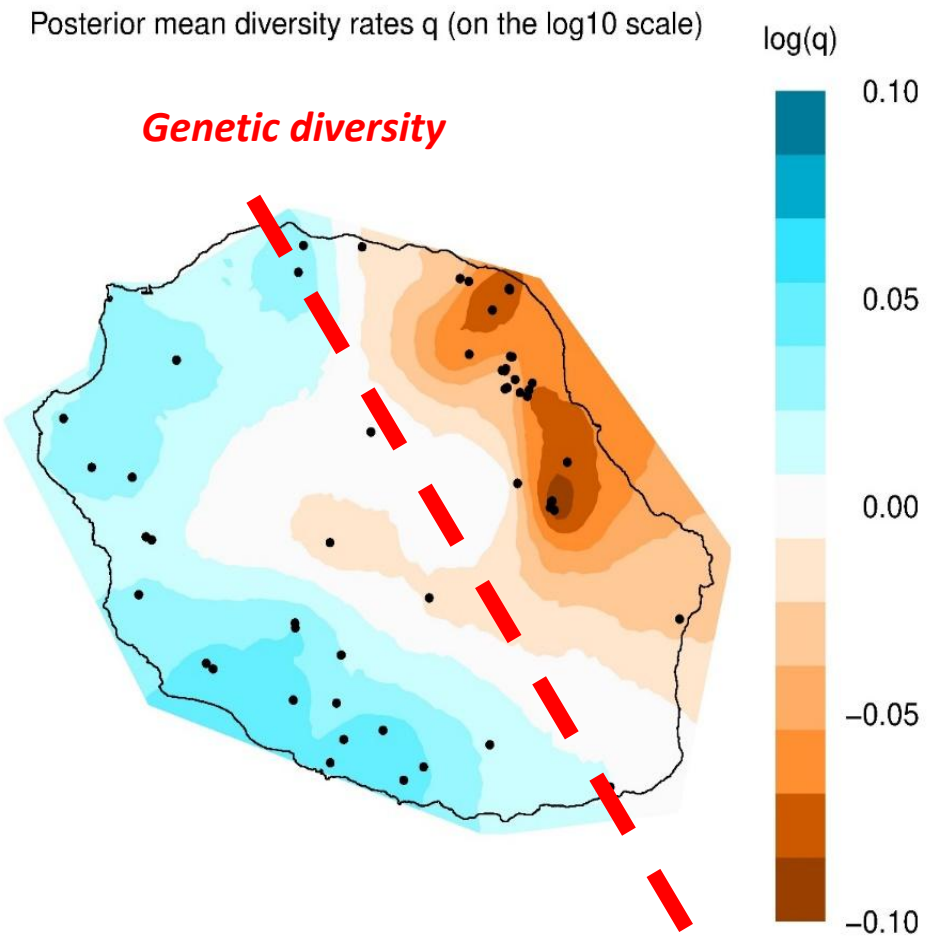
Population structure



Population structure

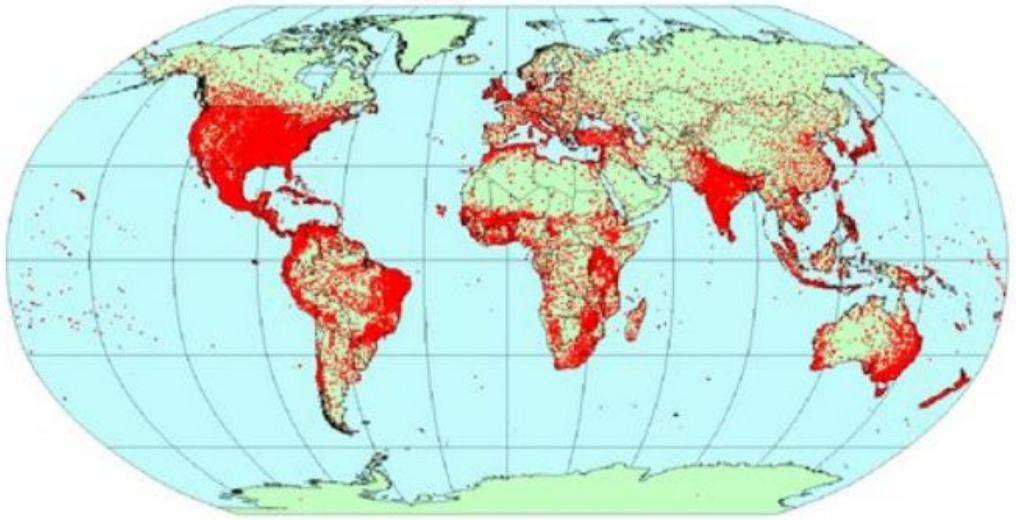
This is useful when studying waves of expansion or demography in space



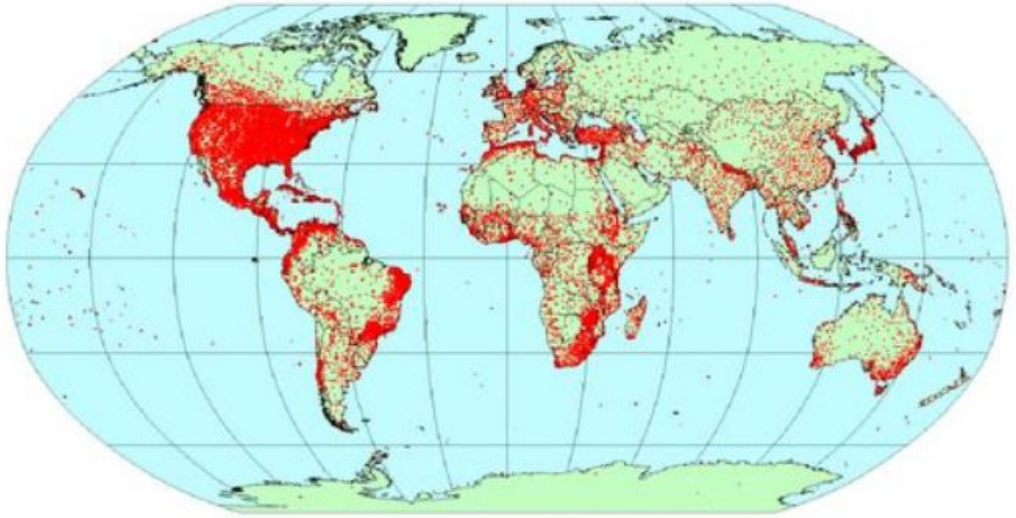


Including the environment: where can I find data?

- 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.



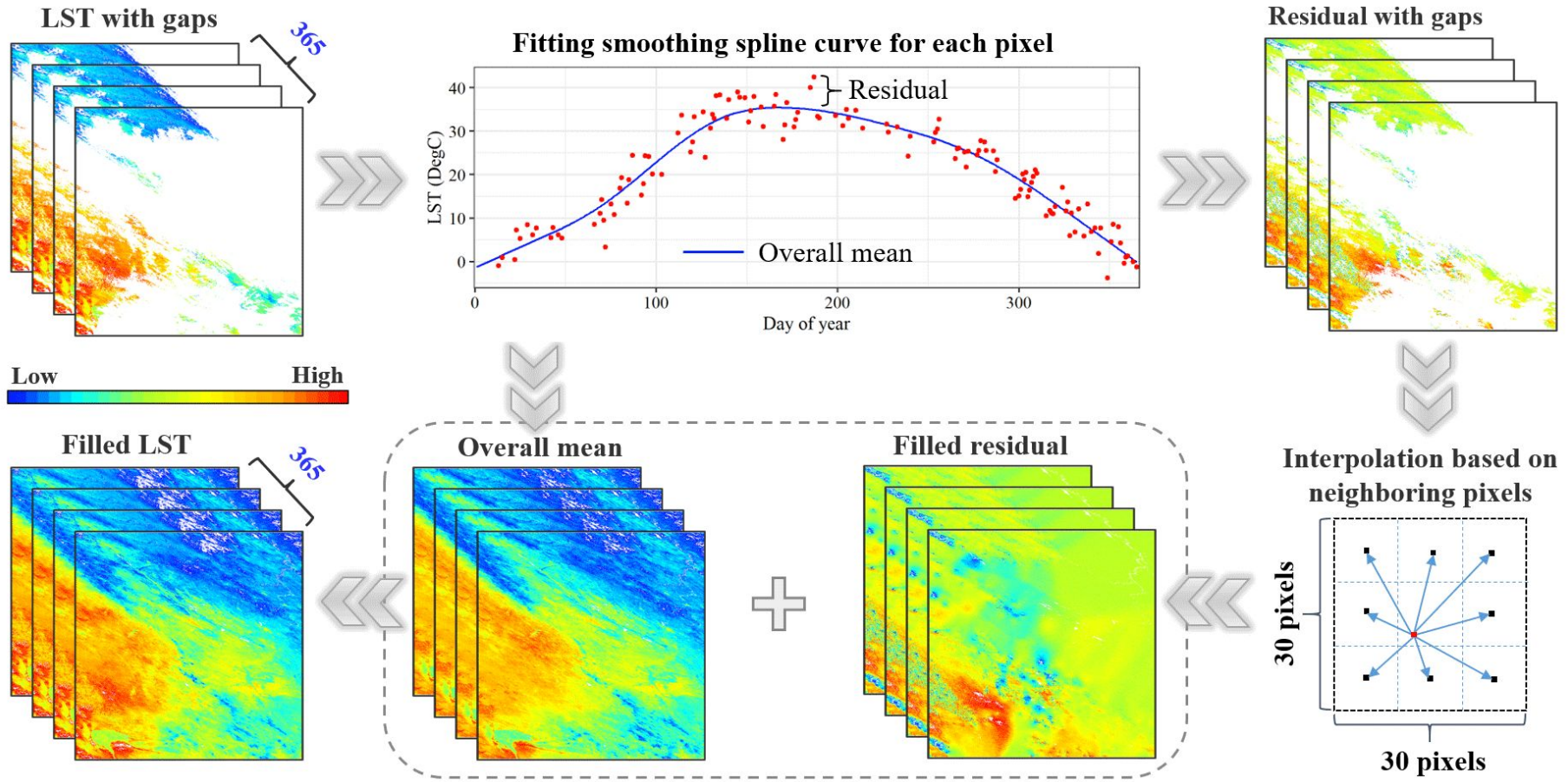
Locations of climate stations with mean temperature data.

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?

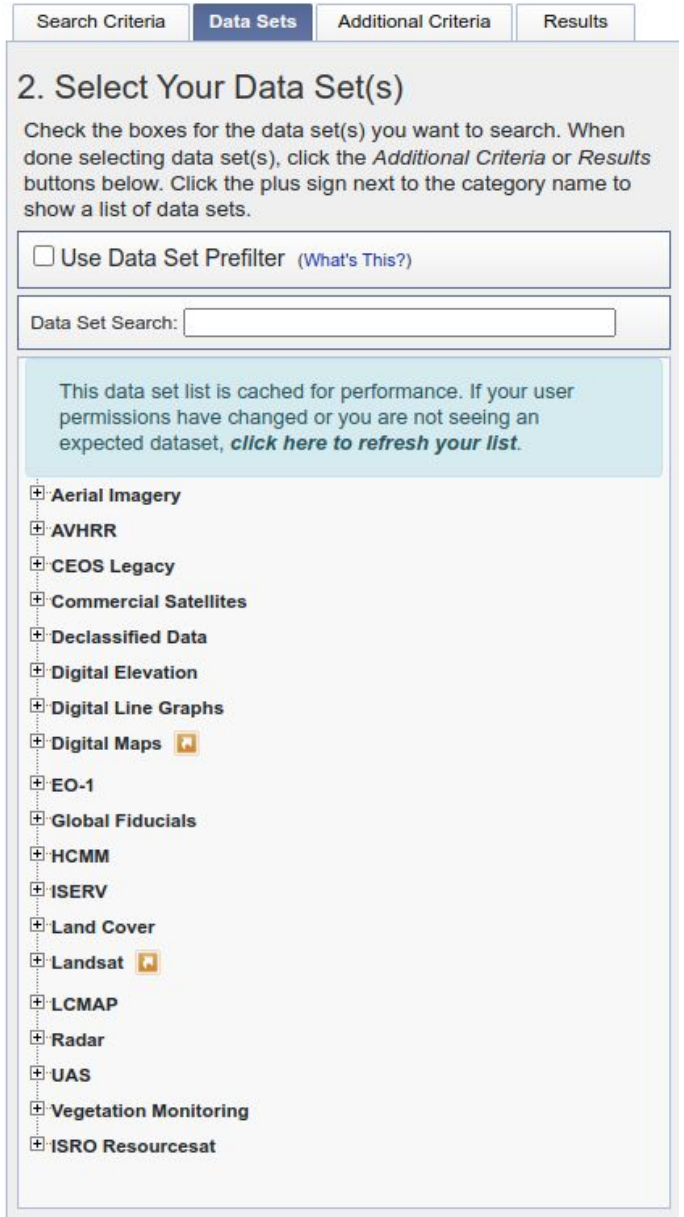
- Keep in mind that these data are obtained through **interpolation of residuals**: the accuracy of the values therefore depend on the underlying model and density of stations in the study area



An example of interpolation for missing data

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.



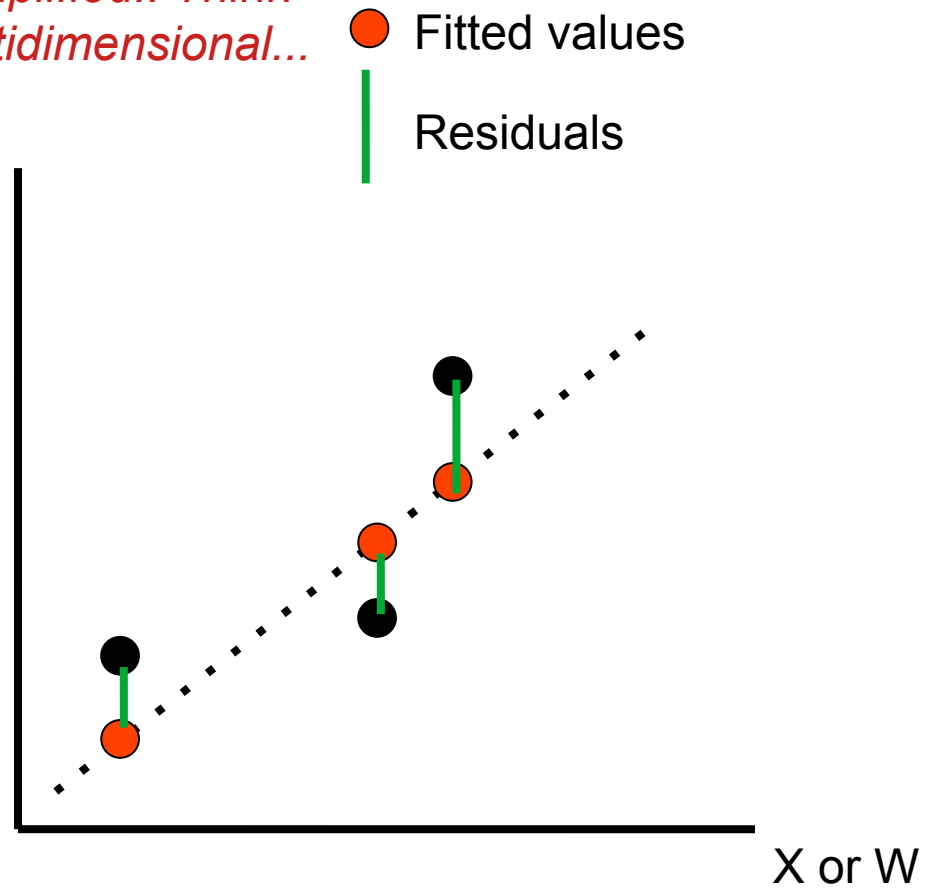
The environment: where can I find data?

- 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 (≈90m)	1 arcsec (≈30m)
DEM accuracy (stdev.)	±16m	±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)**

Simple redundancy analysis (RDA)

p loci | *m variables*

n indiv/pop: **Y** *Response* ~ **X** *Explanatory variables*

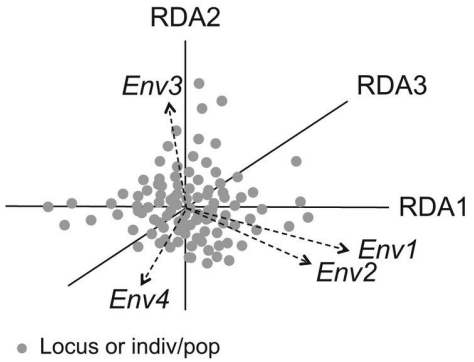
Here you can also inject dbMEM scores...

Linear regression

p loci

n indiv/pop: **Y′** *Fitted values*

PCA



Partial redundancy analysis (pRDA)

p loci | *m variables* | *z covariables*

n indiv/pop: **Y** *Response* ~ **X** *Explanatory variables* + **W** *Covariables*

Linear regression

p loci | *m variables*

n indiv/pop: **$Y_{res|W}$** *Residuals of Y ~ W* ~ **$X_{res|W}$** *Residuals of X ~ W*

Population structure

Redundancy analyses

- Can be used to decide which environmental variables matter:
- - First fit an empty model Y ~ 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

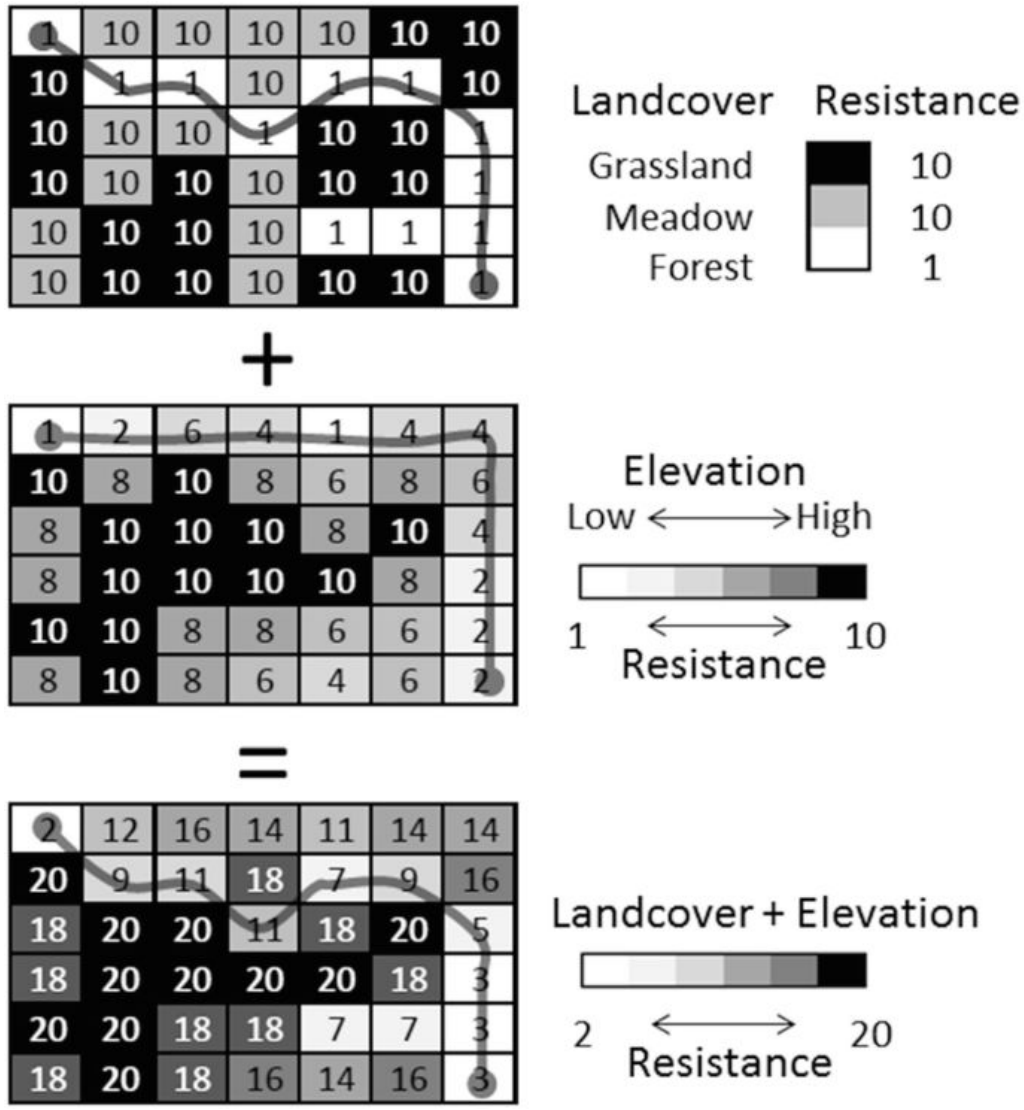
- Optimizes variance explained.
- What explains the most variance is not necessarily causal.
- Not popgen model-based, so deviations from population genetics expectations are not going to show.
- Default RDA expects a linear relationship between X and Y (but extensions can deviate from that).
- As usual: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

More on isolation by resistance

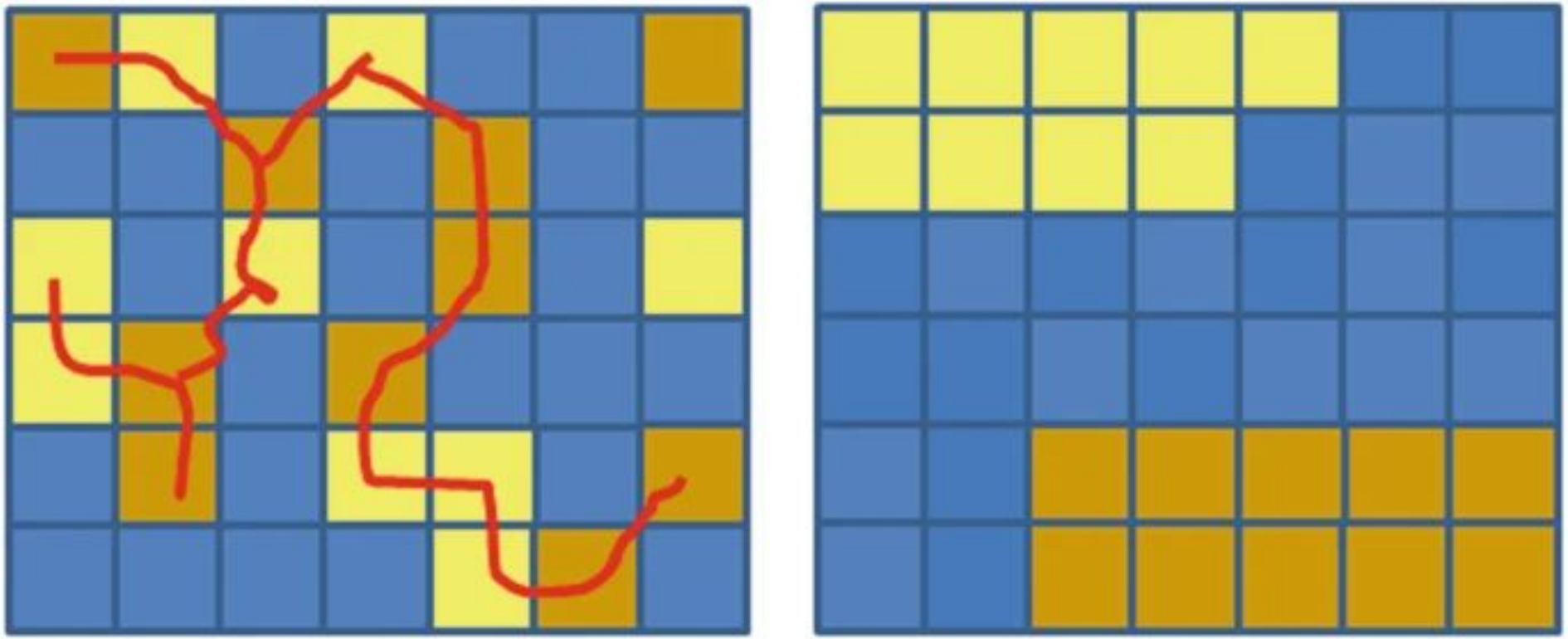
Create a combination of maps where each cell is given a resistance score depending on the feature it harbours

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

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

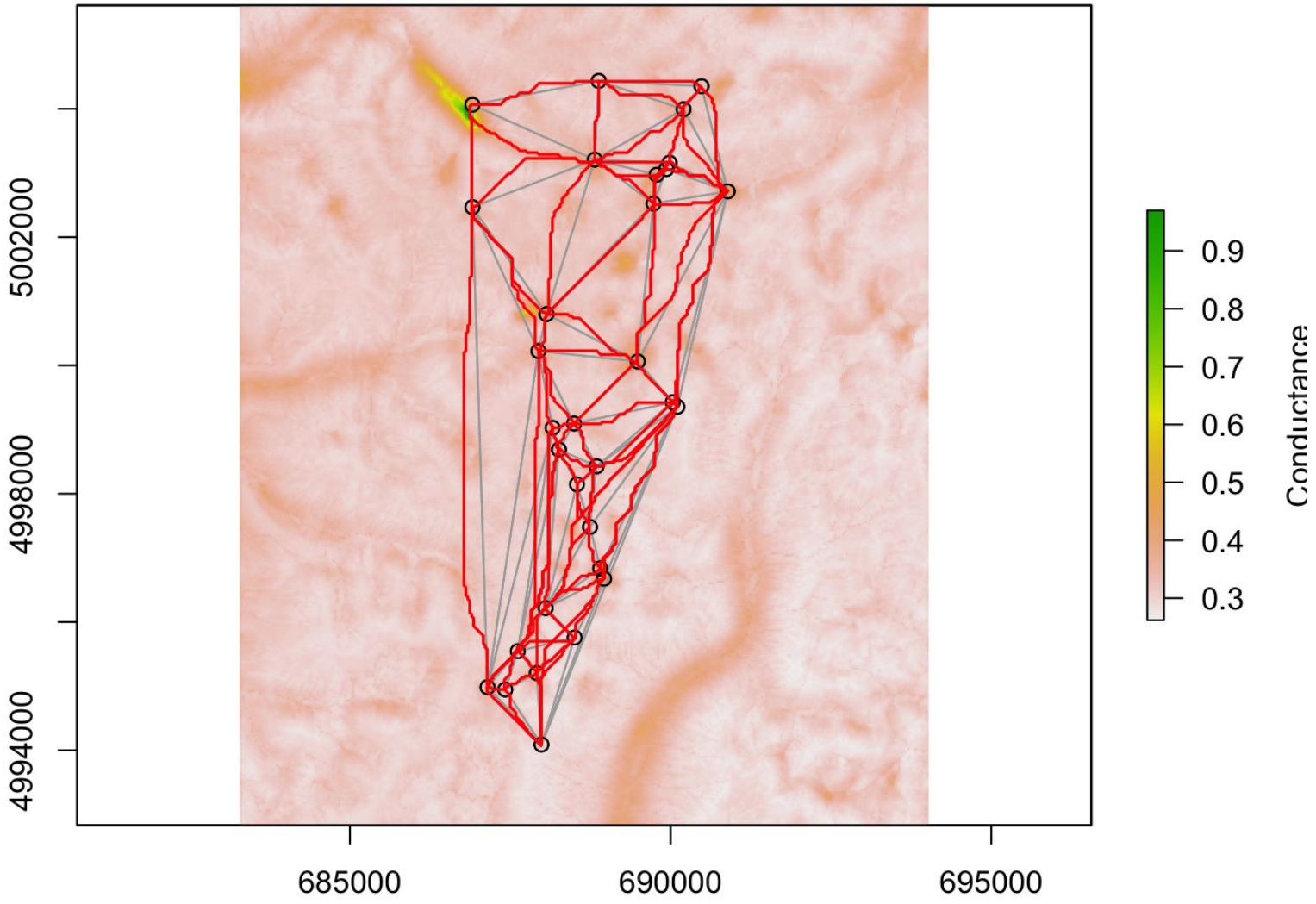


Functional connectivity != landscape fragmentation.



https://link.springer.com/chapter/10.1007/13836_2017_2

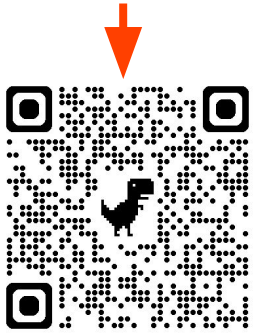
Euclidean distance and shortest paths

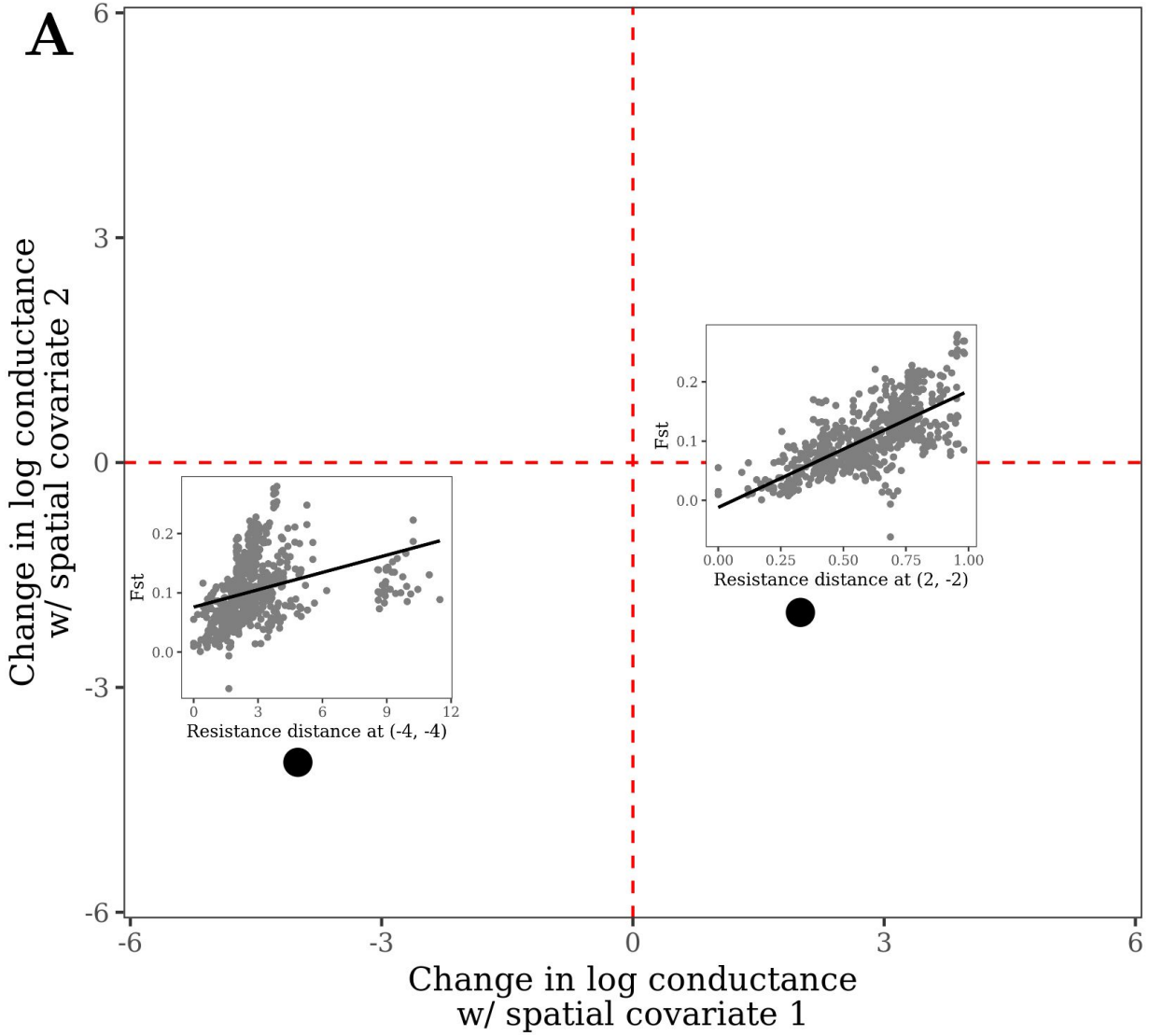


Testing isolation-by-resistance: how can we calculate resistance?

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.

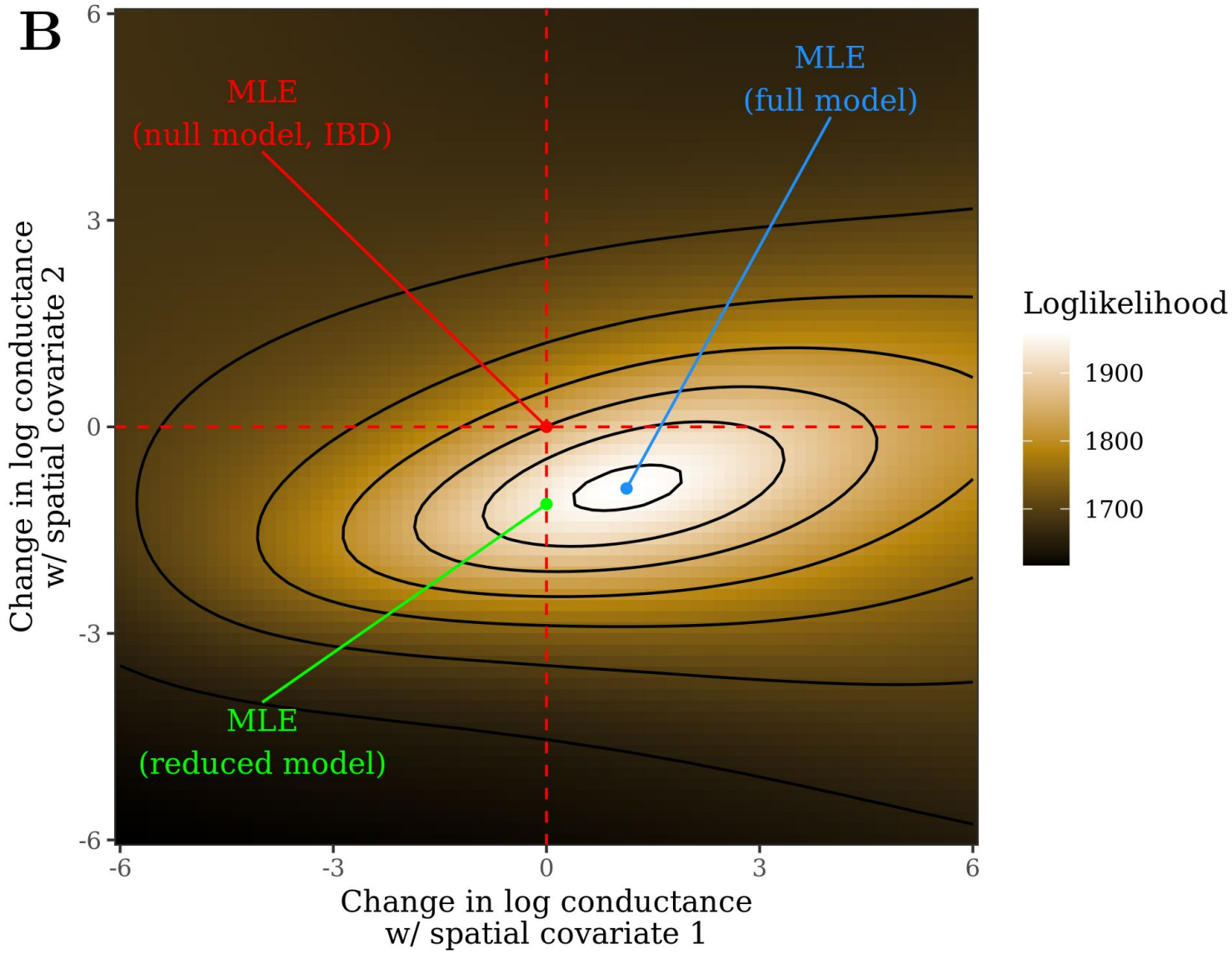
Excellent tutorial here:





Testing isolation-by-resistance

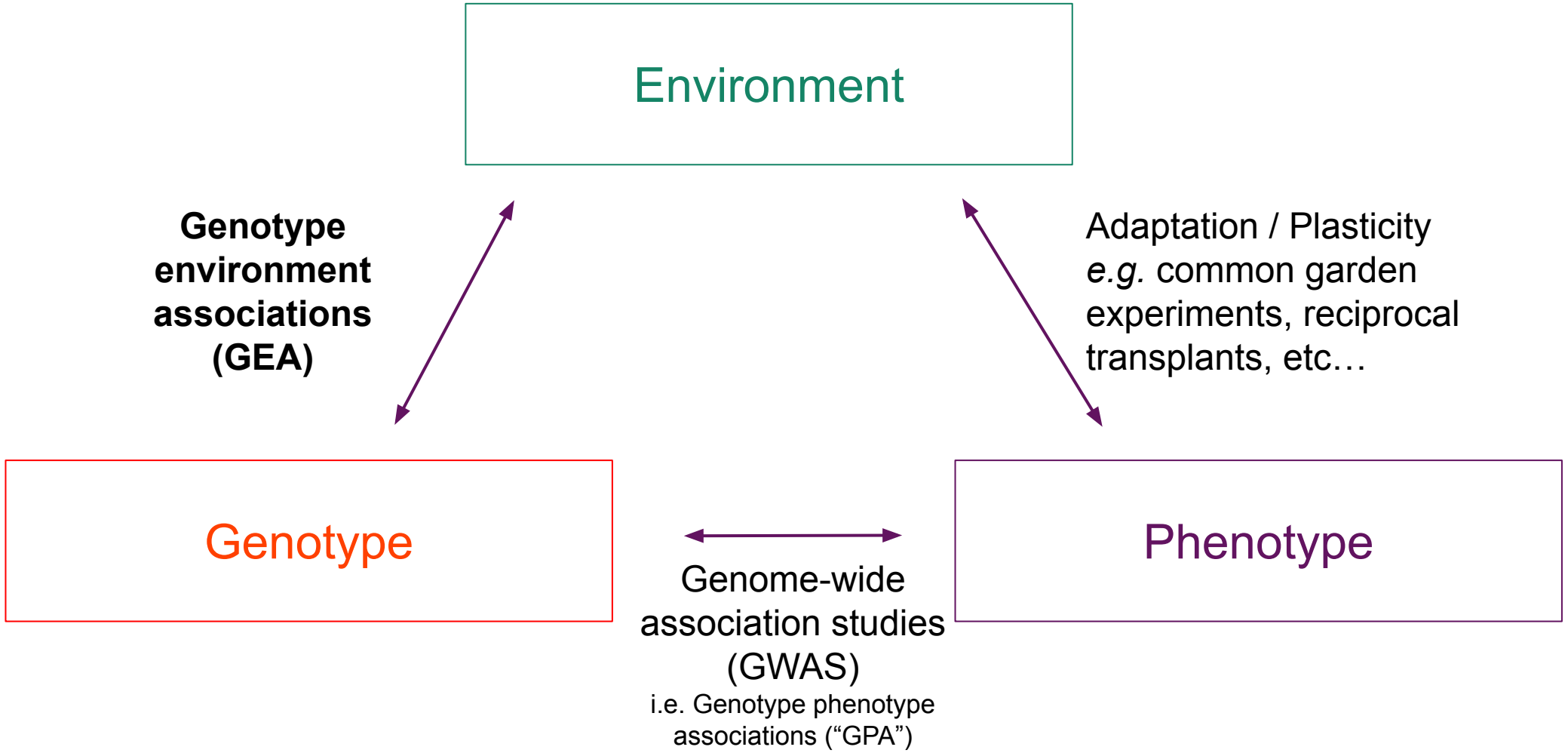
Step 4: Compare the likelihoods of different models
This last step is mathematically complex. The choice of optimizer in particular is still a matter for discussions.



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

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)

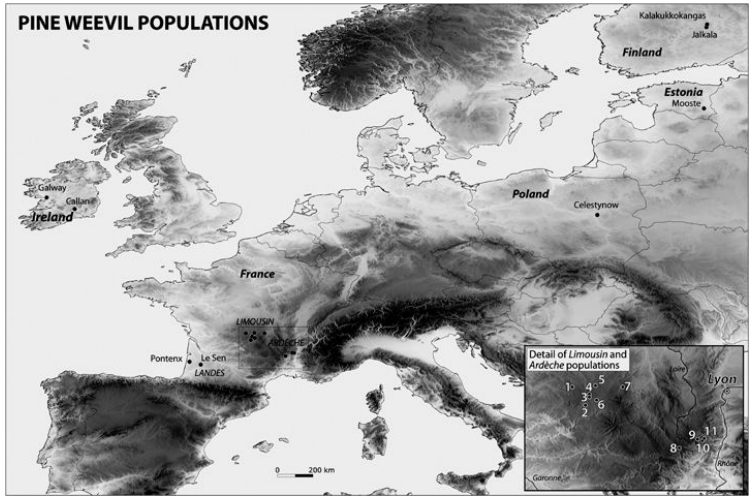
A quite recent field, enabled by the advances in genome sequencing made over the last one or two decades (massive and relatively low-cost sequencing), such as Illumina sequencing:

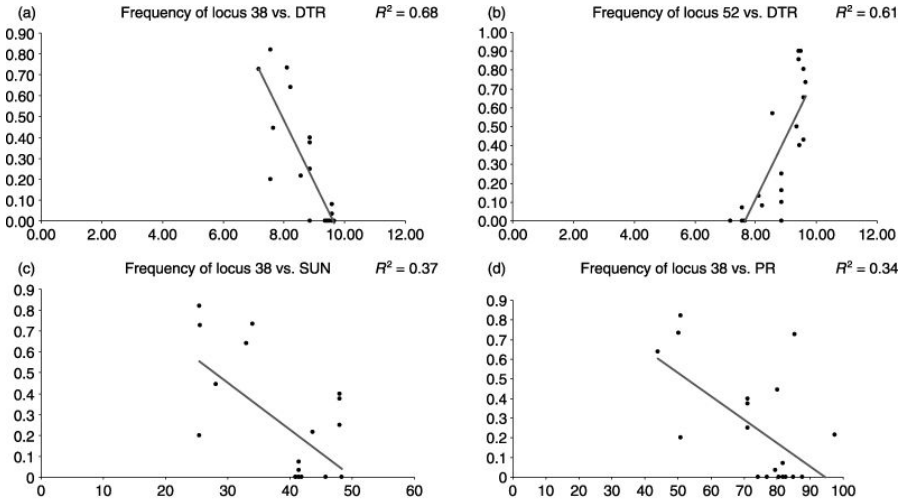
Sequencing of one or few populations -> sequencing of many pop covering different environments/climates

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





Joost et al. 2007 Mol Ecol

BayPass core model

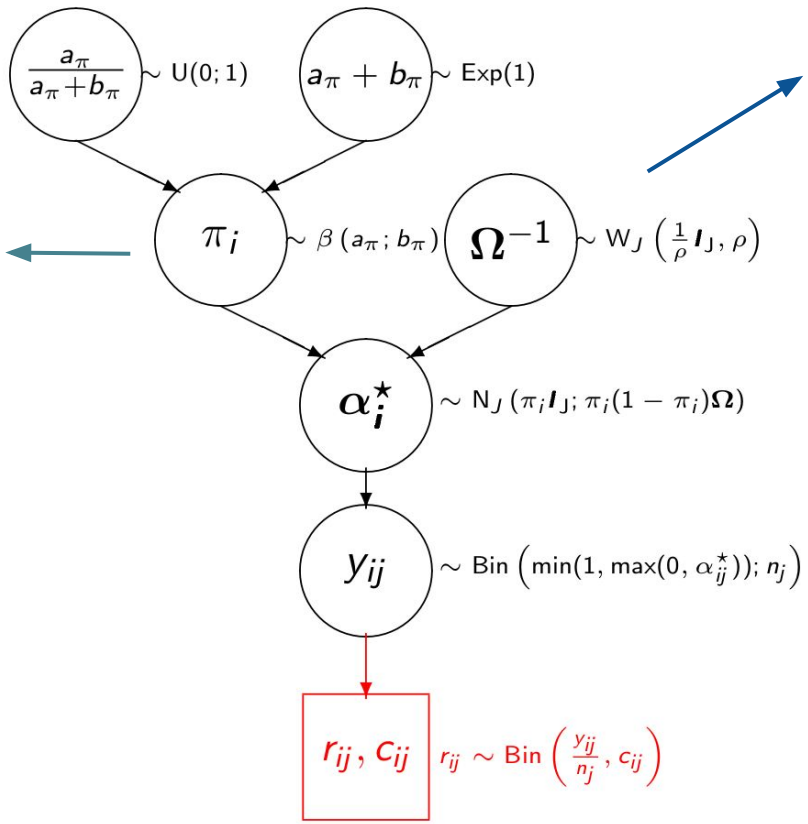
GENETICS | INVESTIGATION

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

Mathieu Gautier[1]

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)



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

Ancestral allele frequency (unobserved)

(allele counts at locus i for population j)

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

Gautier, 2015 Genetics

BayPass core model

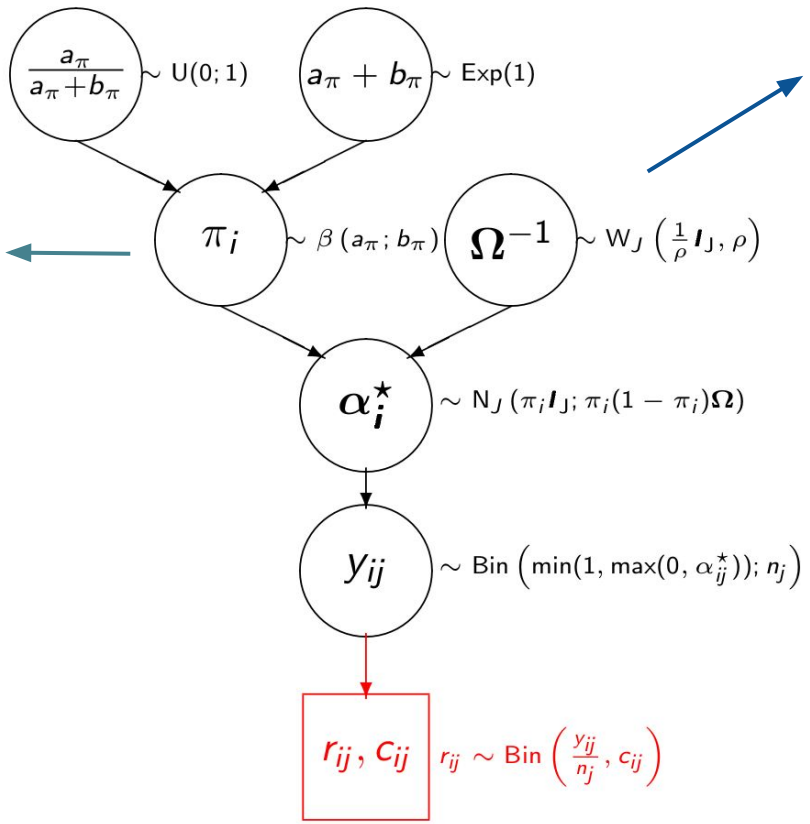
GENETICS | INVESTIGATION

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

Mathieu Gautier[1]

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)

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)

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

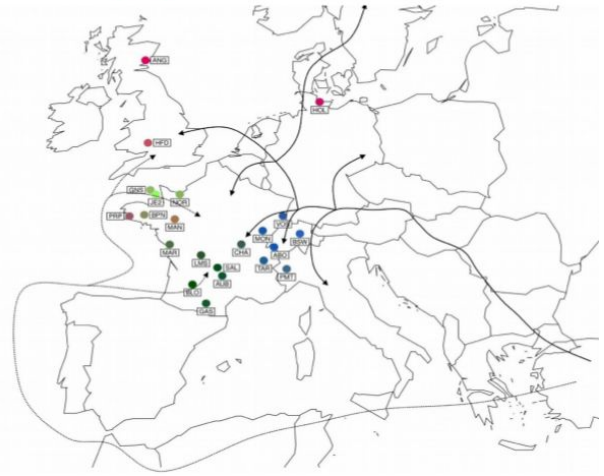
Gautier, 2015 Genetics

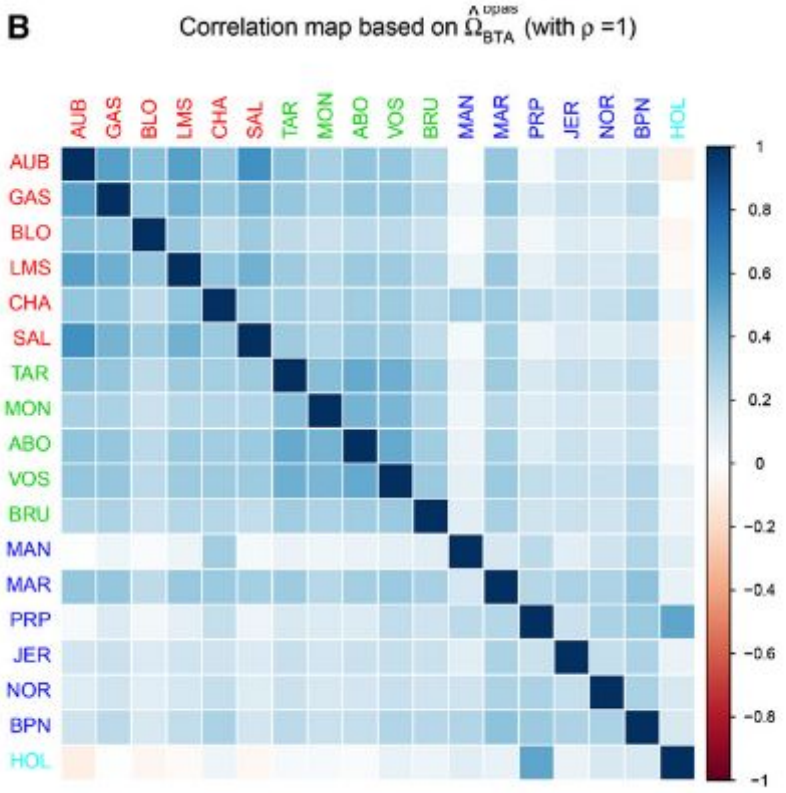
BayPass core model

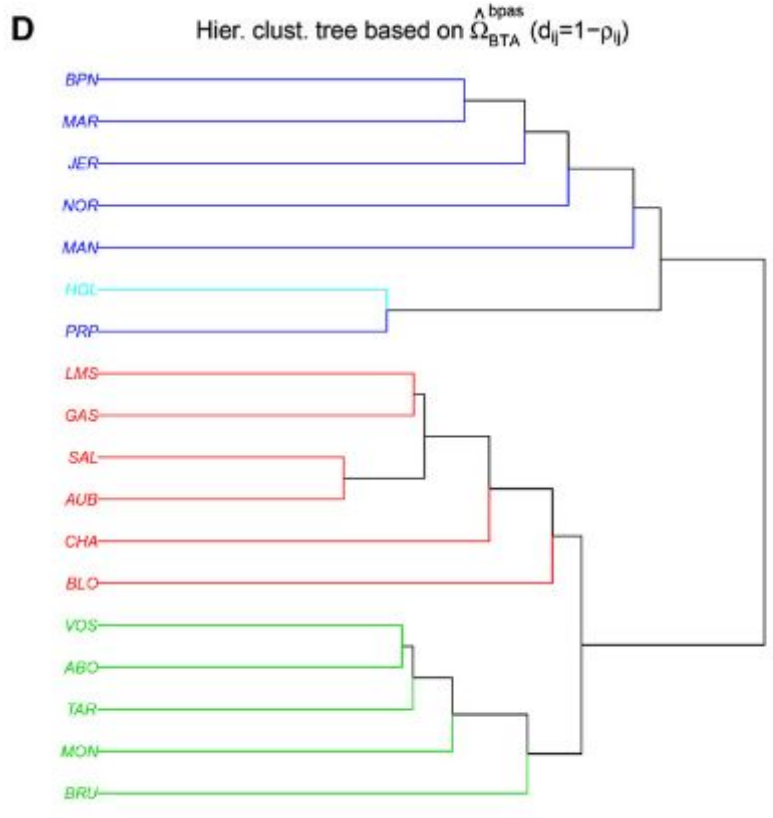
omega matrix (ω, among-pop covariance in allele frequencies) captures the overall population structure in the data

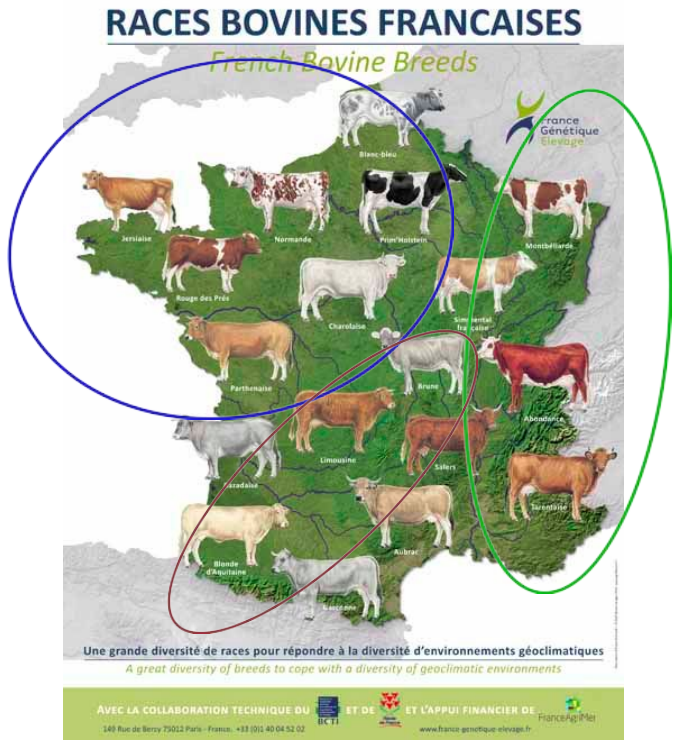
453 individuals from 18 French cattle breeds genotyped at 42,056 SNPs

18 French pops ; 42,046 SNPs (from Gautier et al., 2010)







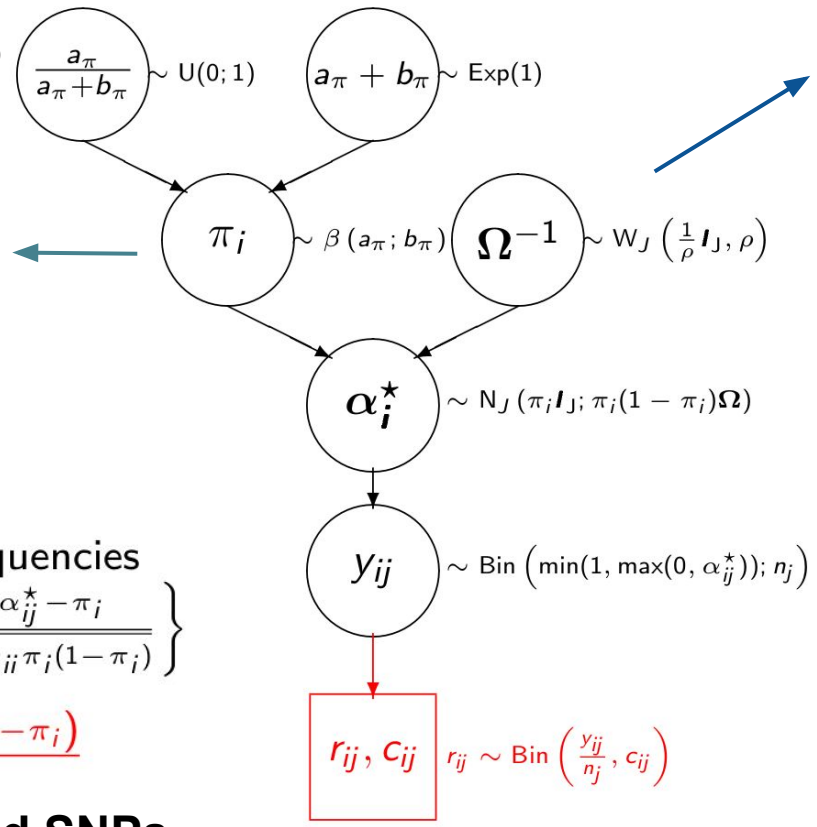


Gautier, 2015 Genetics

BayPass core model

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)



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

Ancestral allele frequency (unobserved)

XtX= SNP–specific "F_{ST}" corrected for population history (omega matrix)

$\mathbf{X_i} \simeq$ vector of scaled pop. allele frequencies

e.g., if $\boldsymbol{\Omega}$ *diagonal (i.e.,* $\omega_{i\neq j} = 0$*),* $\mathbf{X_i} = \left\{ \frac{\alpha_{ij}^\star - \pi_i}{\sqrt{\omega_{ii}\pi_i(1-\pi_i)}} \right\}$

$$X^t X_i = \mathrm{Var}(\mathbf{X_i}) = \frac{(\boldsymbol{\alpha}_i^\star - \pi_i)\boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha}_i^\star - \pi_i)}{\pi_i(1-\pi_i)}$$

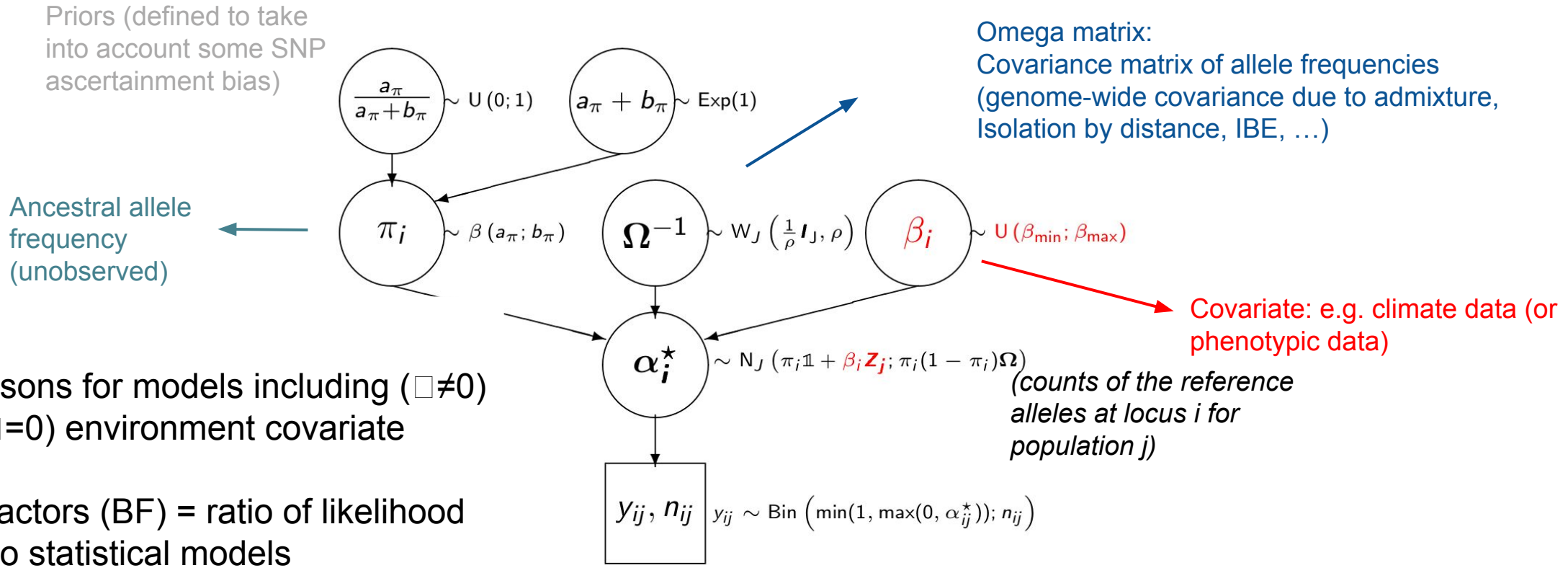
High XtX = Highly differentiated SNPs

(counts of the reference alleles at locus i for population j)

Günther & Coop, 2013 Genetics; Gautier, 2015 Genetics

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



Comparisons for models including (□≠0) or not (□=0) environment covariate

Bayes Factors (BF) = ratio of likelihood of the two statistical models
Higher BF = higher support for GEA

Gautier, 2015 Genetics

BayPass core model & detection of GEA

OXFORD

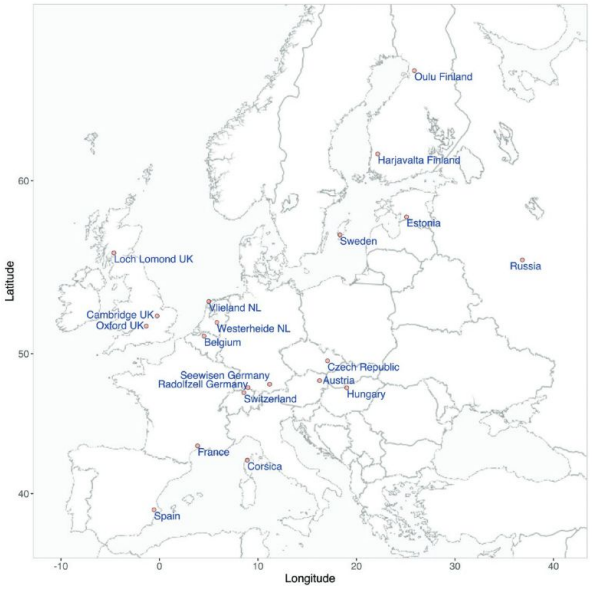
Evolution Letters, 2023, XX(XX), 1–11
https://doi.org/10.1093/evlett/qrad043
Letter

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

Joanne C. Stonehouse[1], Lewis G. Spurgin[2], Veronika N. Laine[3,4], Mirte Bosse[5,6], The Great Tit HapMap Consortium, Martien A.M. Groenen[5], Kees van Oers[3], Ben C. Sheldon[7], Marcel E. Visser[3], Jon Slate[1]

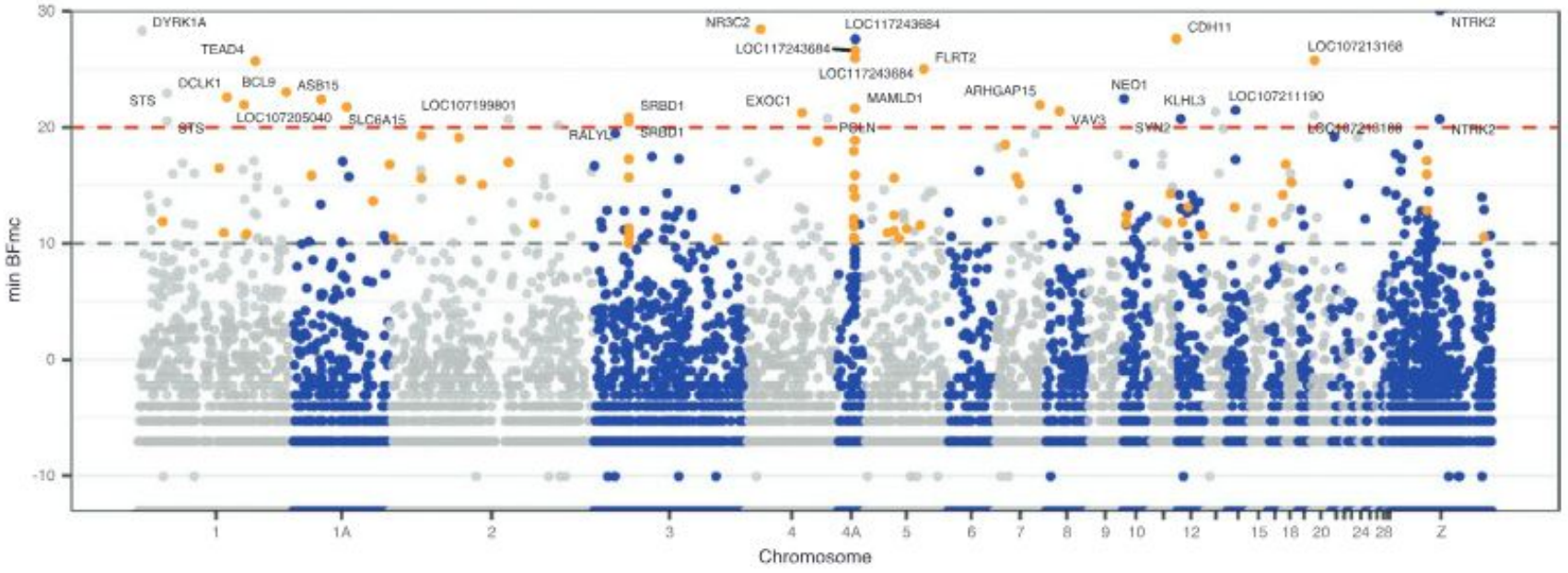
[1]School of Biosciences, University of Sheffield, Sheffield, United Kingdom
[2]School of Biological Sciences, University of East Anglia, Norwich Research Park, Norwich, United Kingdom
[3]Department of Animal Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands
[4]Finnish Museum of Natural History, University of Helsinki, Helsinki, Finland
[5]Animal Breeding and Genomics, Wageningen University & Research, Wageningen, The Netherlands
[6]Amsterdam Institute for Life and Environment (A-LIFE), Section Ecology and Evolution, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands
[7]Edward 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: jslate@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 adaptations 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

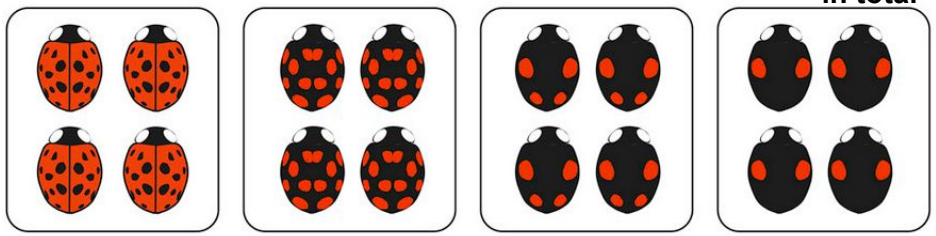
Report

Current Biology

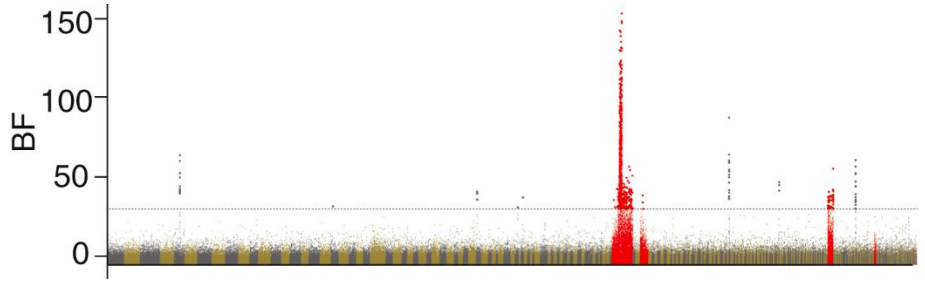
The Genomic Basis of Color Pattern Polymorphism in the Harlequin Ladybird

Pooled-DNA samples from worldwide populations

12 pools in total



Genome-wide association study with color form as covariate



Gautier et al. 2018 Current Biology

OPEN

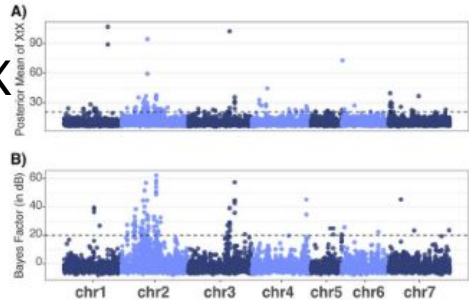
Cost-effective detection of genome-wide signatures for 2,4-D herbicide resistance adaptation in red clover

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

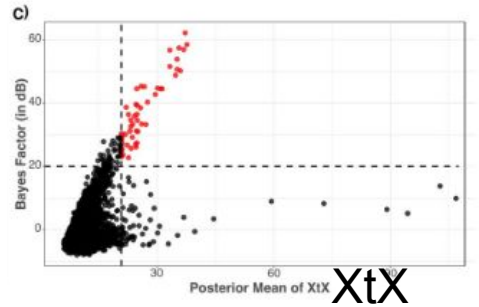
categorical pool-specific covariable (R/S)

XtX

BF



BF



XtX

GEA & Genomic offsets

Back to GEA!

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

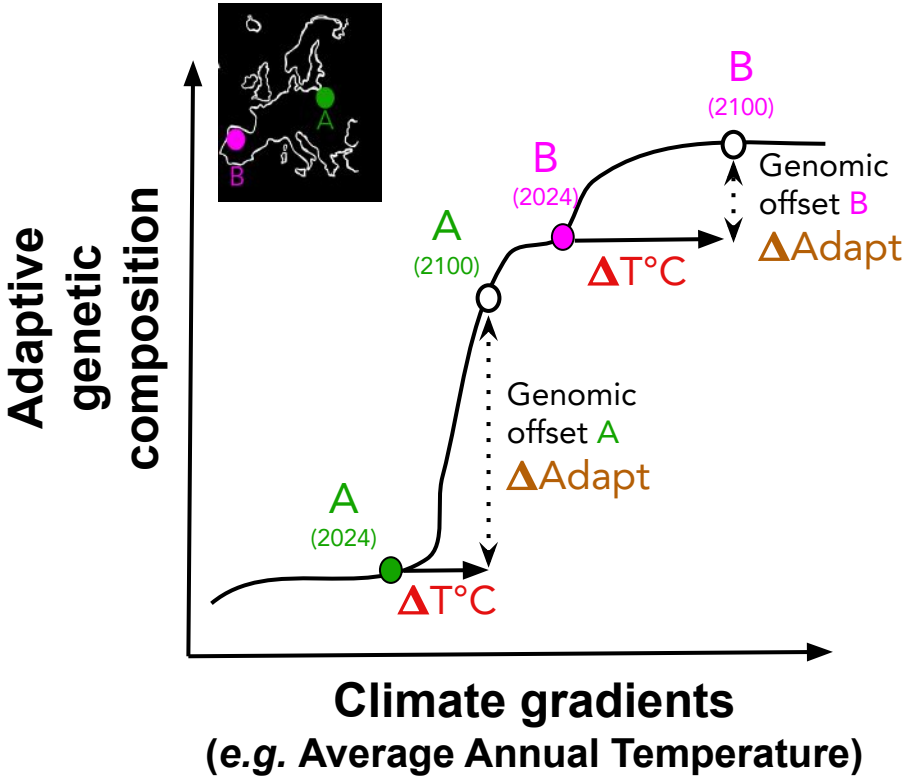
- → Assumes that you have uncovered genetic basis of adaptation
  - → Set of adaptive loci (knowledge of allele frequency at adaptive loci in many pops)
- → Assumes that you have precise information about how the environment is expected to change

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)

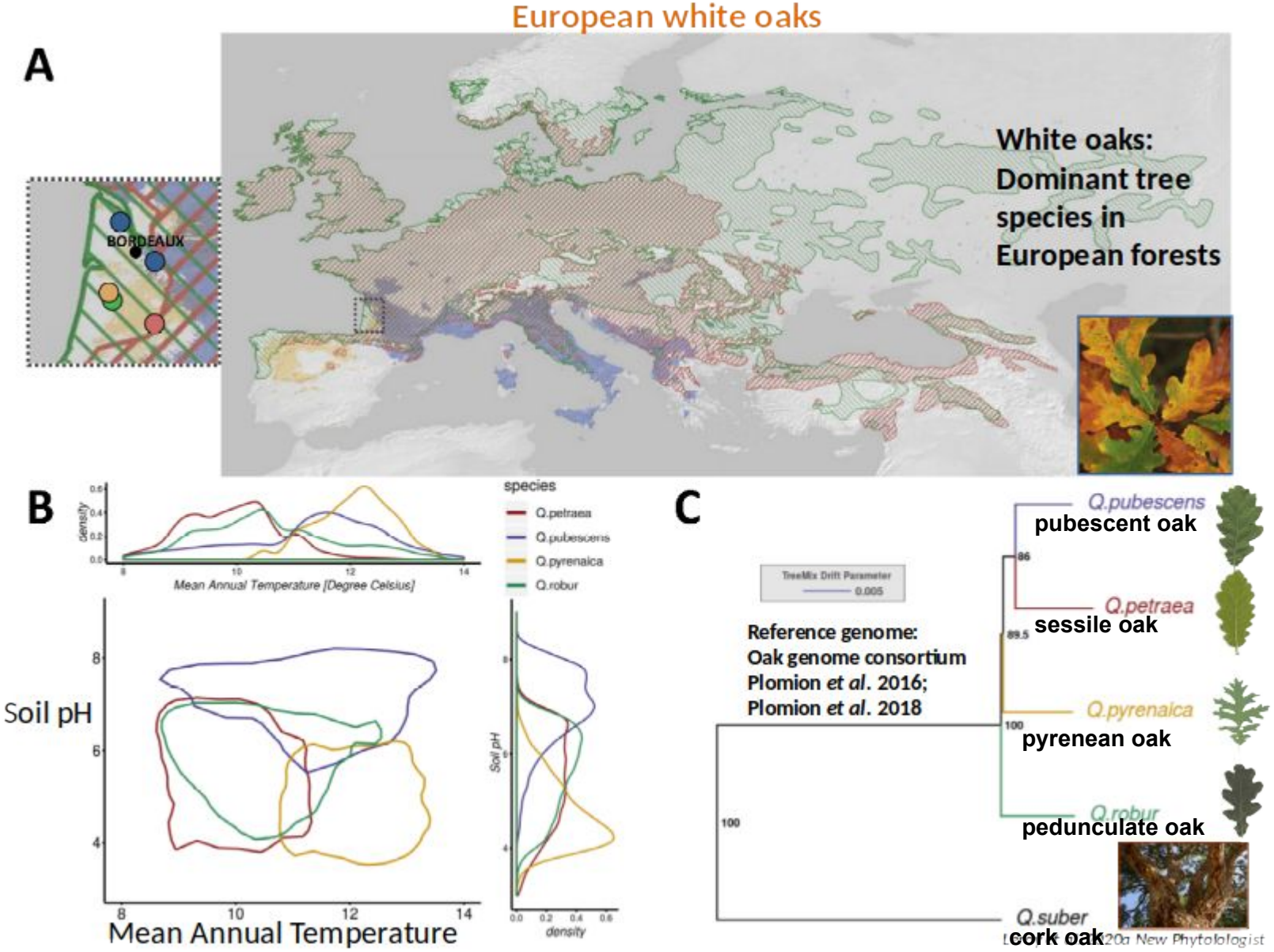
WorldClim METEO FRANCE ipcc

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…

Introducing the practical: European white oaks



Leroy et al. 2020a New Phytologist

Introducing the practical: European white oaks

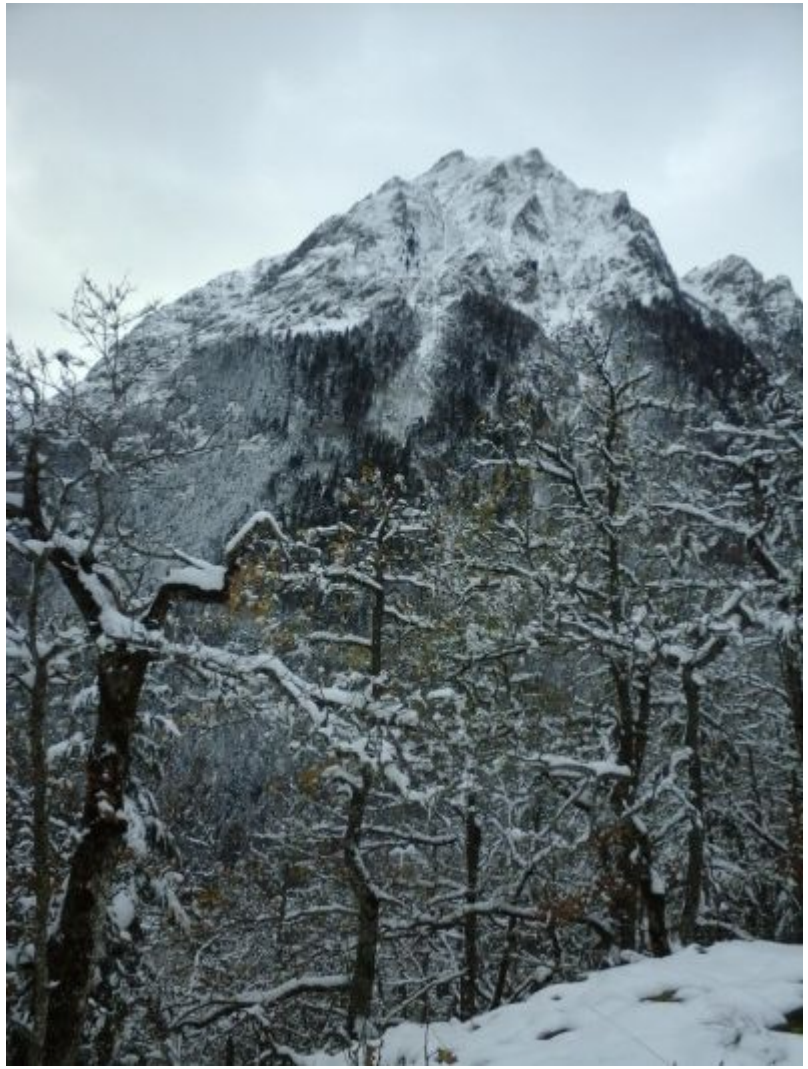
- Genomic data (Pool-seq):

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)

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

- Phenotypic data:
leaf unfolding in common gardens



Leroy et al. 2020b New Phytologist

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	Killarney, 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	Göhrde, 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.

Physalia

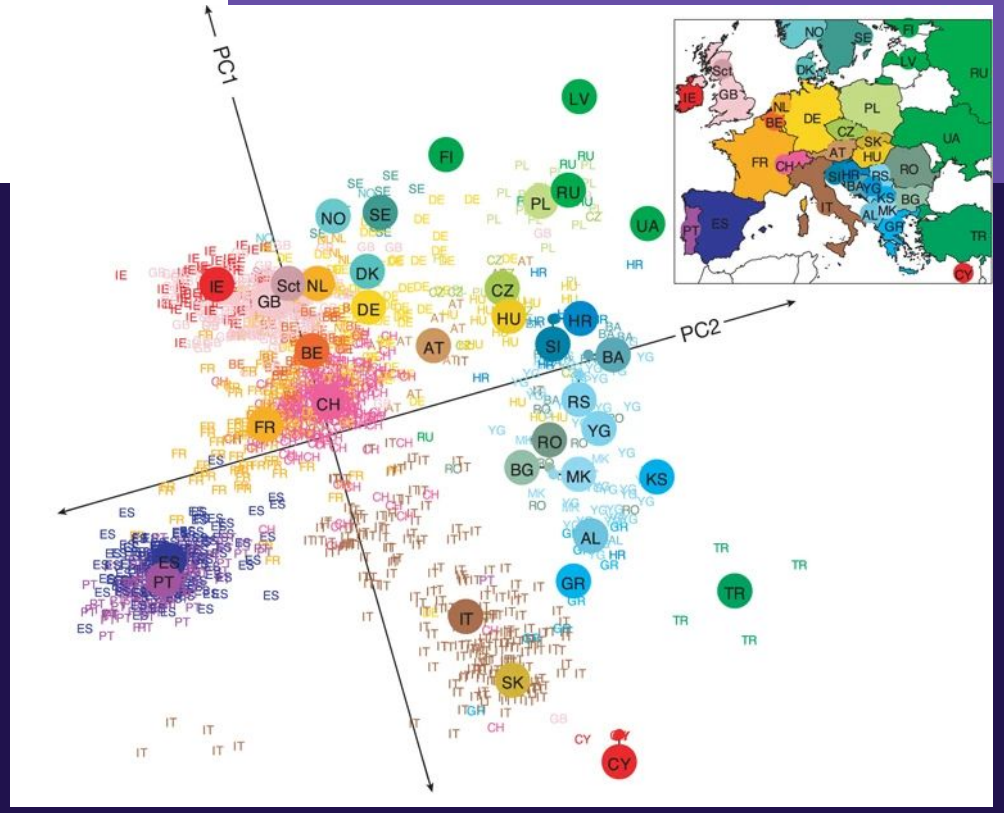
Courses

Landscape genomics

29/11/2024

Physalia course

Yann Bourgeois, Thibault Leroy



Physalia

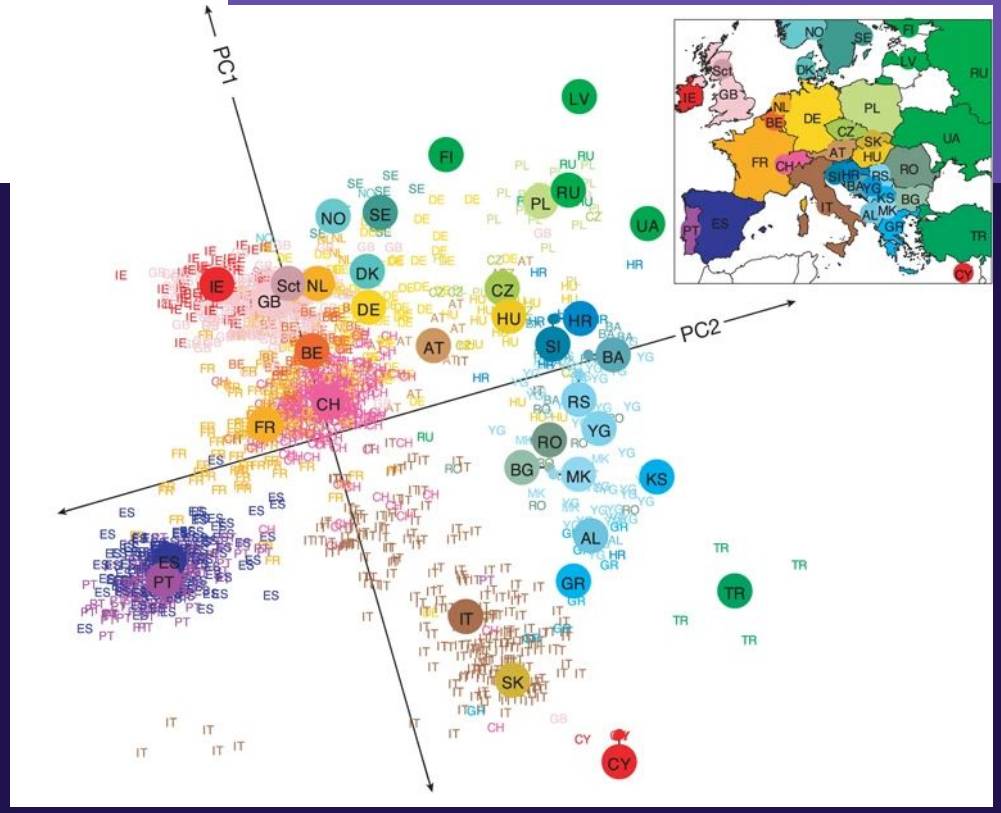
Courses

Landscape genomics Recap - Practical

29/11/2024

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

```
Sc00000003338 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
Sc00000004078 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
```

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(50,50,50,50,50,44,50,50,50,50,40,40,40,40,40,40,20,36),
+ 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÷213319 ~ **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

Leroy et al. 2020b New Phytologist

Map & climate data

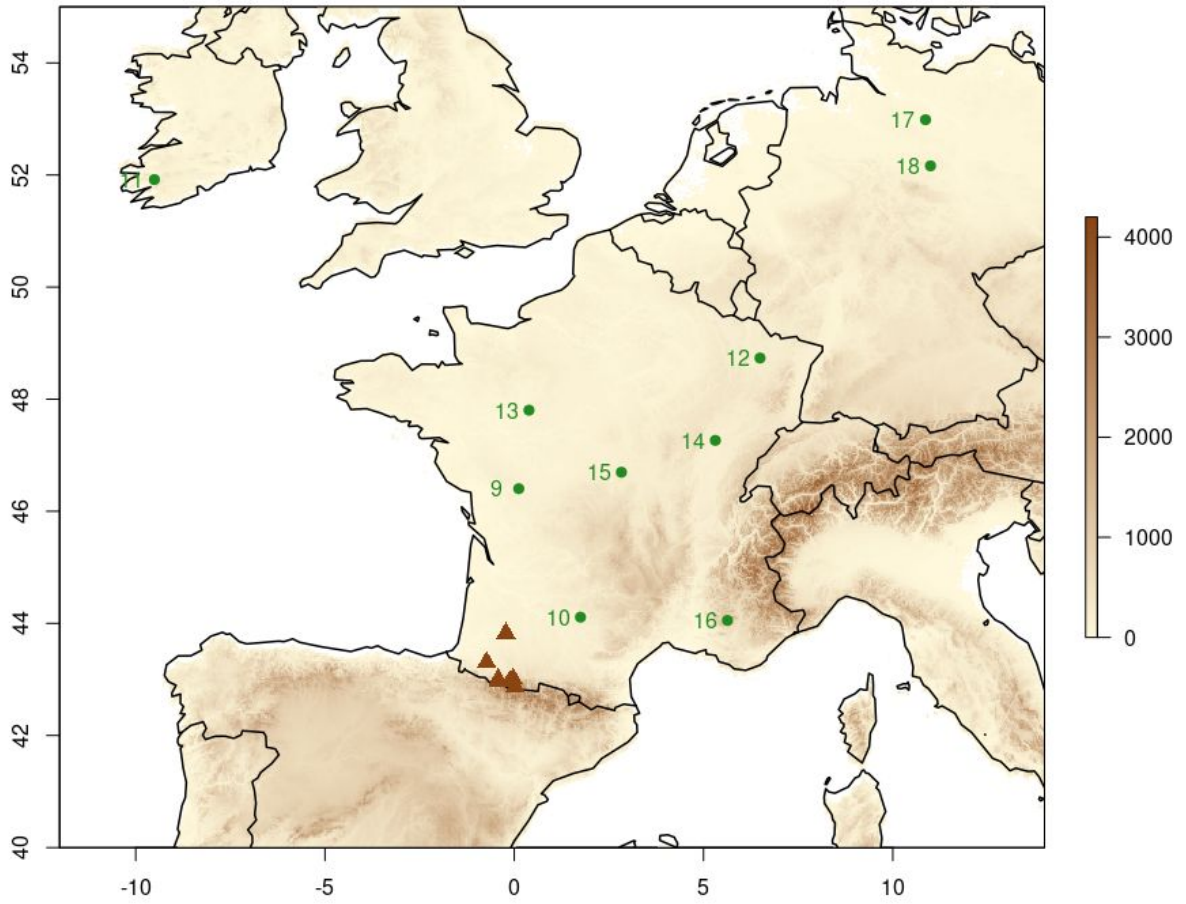
- Climate data (1950-2000, worldclim data):
Mean annual temperature & precipitation sums

- 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	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	Killarney, 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	Göhrde, 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

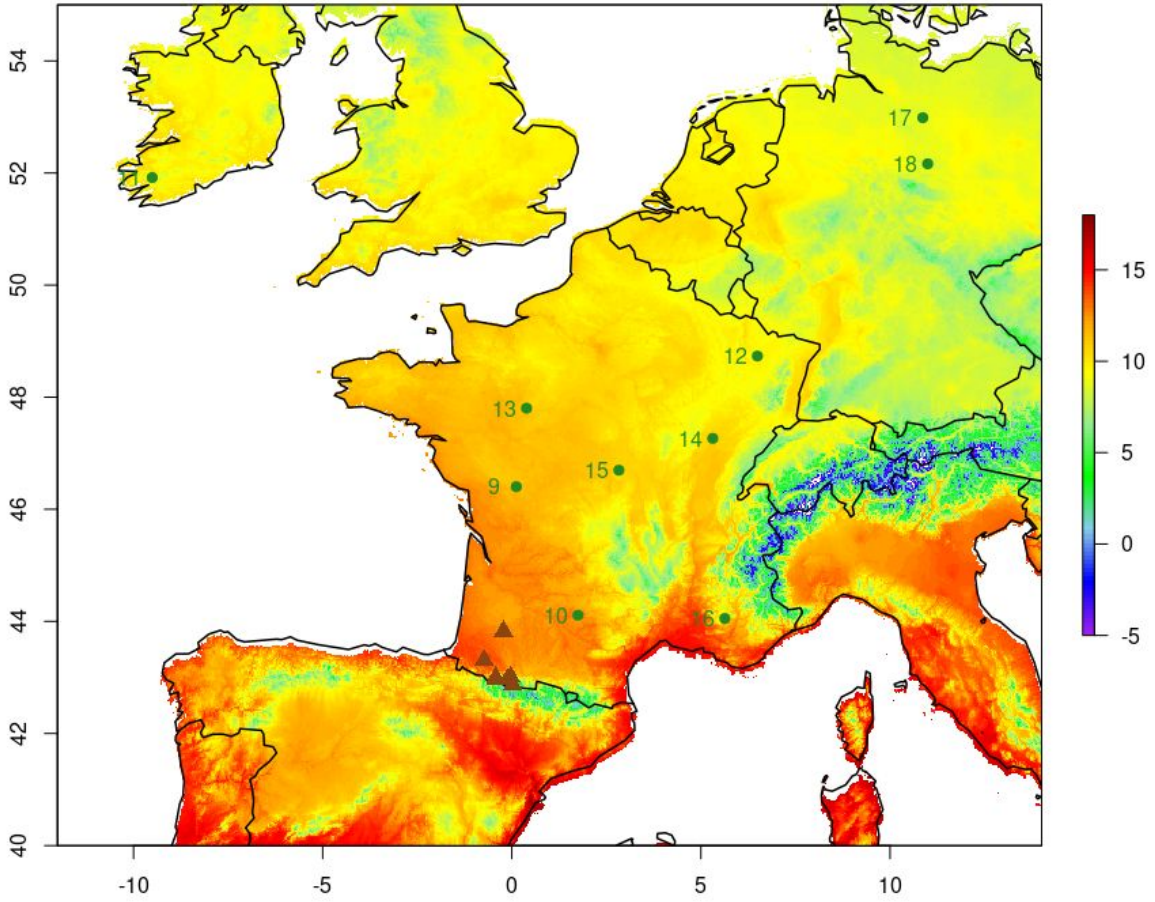
- Climate data (1950-2000, worldclim data):
Mean annual temperature & precipitation sums

- 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	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	Killarney, 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	Göhrde, 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

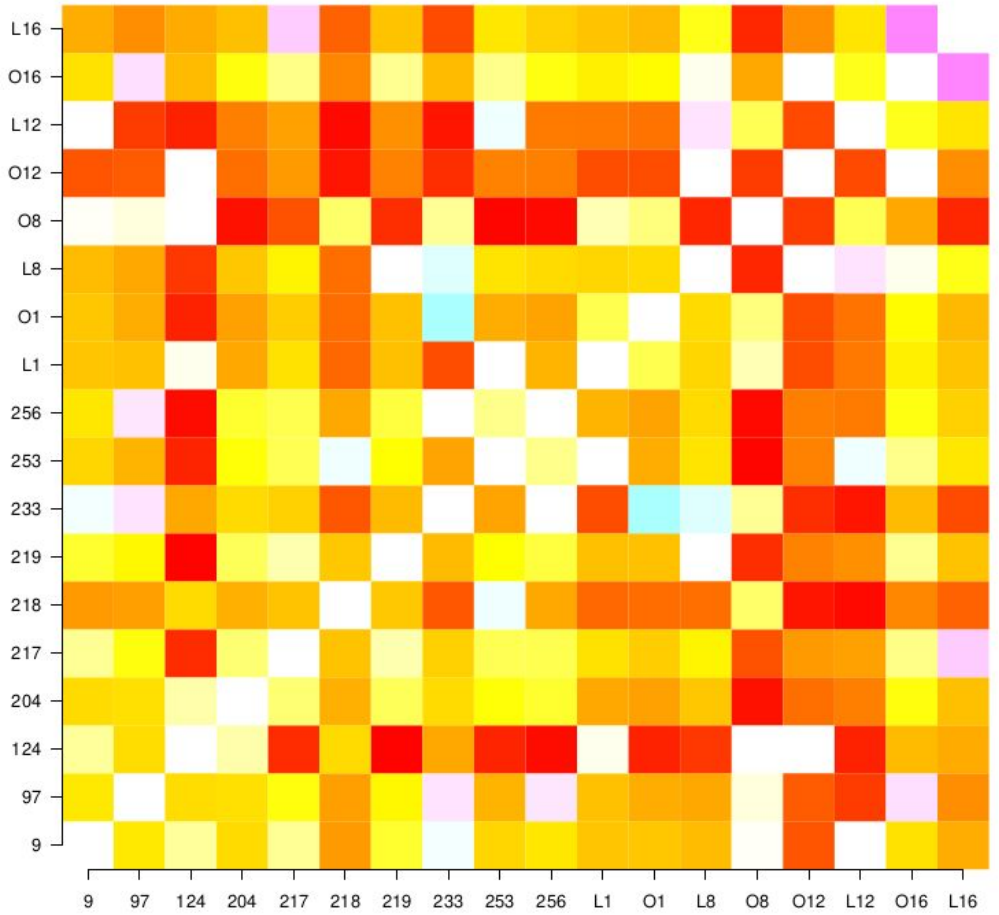
```
pairwise_fst_results <- compute.pairwiseFST(pooldata, method = "Anova",output.snp.values = FALSE)
Fstmatrix <- pairwise_fst_results@PairwiseFSTmatrix

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, …)

poolfstat: Computing f-Statistics and Building Admixture Graphs Based on Allele Count or Pool-Seq Read Count Data

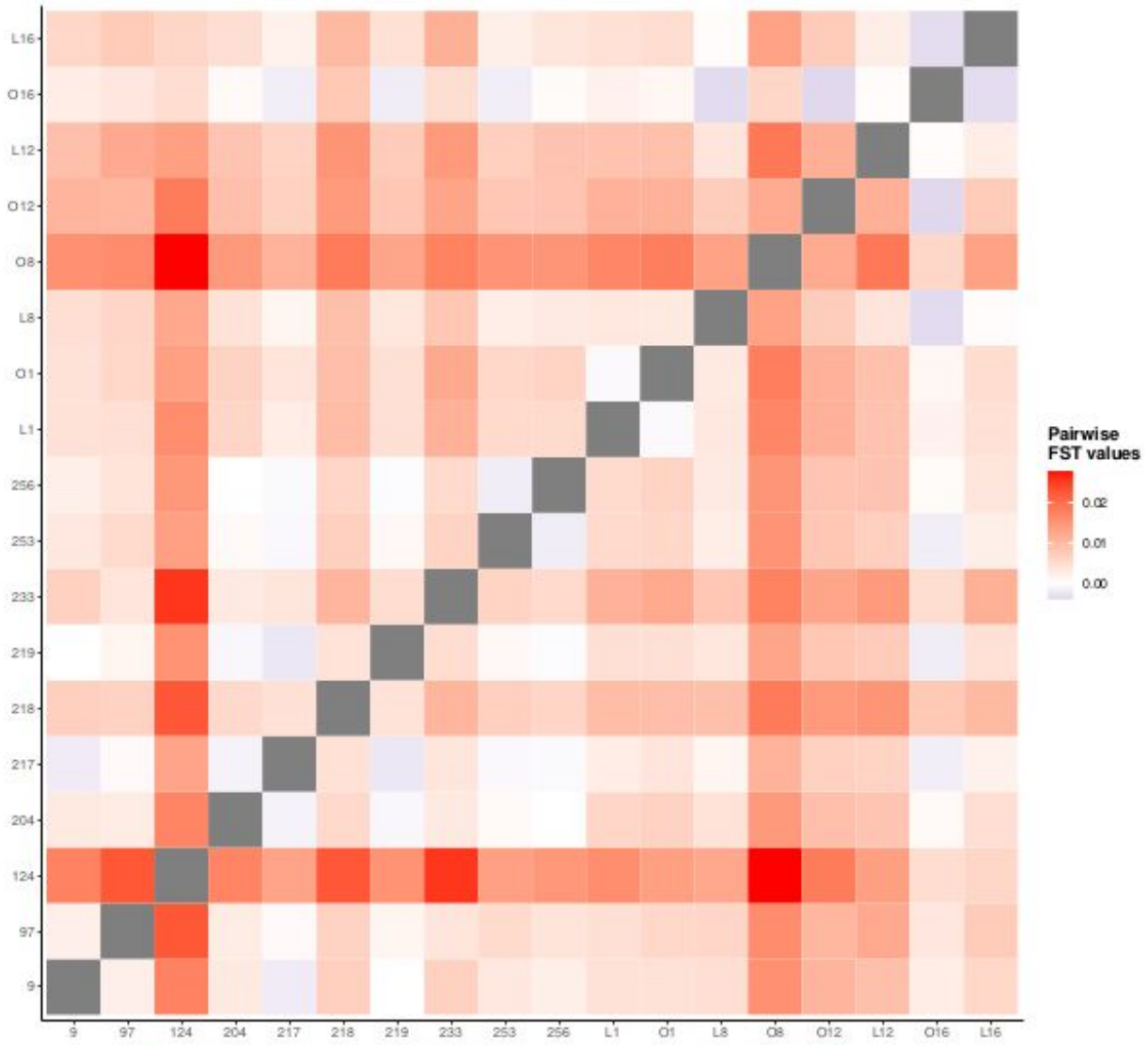
Functions for the computation of F-, f- and D-statistics (e.g., Fst, hierarchical F-statistics, Patterson's F2, F3, F3*, F4 and D parameters) in population genomics studies from allele count or Pool-Seq read count data and for the fitting, building and visualization of admixture graphs. The package also includes several utilities to manipulate Pool-Seq data stored in standard format (e.g., such as 'vcf' files or 'rsync' files generated by the the 'PoPoolation' software) and perform conversion to alternative format (as used in the 'BayPass' and 'SelEstim' software). As of version 2.0, the package also includes utilities to manipulate standard allele count data (e.g., stored in TreeMix, BayPass and SelEstim format).

Version:	3.0.0
Depends:	R (≥ 3.0)
Imports:	Rcpp (≥ 1.0.5), methods, data.table, utils, foreach, doParallel, parallel, DiagrammeR, ape, stats, Ryacas, Matrix, RcppProgress, progress, nnls
LinkingTo:	Rcpp, RcppProgress
Published:	2024-11-23
DOI:	10.32614/CRAN.package.poolfstat
Author:	Mathieu Gautier [aut, cre]

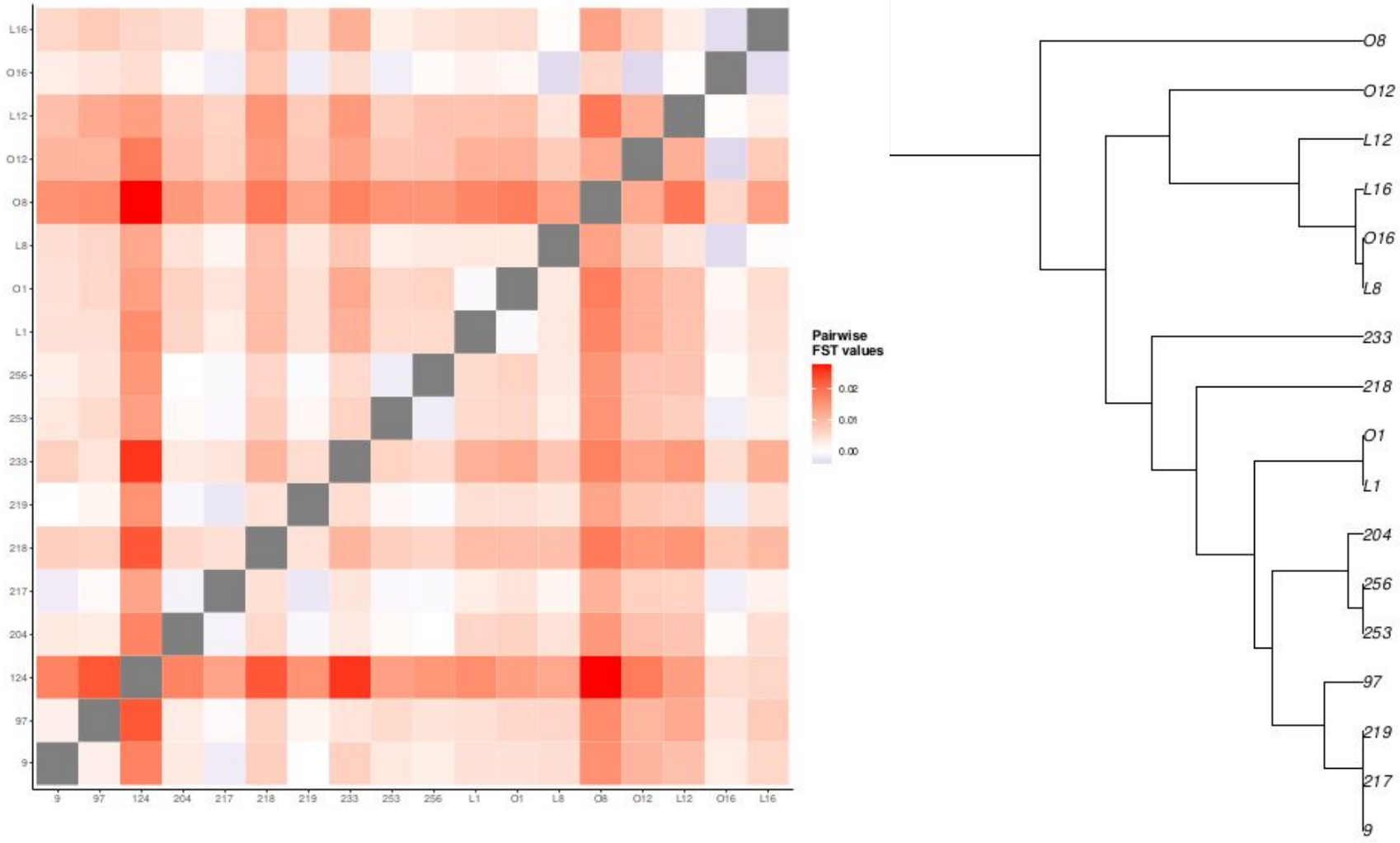


Population structure -> Day 2

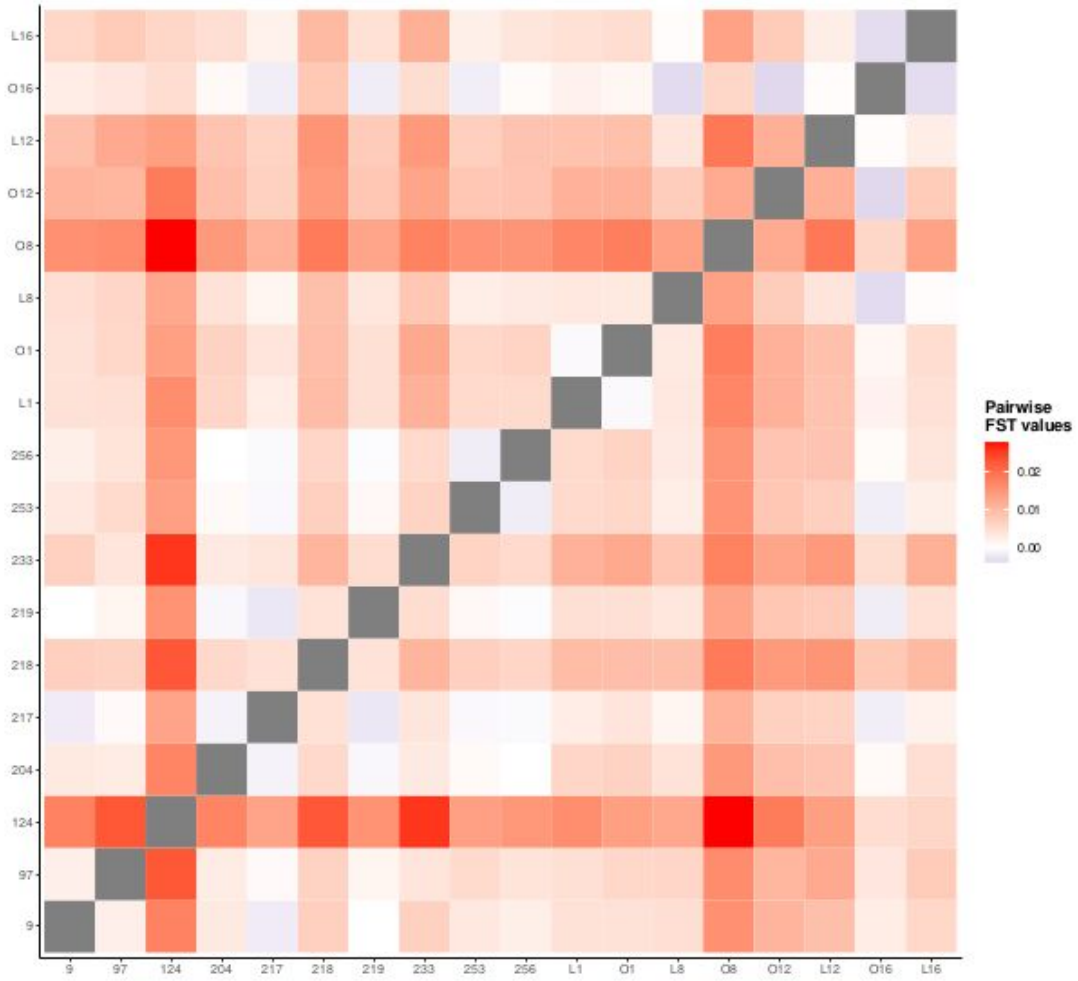
```
ggplot(data = melted_Fstmatrix, aes(x=X1, y=X2, fill=value))+
geom_tile()+
scale_fill_gradient2("Pairwise\nFST values", low = "navyblue",
mid = "white",high = "red", midpoint = 0)+
xlab("")+ylab("")+
scale_x_discrete(labels=c("Pool1" = "9", "Pool2" = "97",
"Pool3" = "124","Pool4" = "204", "Pool5" = "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",
4"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"))
```



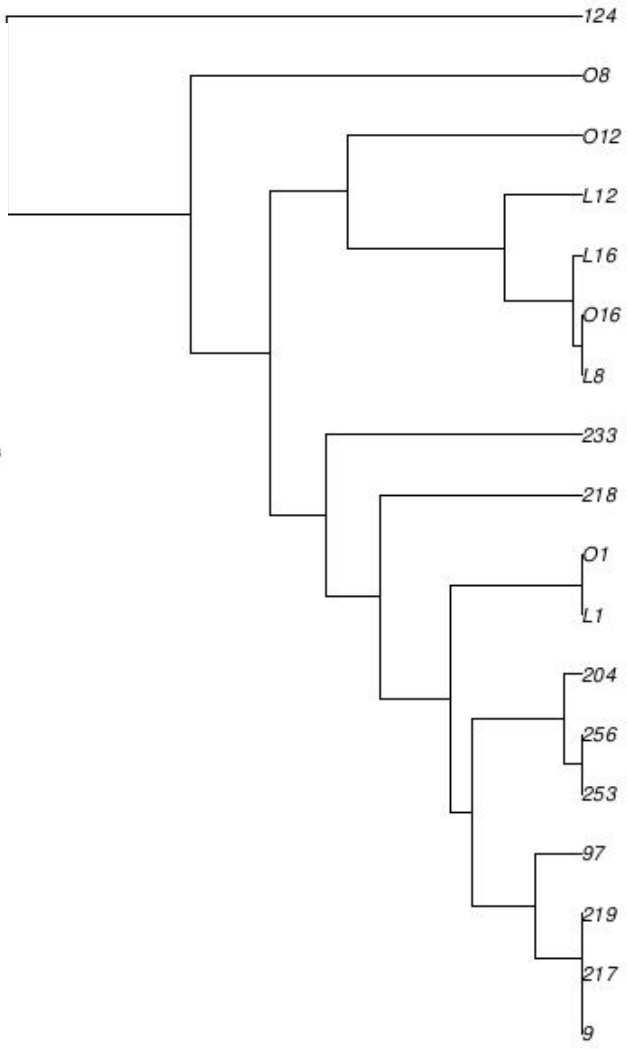
Population structure -> Day 2

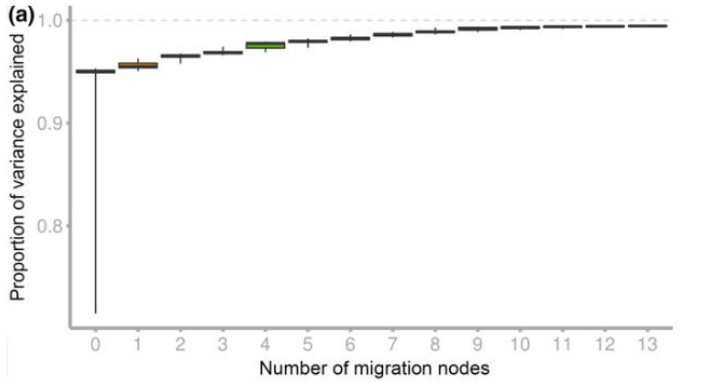


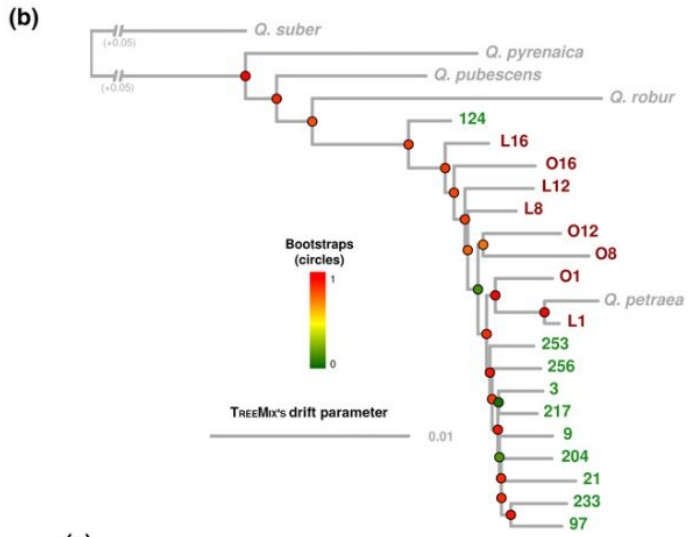
Population structure -> Day 2

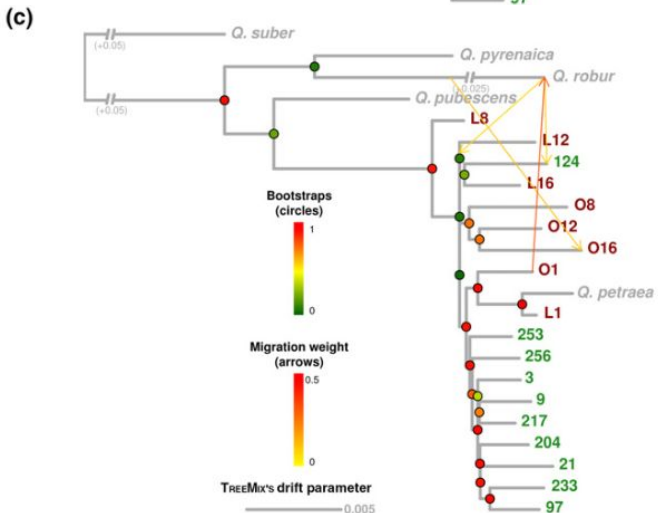


TreeMix



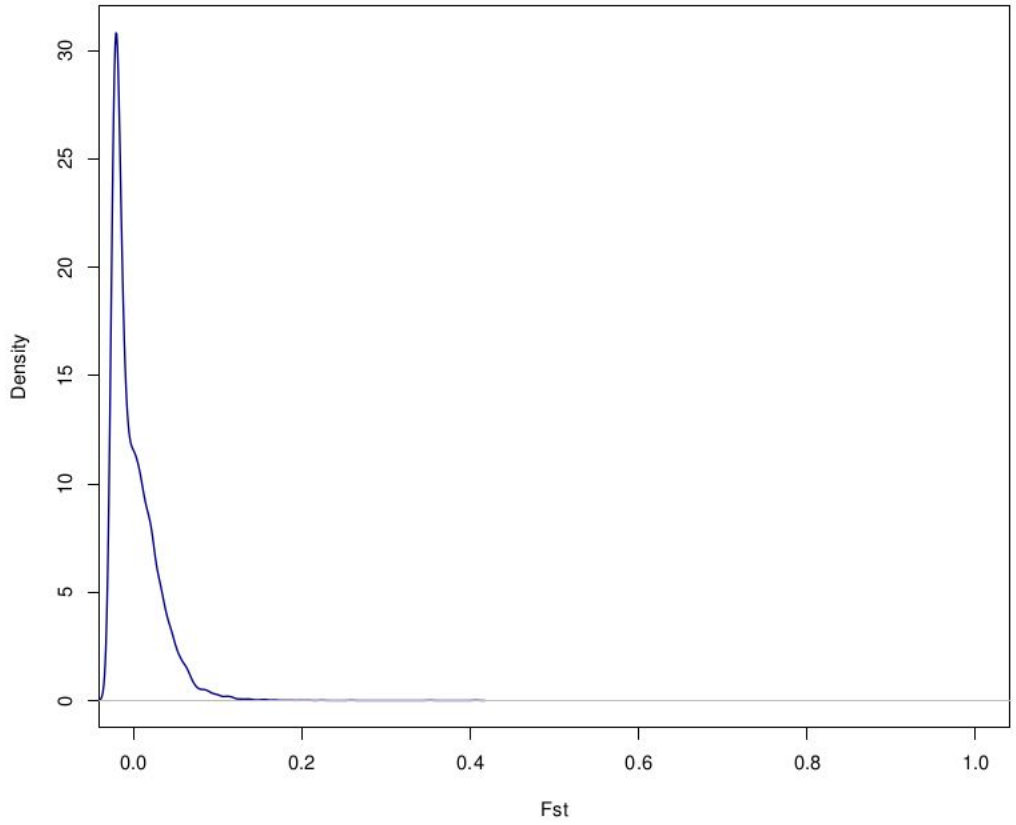






Global variation in F_{ST} -> (Day3 &) Day 4

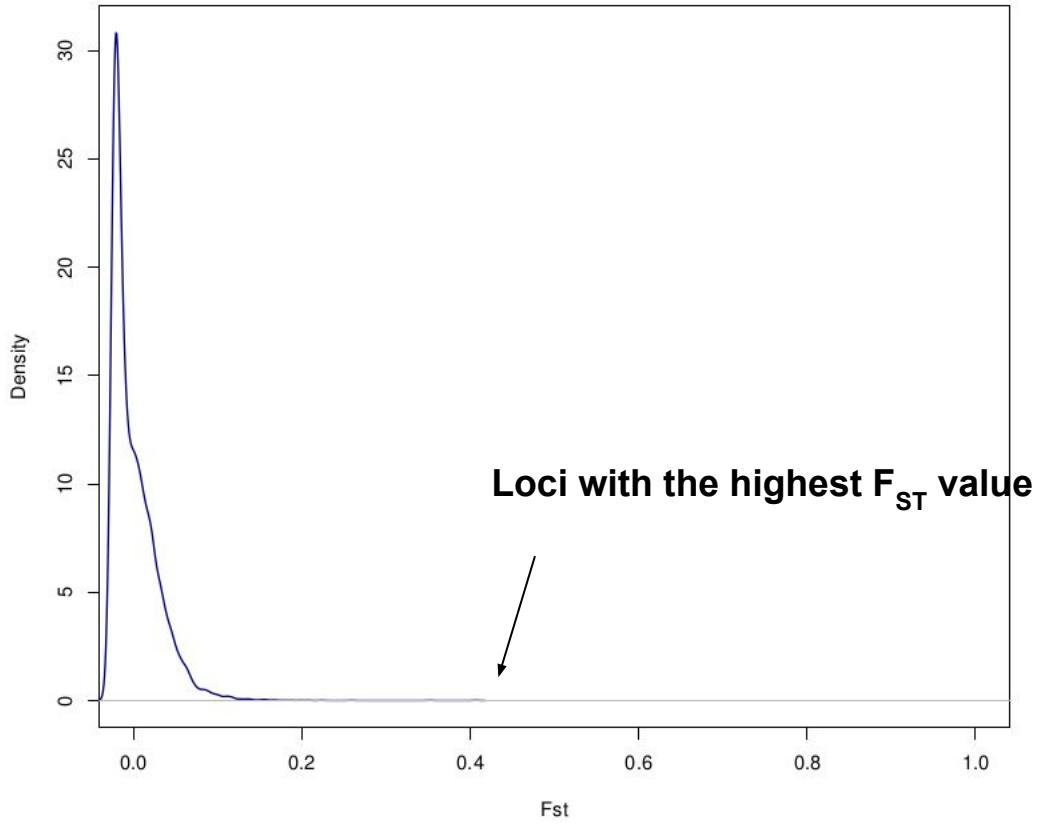
```
SNPvalues=computeFST(pooldata, method="Anova")$snp.FST
SNPvalues=as.data.frame(SNPvalues)

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: **0.406**512841482787"

```
 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()
```

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
```

```
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$SNPvalues==max(SNPvalu
es)),]/pooldata@readcoverage[which(SNPvalues$SNPvalues==max(SNPvalues)),]
)
[1] 0.8461538        (i.e. Pop L1)
>
min(pooldata@refallele.readcount[which(SNPvalues$SNPvalues==max(SNPvalue
s)),]/pooldata@readcoverage[which(SNPvalues$SNPvalues==max(SNPvalues)),])
[1] 0.009433962    (i.e. Pop 219)
```

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?)

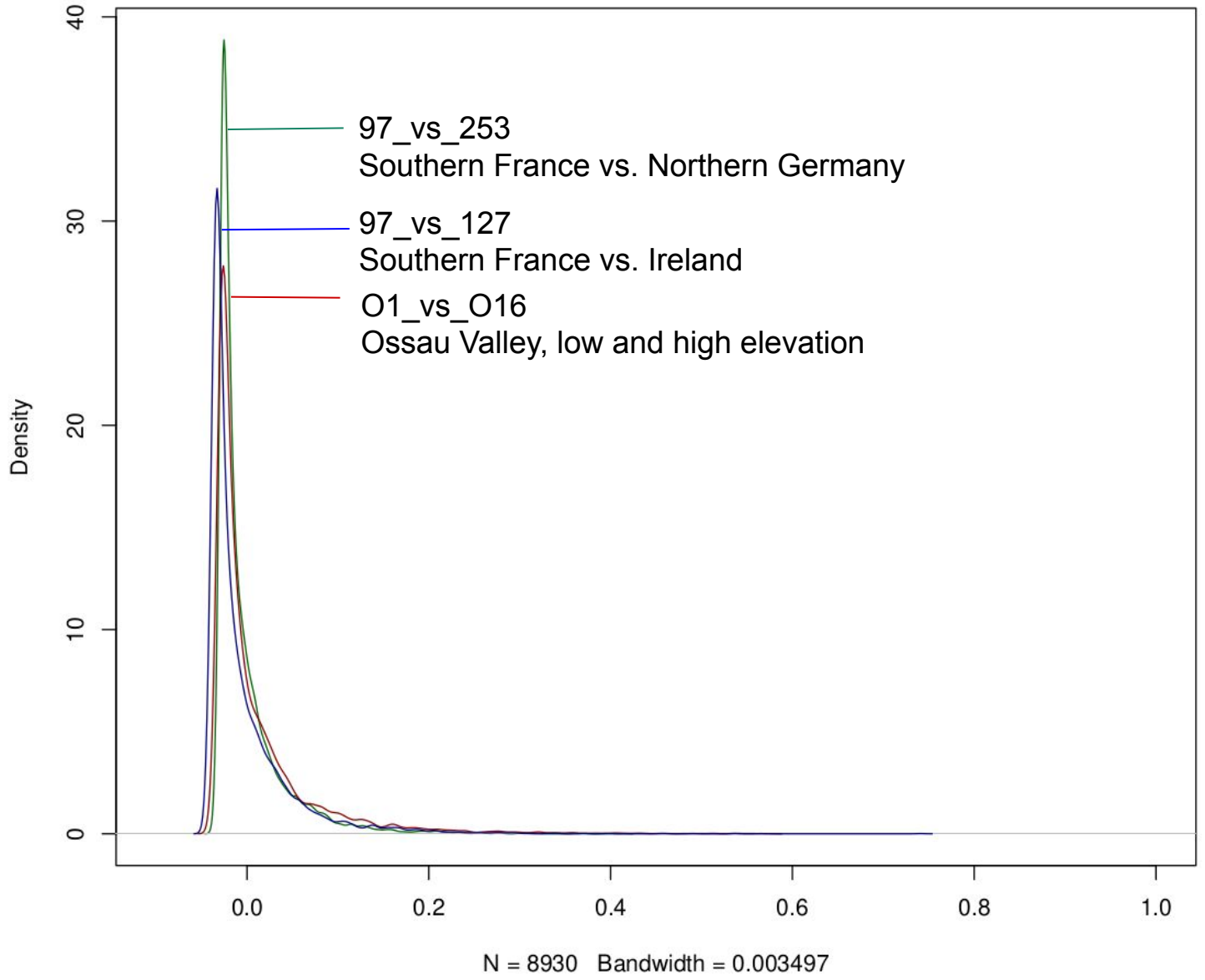
pooldata@snp.info[which(SNPvalues$SNPvalues>quantile(SNPvalues$SNPvalues, probs = 0.99, na.rm = TRUE)),]

	Chromosome	Position	RefAllele	AltAllele
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



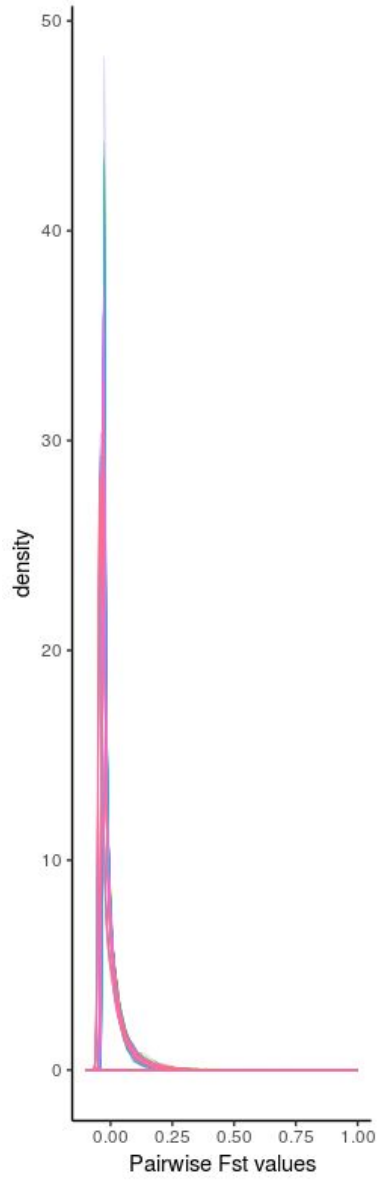
Near-zero F_{ST} at most loci, even between populations at quite different latitude

-> “Recent” post-glacial recolonisation from Southern refugia

\+ large population sizes

Empirical distribution of F_{ST} over pairs of pop

-> Day4



comparison_pops

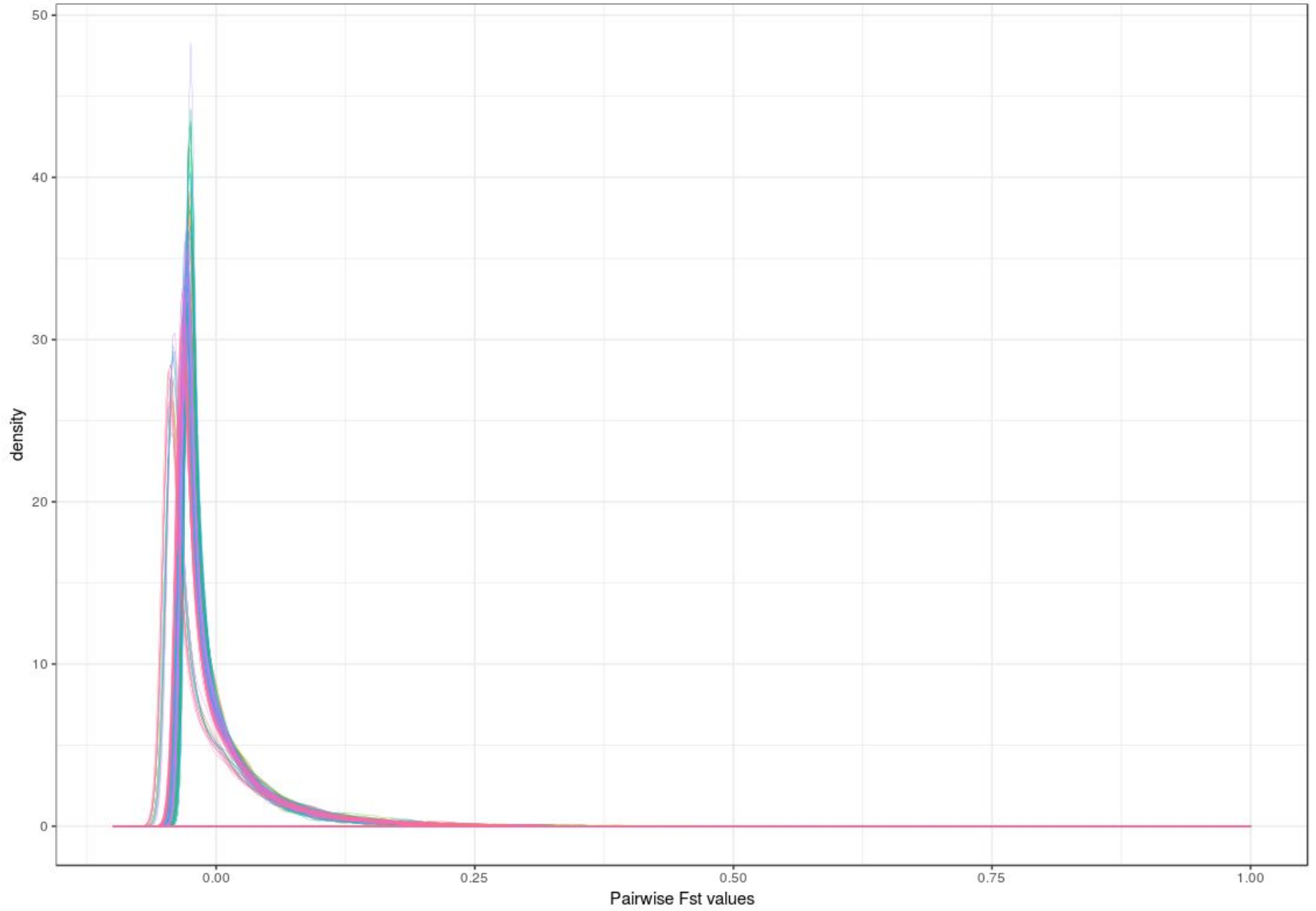
9_vs_97	97_vs_218	124_vs_L1	204_vs_O16	218_vs_O1	233_vs_L1	256_vs_O8	L8_vs_L12
9_vs_124	97_vs_219	124_vs_O1	204_vs_L16	218_vs_L8	233_vs_O1	256_vs_O12	L8_vs_O16
9_vs_204	97_vs_233	124_vs_L8	217_vs_218	218_vs_O8	233_vs_L8	256_vs_L12	L8_vs_L16
9_vs_217	97_vs_253	124_vs_O8	217_vs_219	218_vs_O12	233_vs_O8	256_vs_O16	O8_vs_O12
9_vs_218	97_vs_256	124_vs_O12	217_vs_233	218_vs_L12	233_vs_O12	256_vs_L16	O8_vs_L12
9_vs_219	97_vs_L1	124_vs_L12	217_vs_253	218_vs_O16	233_vs_L12	L1_vs_O1	O8_vs_O16
9_vs_233	97_vs_O1	124_vs_O16	217_vs_256	218_vs_L16	233_vs_O16	L1_vs_L8	O8_vs_L16
9_vs_253	97_vs_L8	124_vs_L16	217_vs_L1	219_vs_233	233_vs_L16	L1_vs_O8	O12_vs_L12
9_vs_256	97_vs_O8	204_vs_217	217_vs_O1	219_vs_253	253_vs_256	L1_vs_O12	O12_vs_O16
9_vs_L1	97_vs_O12	204_vs_218	217_vs_L8	219_vs_256	253_vs_L1	L1_vs_L12	O12_vs_L16
9_vs_O1	97_vs_L12	204_vs_219	217_vs_O8	219_vs_L1	253_vs_O1	L1_vs_O16	L12_vs_O16
9_vs_L8	97_vs_O16	204_vs_233	217_vs_O12	219_vs_O1	253_vs_L8	L1_vs_L16	L12_vs_L16
9_vs_O8	97_vs_L16	204_vs_253	217_vs_L12	219_vs_L8	253_vs_O8	O1_vs_L8	O16_vs_L16
9_vs_O12	124_vs_204	204_vs_256	217_vs_O16	219_vs_O8	253_vs_O12	O1_vs_O8	
9_vs_L12	124_vs_217	204_vs_L1	217_vs_L16	219_vs_O12	253_vs_L12	O1_vs_O12	
9_vs_O16	124_vs_218	204_vs_O1	218_vs_219	219_vs_L12	253_vs_O16	O1_vs_L12	
9_vs_L16	124_vs_219	204_vs_L8	218_vs_233	219_vs_O16	253_vs_L16	O1_vs_O16	
97_vs_124	124_vs_233	204_vs_O8	218_vs_253	219_vs_L16	256_vs_L1	O1_vs_L16	
97_vs_204	124_vs_253	204_vs_O12	218_vs_256	233_vs_253	256_vs_O1	L8_vs_O8	
97_vs_217	124_vs_256	204_vs_L12	218_vs_L1	233_vs_256	256_vs_L8	L8_vs_O12	

length(unique(melted_Fstlocus$comparison_pops))

153 pairwise comparisons -> (n * n-1) /2 *i.e.* (18*17) /2

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?

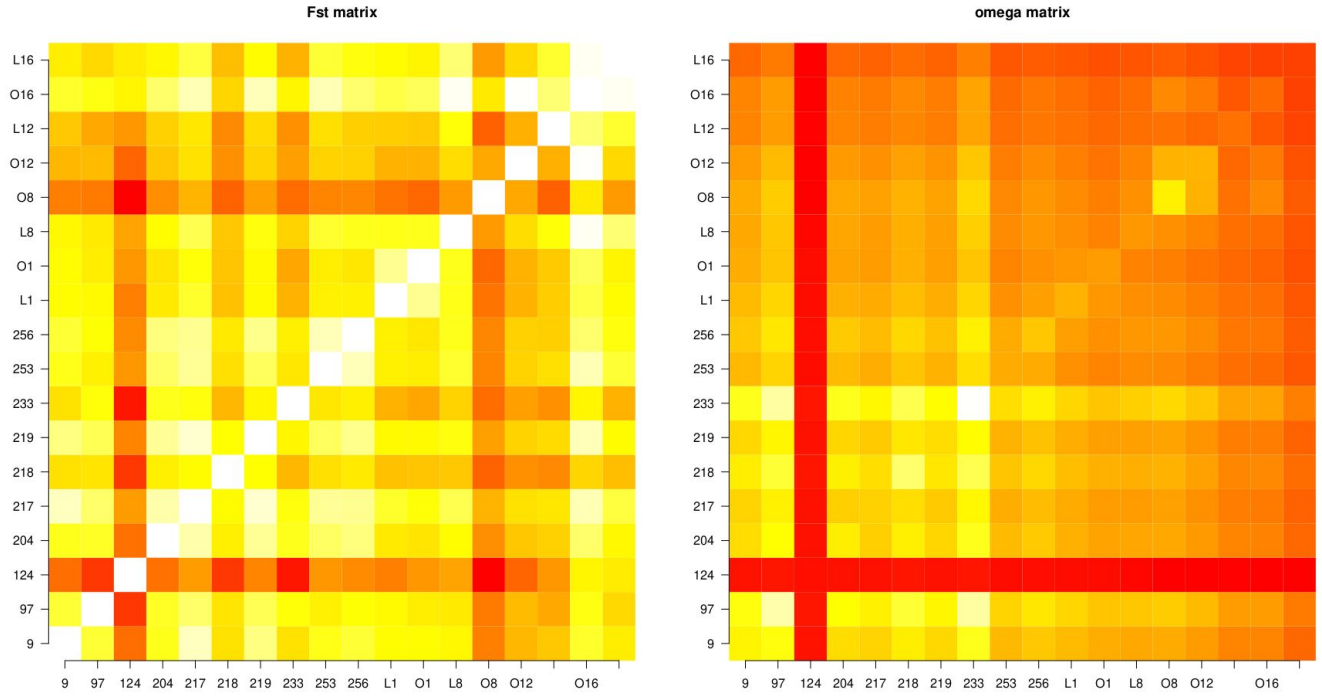
Because we similarly select the top 1% of SNPs with the highest F_{ST} values, corresponding to a given number of SNPs. By doing so, we implicitly assume that 1% of the genome is indeed under positive selection! And we have no idea of this proportion in reality! Could be more, could be less, could be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

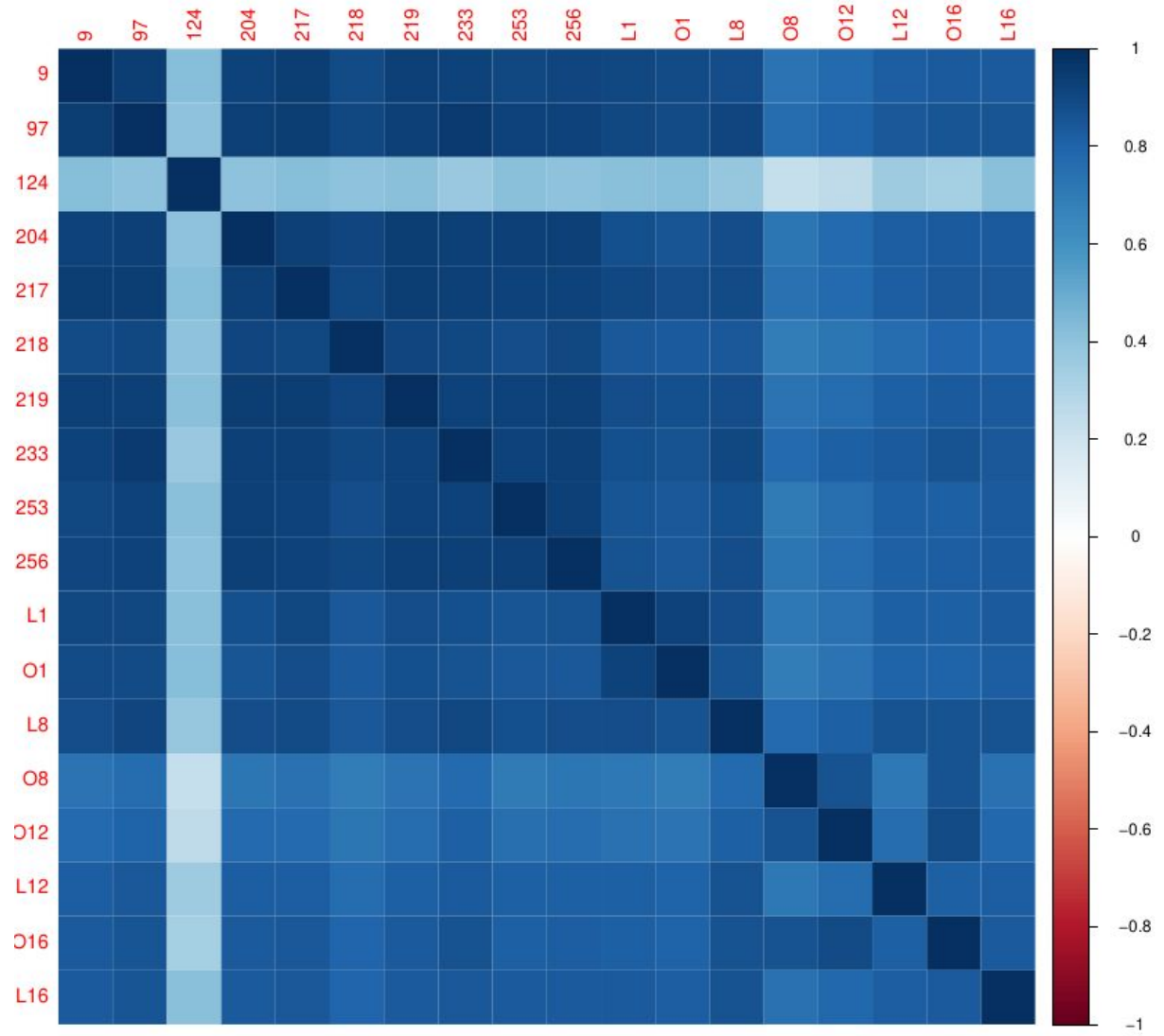
-> Day2

```
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)
```

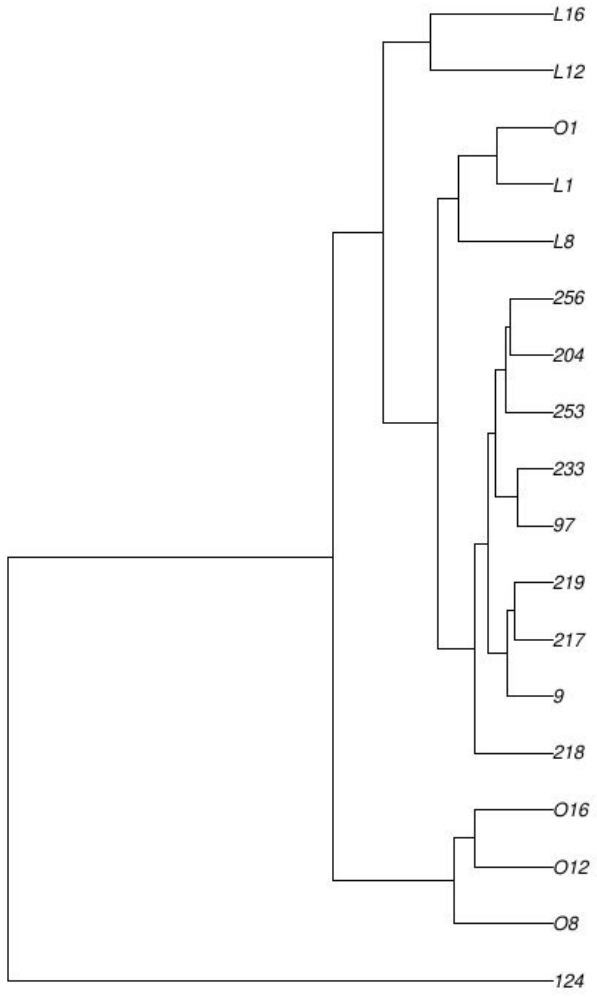


BayPass inference - omega matrix

Correlation matrix

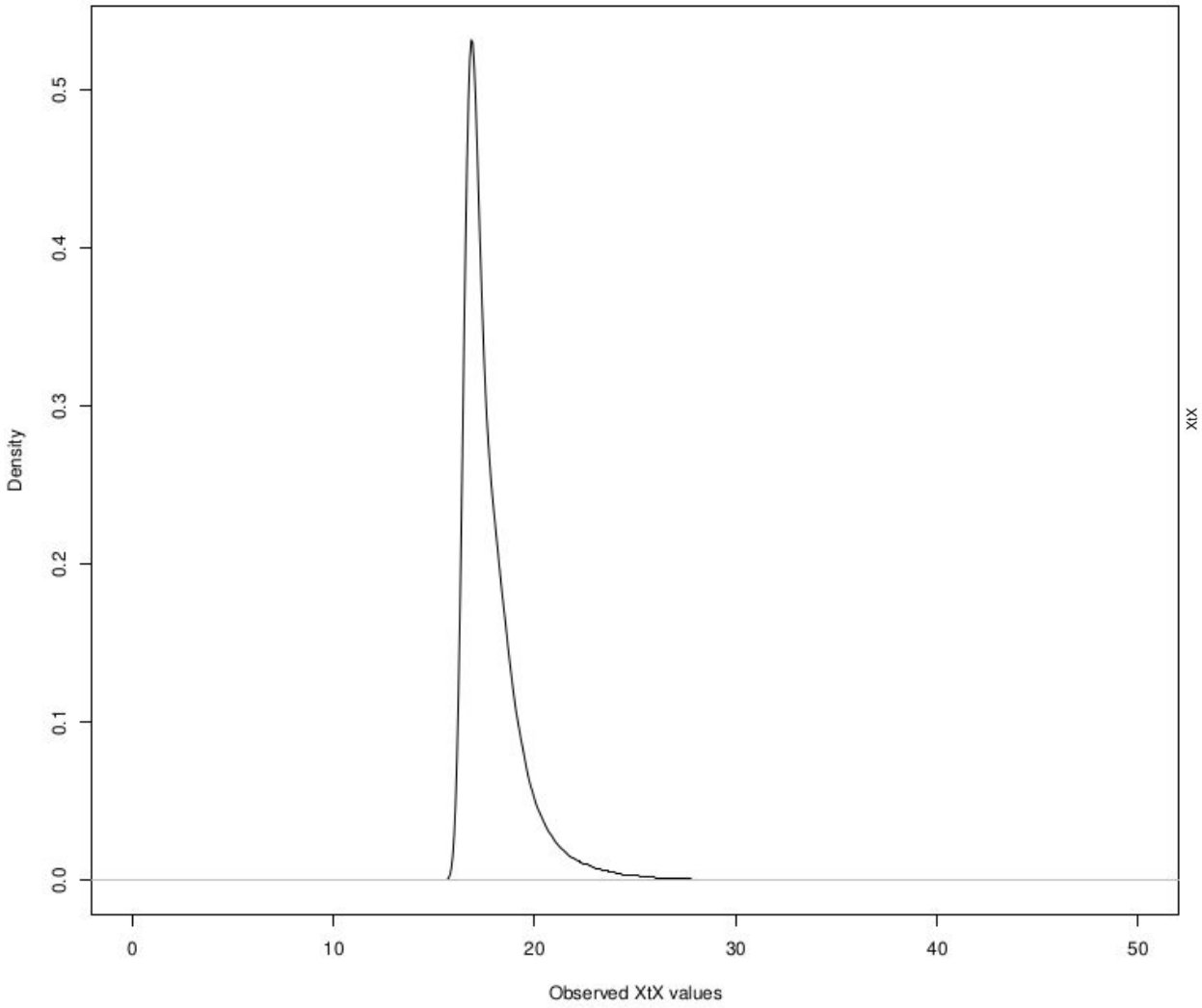


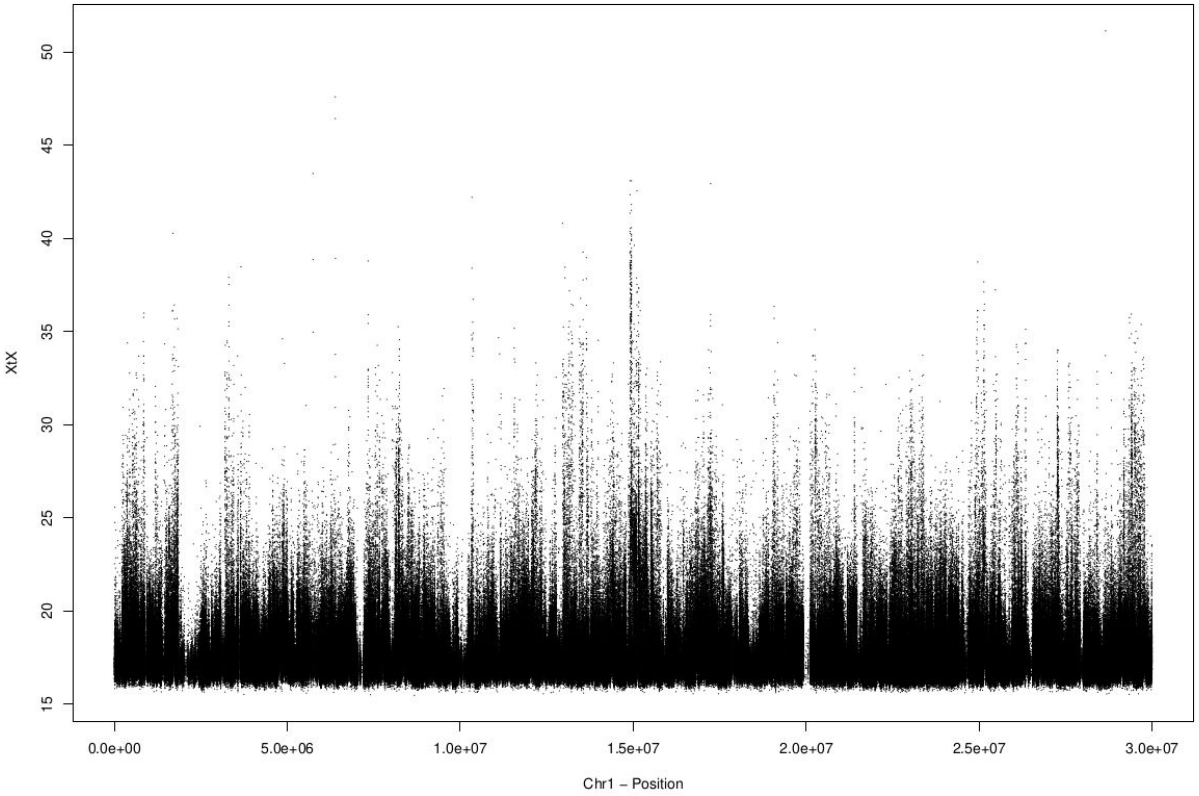
-> Day2



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

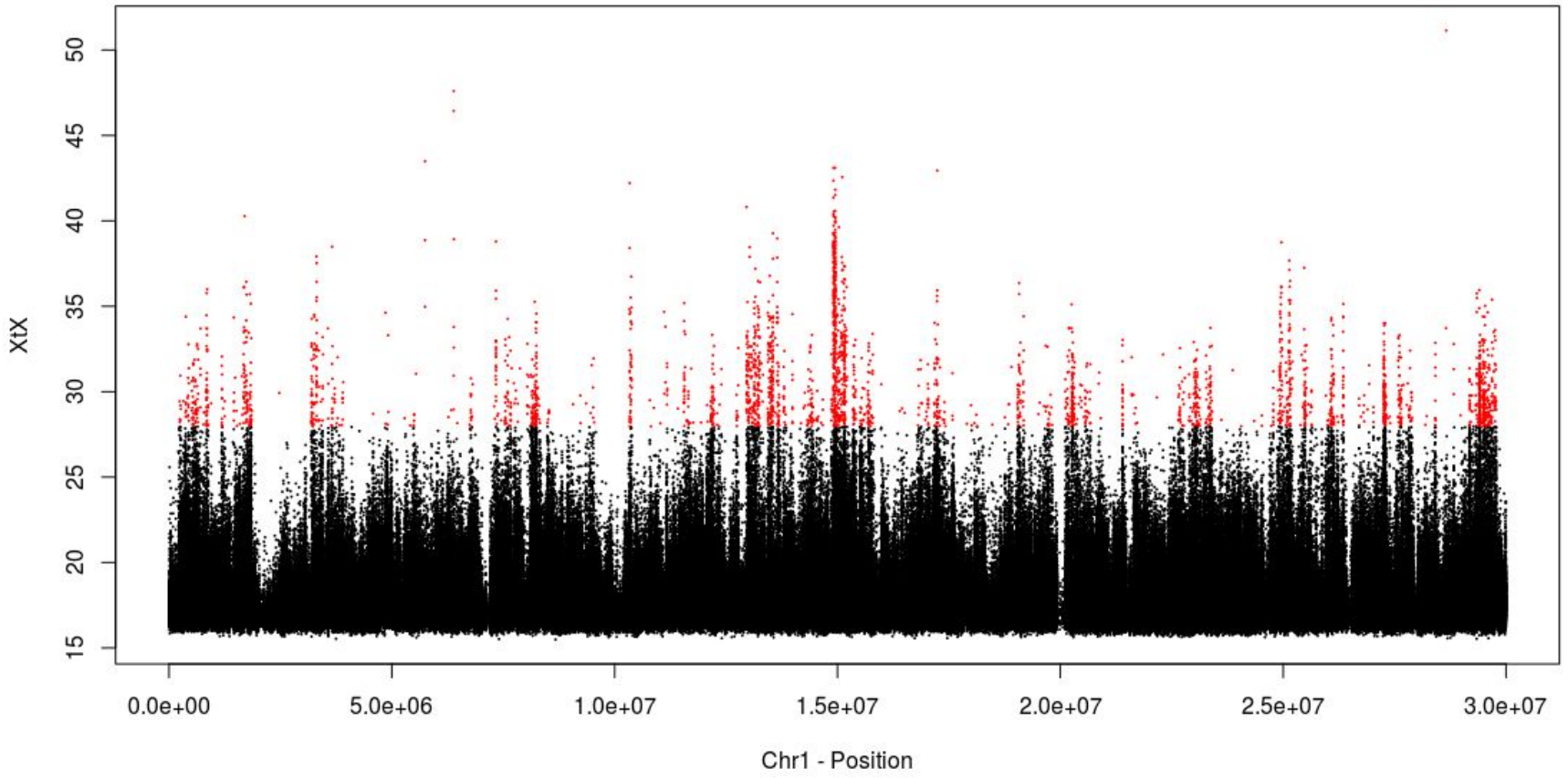
3090/1349416
=> 0.002289879 (**0.23% of SNPs exhibit higher levels of XtX than observed under the neutral simulations**)

⟶ Quite ≠ from the previously assumed 1% !

BayPass - XtX scan (outliers, after calibration)

-> Day4

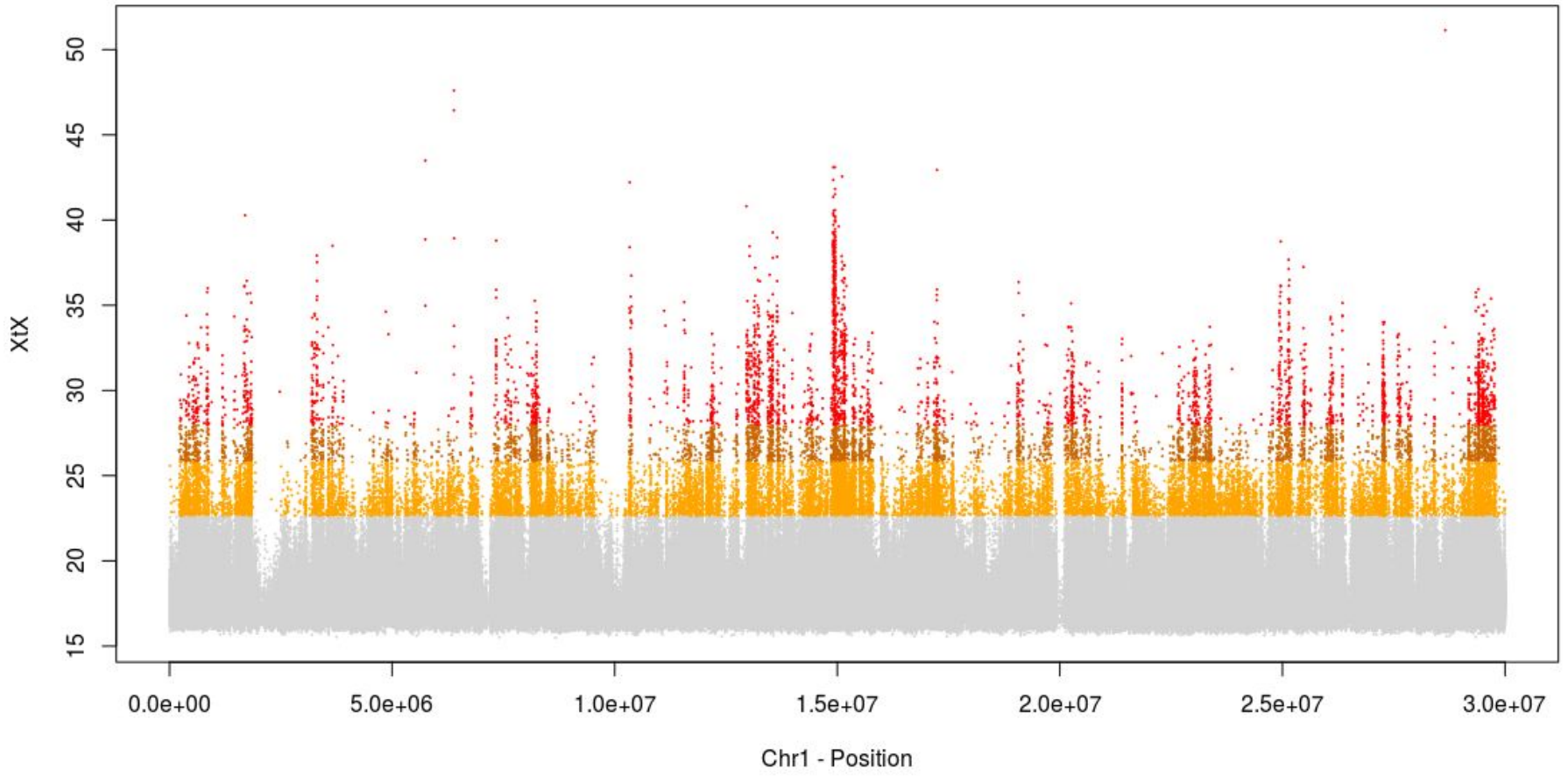
```
plot(XtX_BF$XtX~XtX_BF$position,cex=0.05,pch=20,xlab="Chr1 -
Position",ylab="XtX",col=ifelse(XtX_BF$XtX>=xtx.threshold[3],"red","black"))
```



BayPass - XtX scan (outliers, after calibration)

-> 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"))))
```



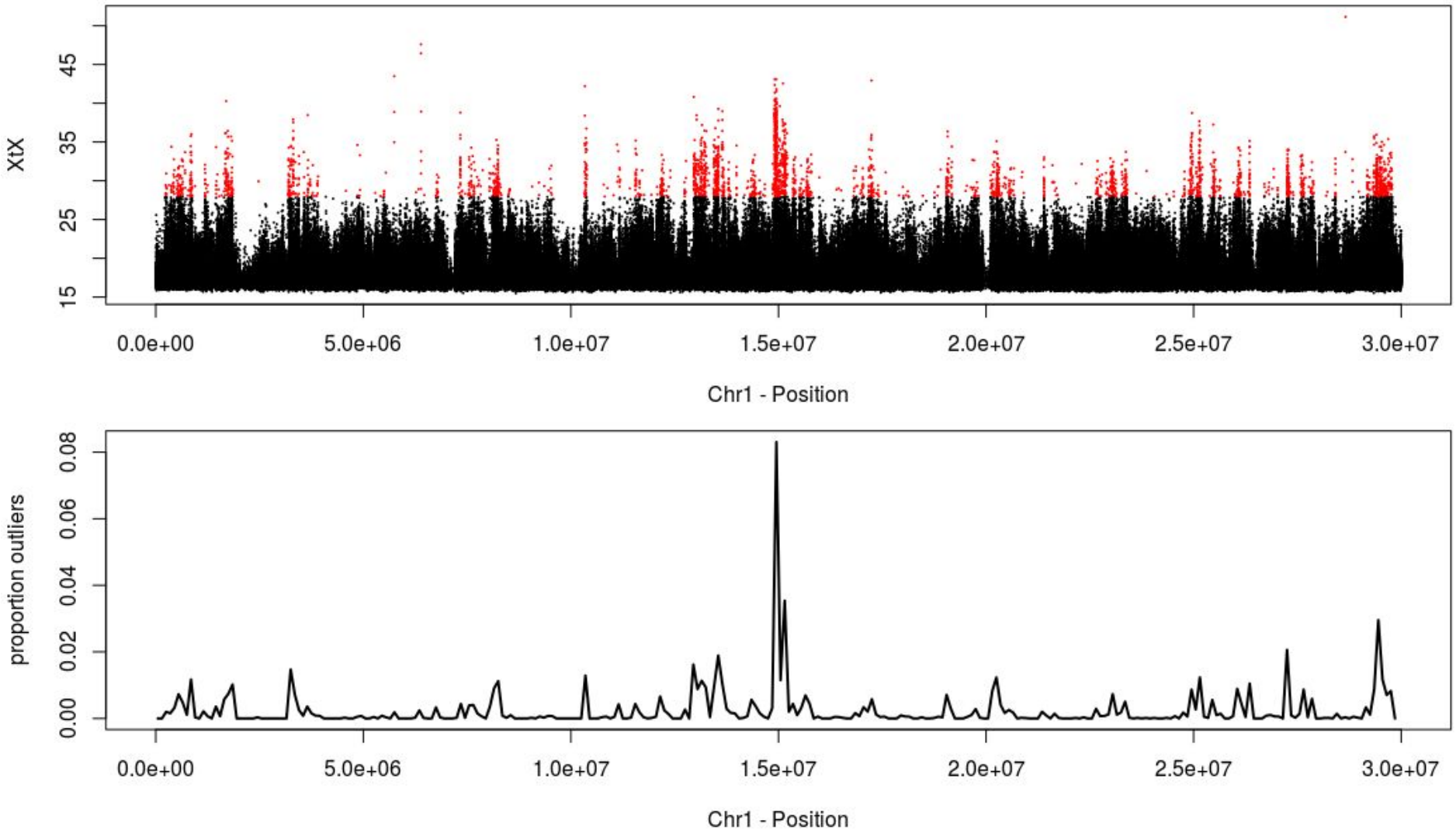
BayPass - XtX - Sliding window approach

Option 2 (very end of the tutorial)

```
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_outliert
hresholdmax))
}
colnames(TablePropOutlier)<-
c("position_median","start_windows","end_windows","outliers_thresholdmax","nb_SNPs_windows","proportion_outliers")
TablePropOutlier<- as.data.frame(TablePropOutlier)
```

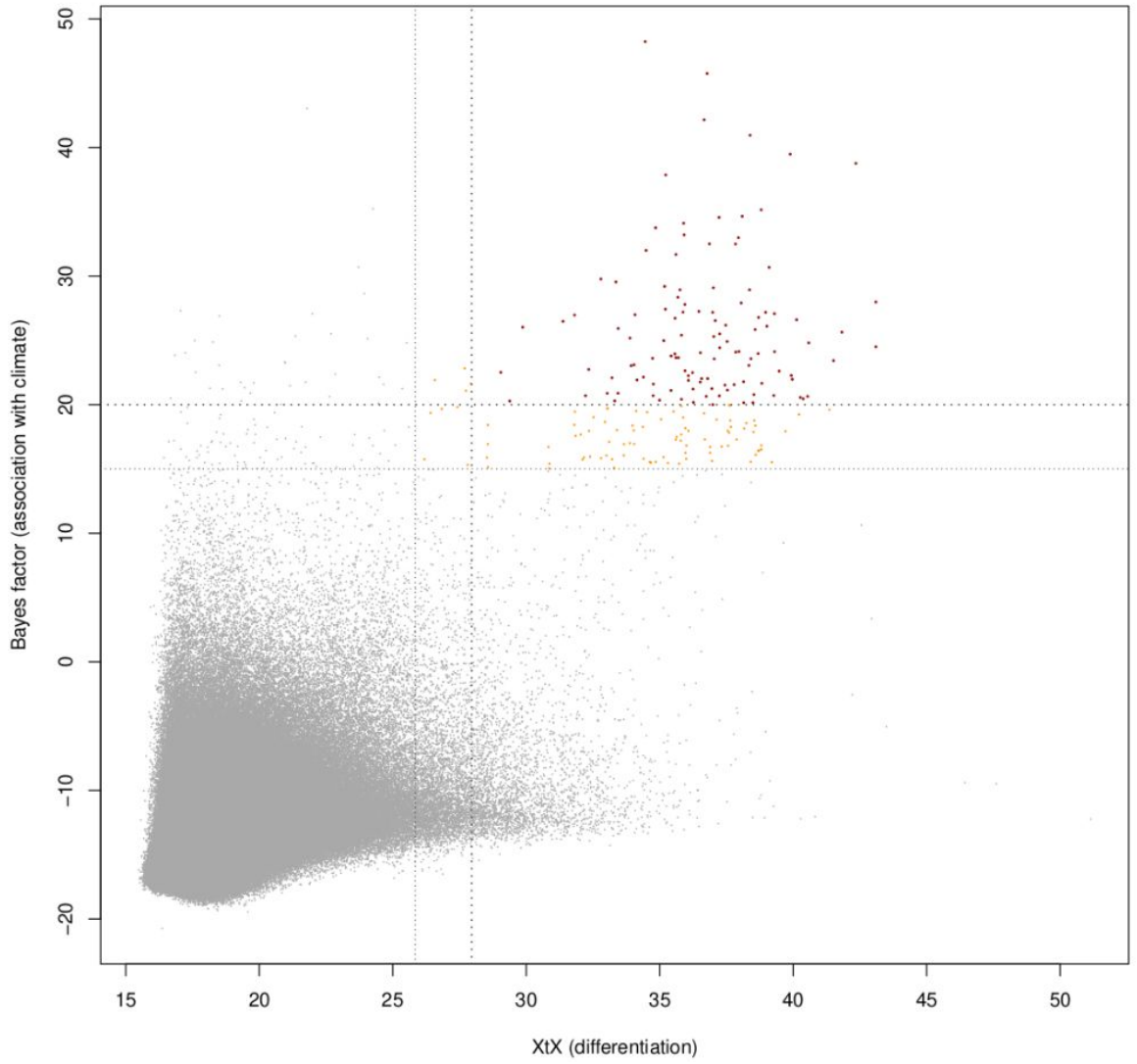
BayPass - XtX - Sliding window approach

Option 2 (very end of the tutorial)



BayPass - combining XtX and GEA !

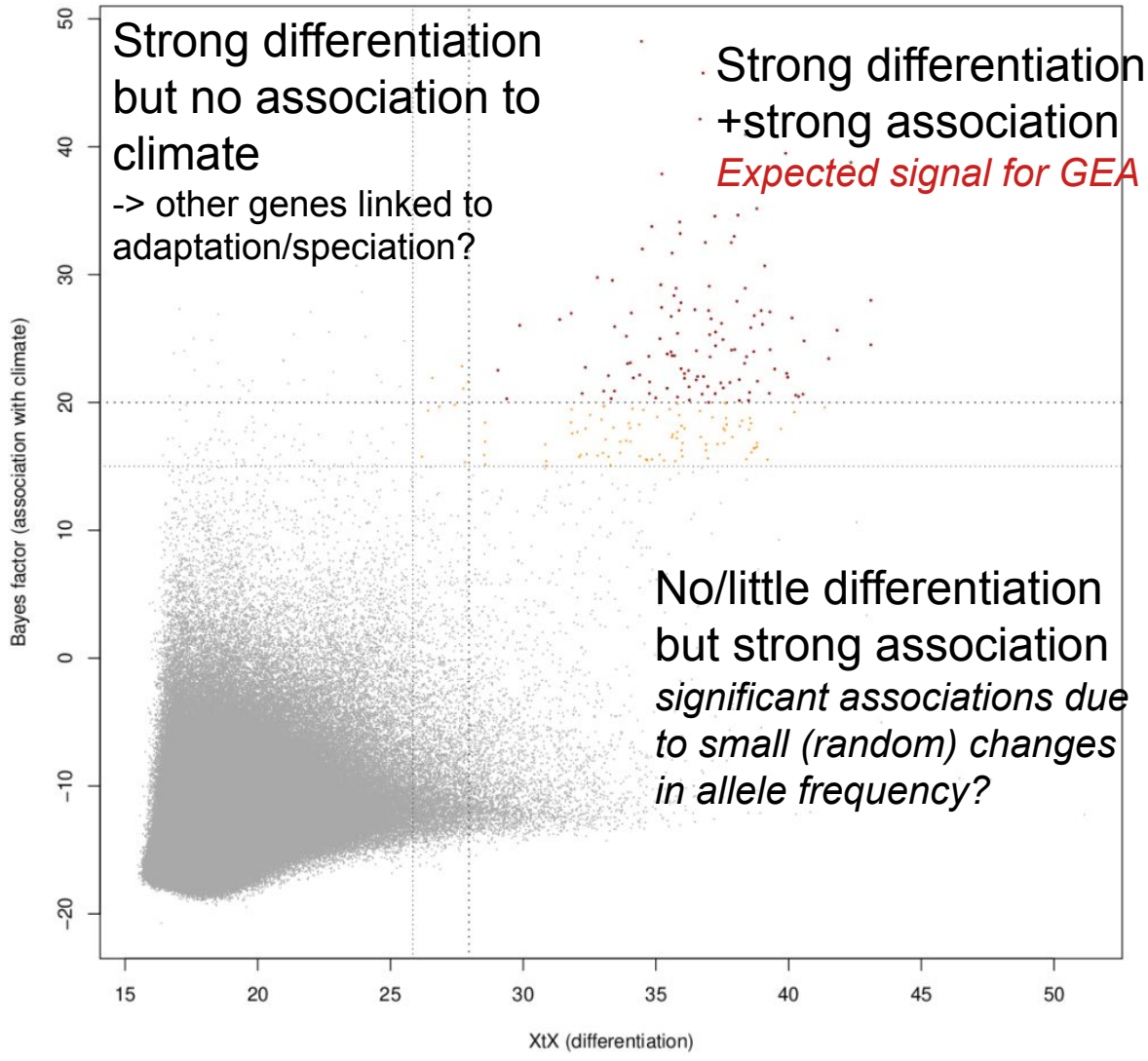
BF



XtX

-> Day5

BayPass - combining XtX and GEA !



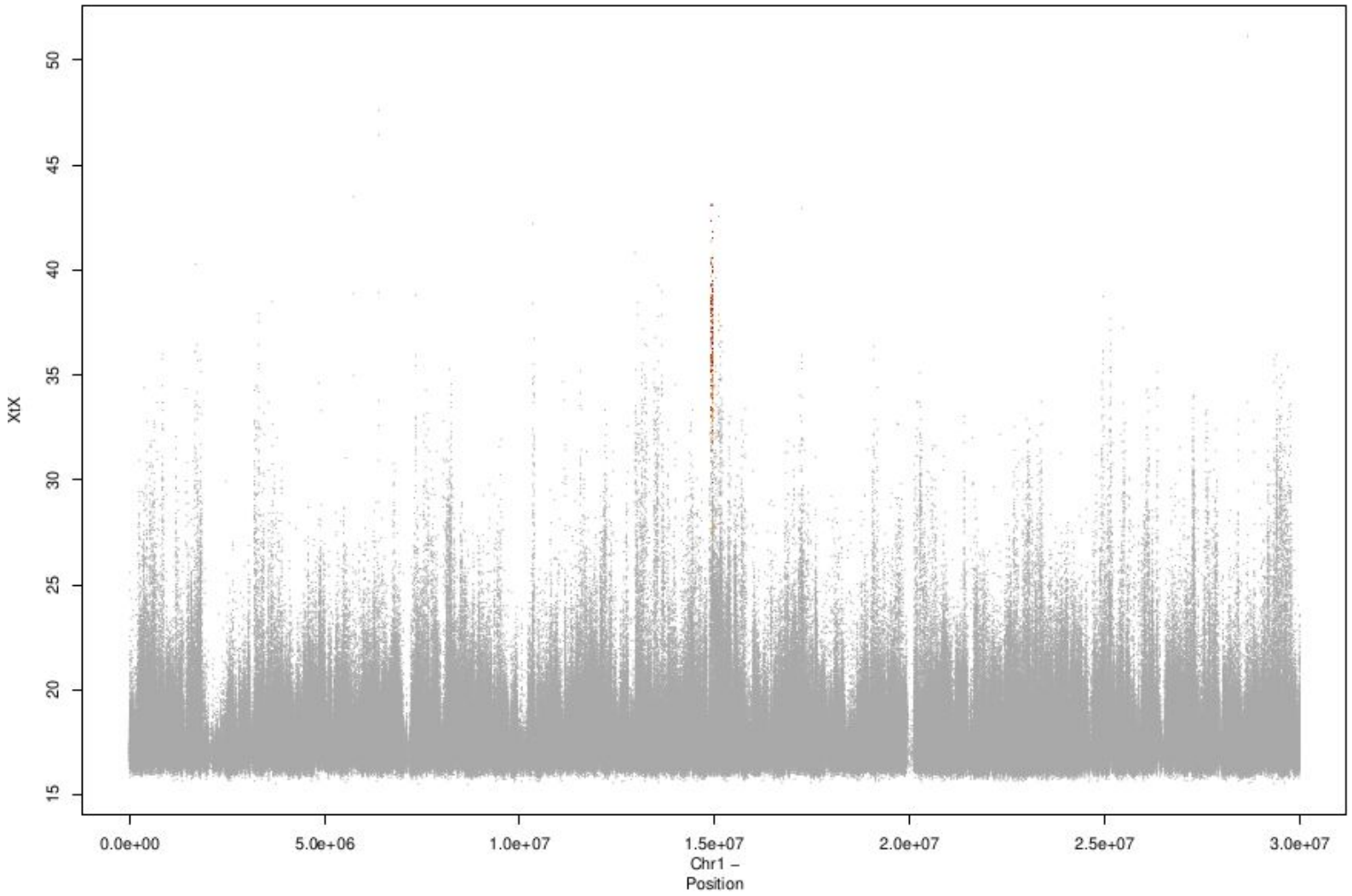
BF

XtX

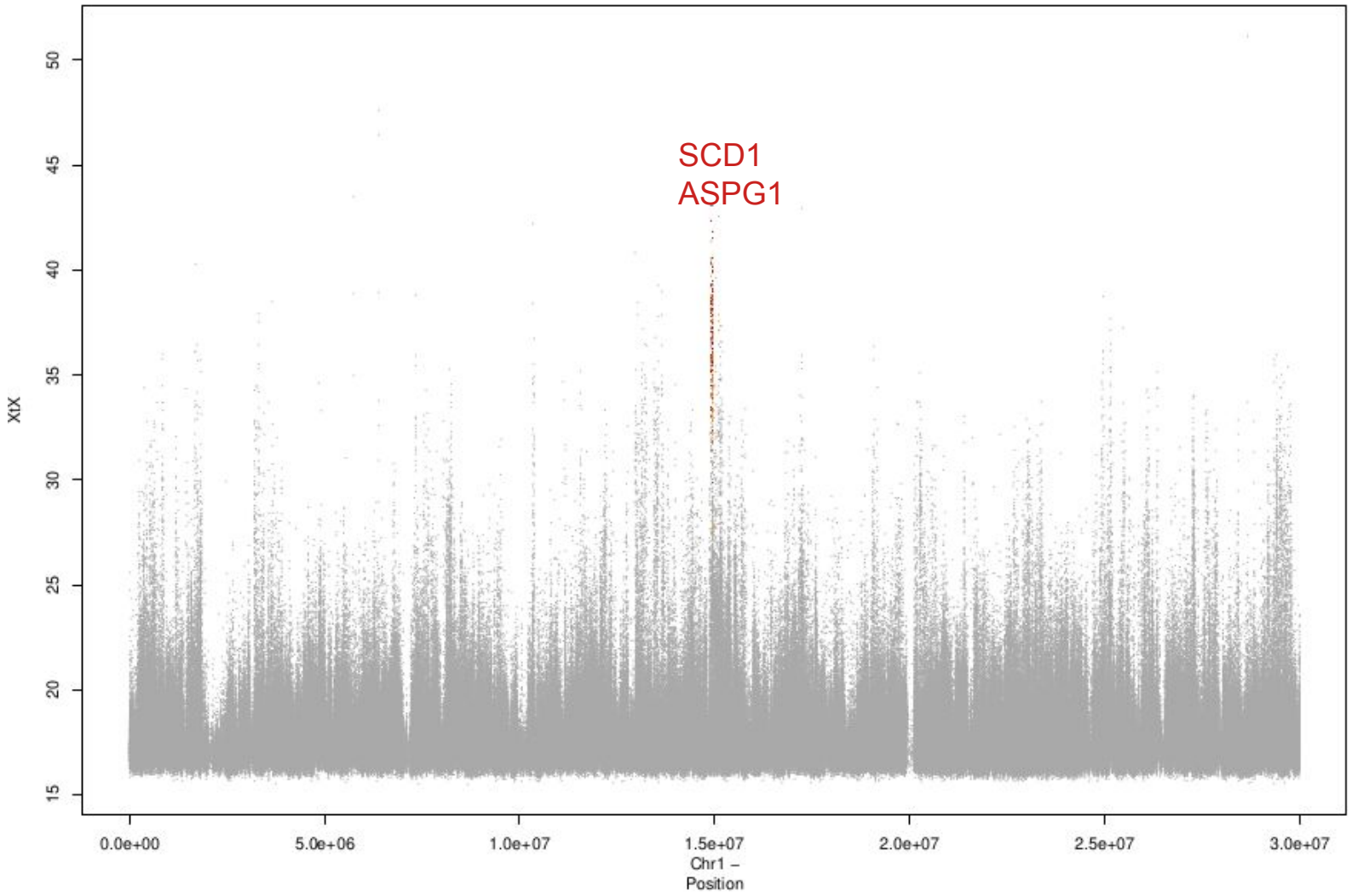
-> Day5

BayPass - XtX + GEA to reveal local adaptation!

-> Day5



BayPass - XtX + GEA to reveal local adaptation!



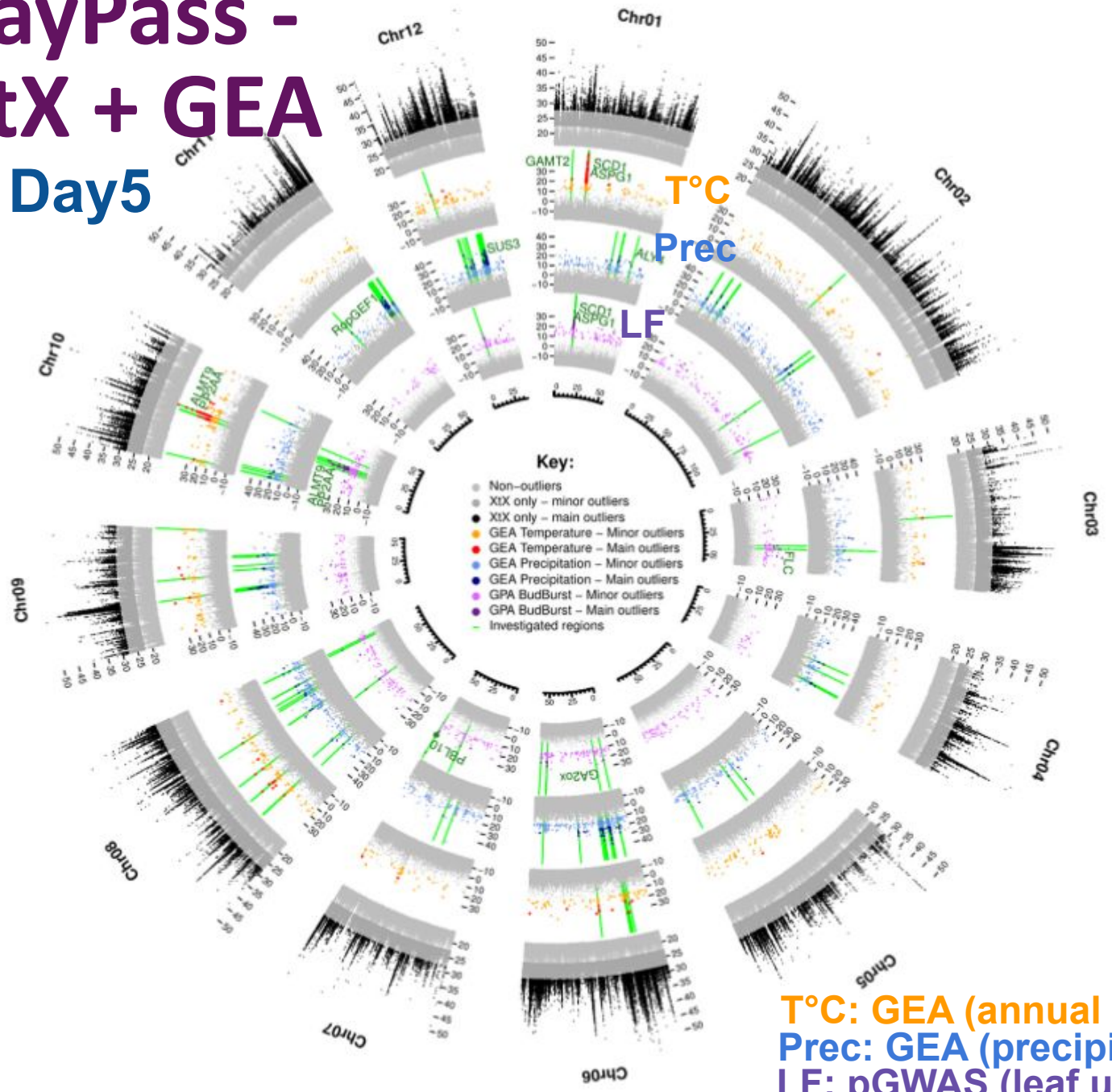
-> Day5

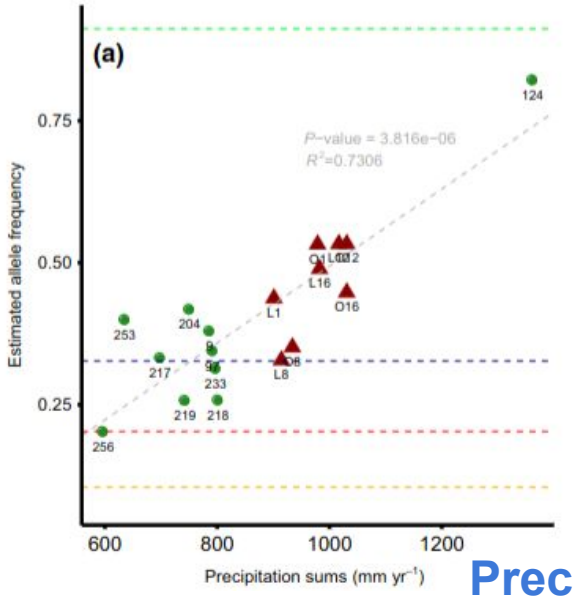
2 candidate genes:
SCD1 is known to be involved in stomatal responses to water stress, making it a plausible candidate for temperature-related associations.

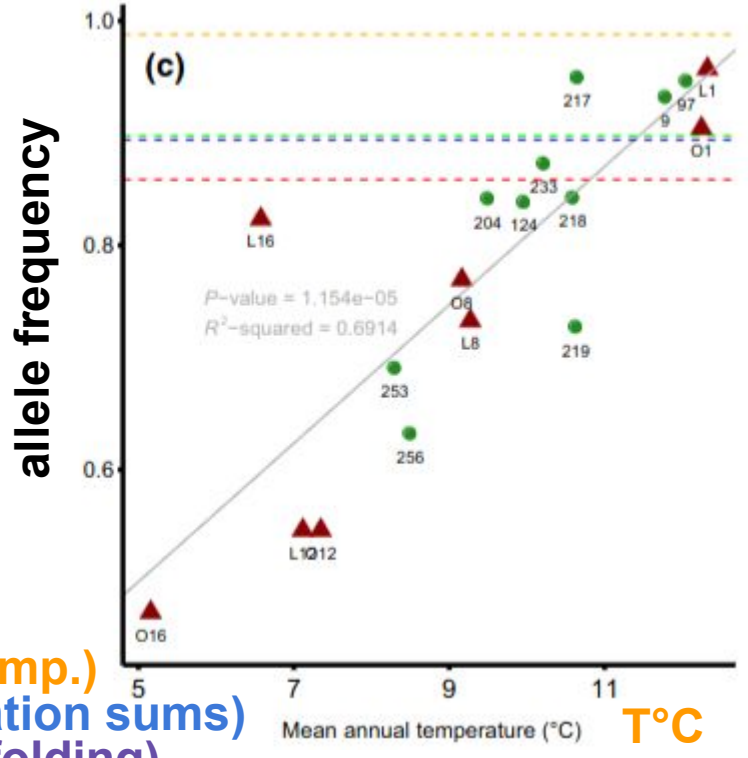
ASPG1 plays a key role in the production of gibberellins, which are crucial for various plant developmental processes, including seed dormancy.
-> determining the optimal timing of seed germination and, consequently, plant survival.

BayPass - XtX + GEA

-> Day5



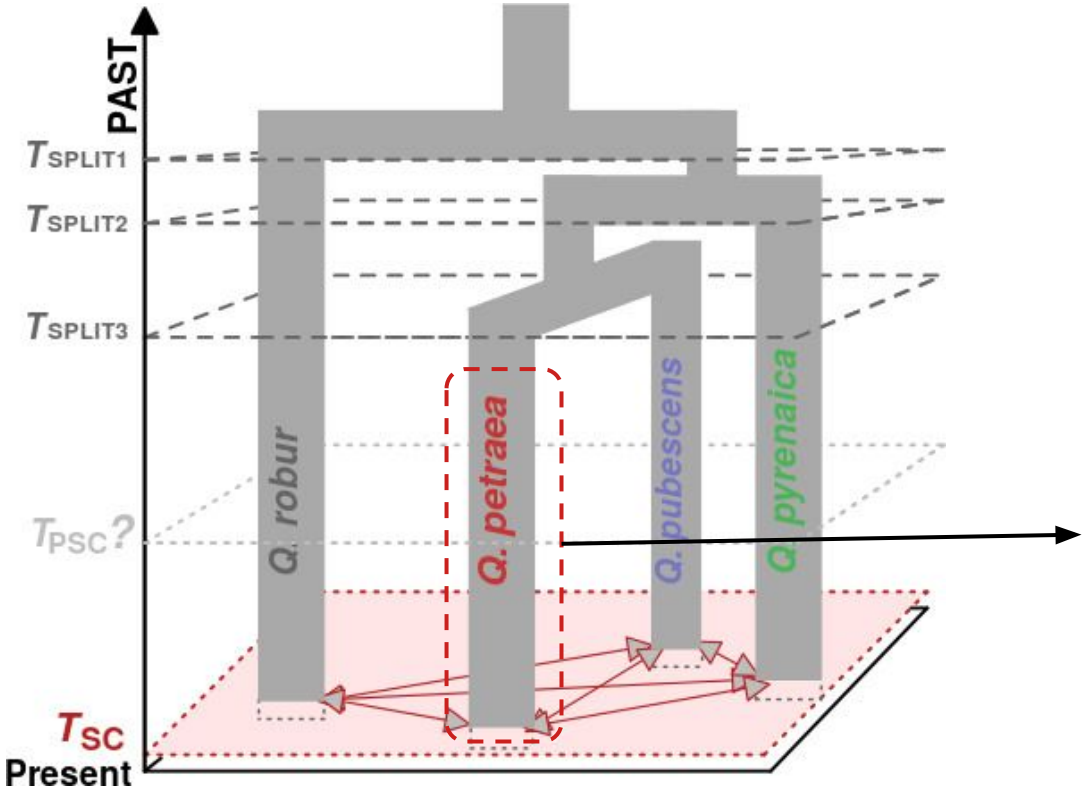




T°C: GEA (annual temp.)
Prec: GEA (precipitation sums)
LF: pGWAS (leaf unfolding)

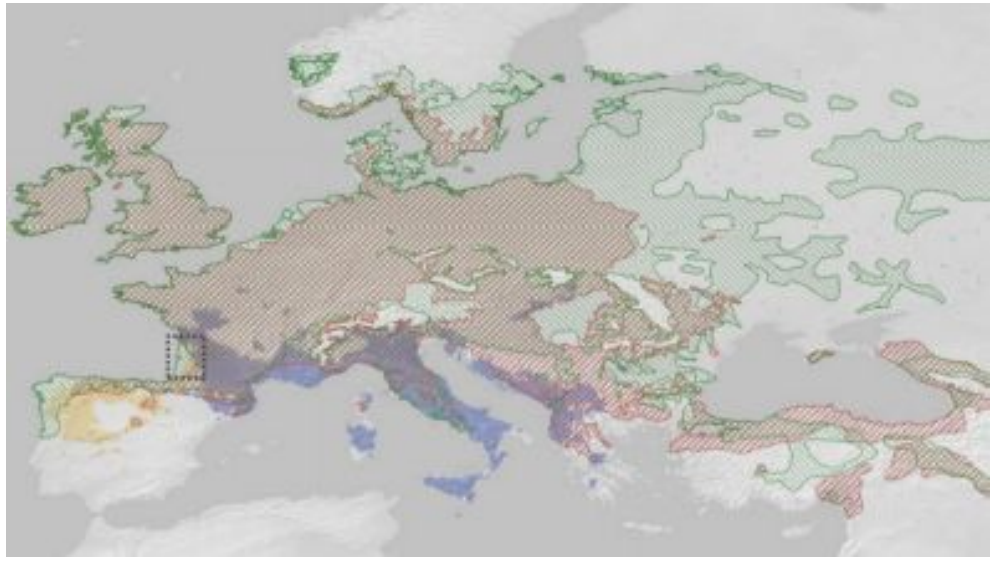
Back to the context: European white oaks

Demographic history **-> Day3**



Leroy et al. 2020a New Phytologist

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



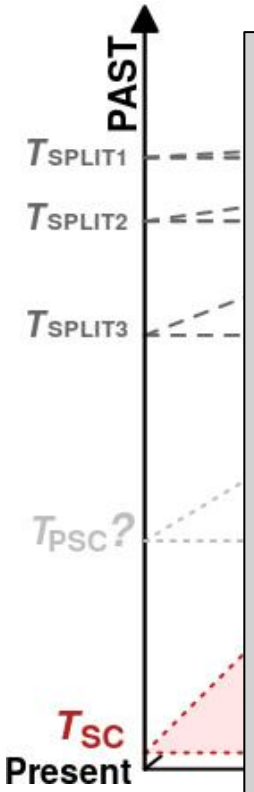
-> adaptive allele transferred from the most cold-adapted species (pedunculate oak, *Q. robur*) to sessile oak (*Q. petraea*)

***i.e.* adaptive introgression**

Leroy et al. 2020b New Phytologist

Back to the context: European white oaks

Demographic history **-> Day3**



My conclusion for the workshop today:
On oaks, it was crucial to combine all these research axes (population structure), demographic reconstruction, selection scans, GEA) to understand the evolutionary processes and how sessile oak populations have persisted in the Pyrenees.

A more general reminder:
It is important to keep in mind that the presence of footprints of past or contemporary selection in genomes biases demographic analyses, and conversely, methods for detecting traces of selection are biased by complex demographic histories

The future:
jointly reconstructing past demography and detecting footprints of selection at once

…red)

…most
…ak, *Q.*

***i.e.* adaptive introgression**

Leroy et al. 2020a New Phytologist

Leroy et al. 2020b New Phytologist