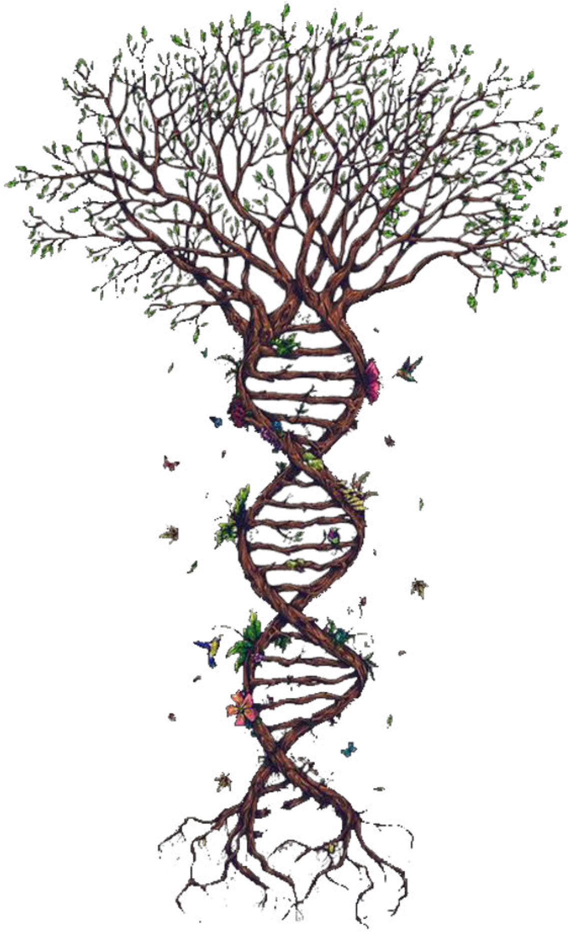
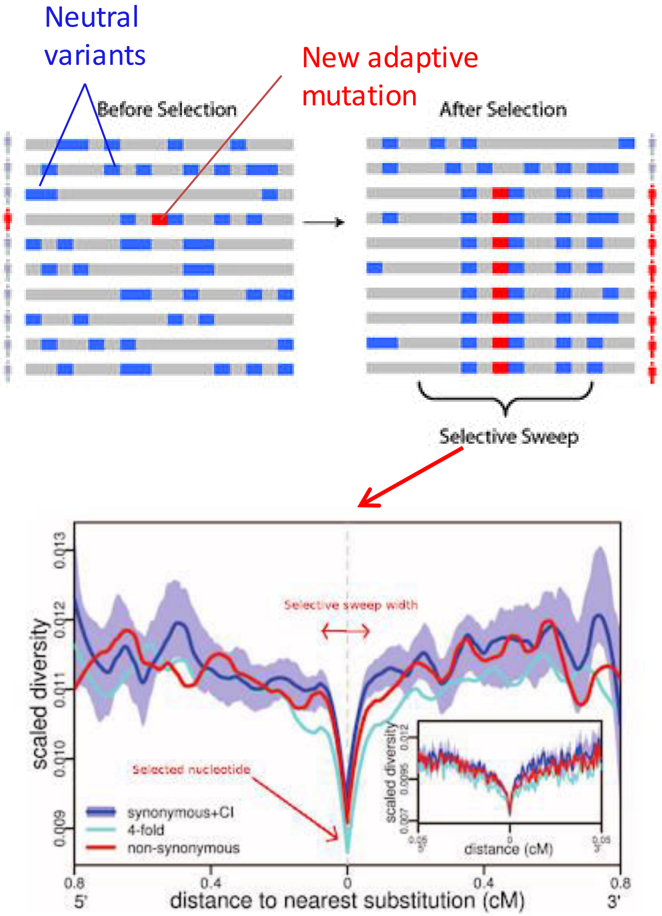
Part 2: Detecting footprints of natural selection in genomes



Thibault Leroy, University Assistant
Genomic approaches
23 April 2020

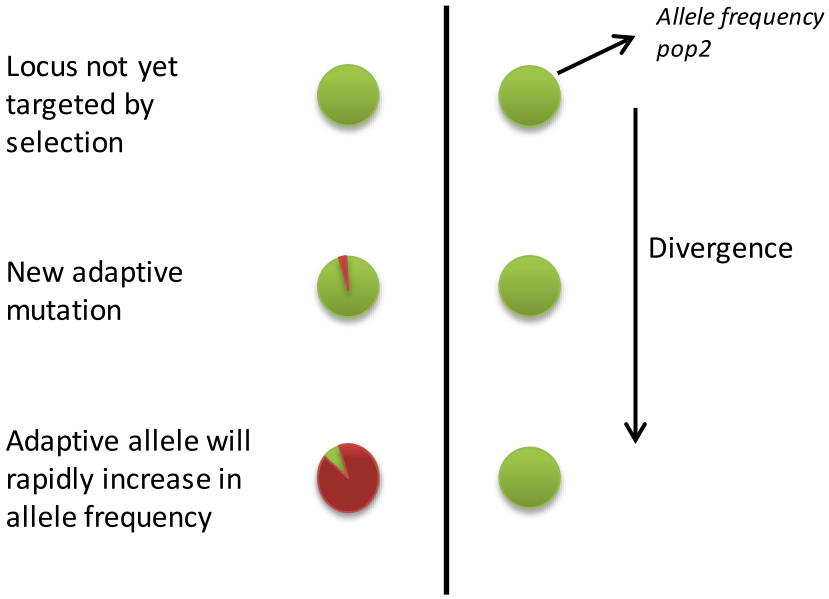
Methods are divided into two main groups:

Selective sweeps
(within-population variation)



Reduction of the diversity at the selected locus (+ SNP in close vicinity = linkage disequilibrium)
Extended haplotype homozygosity (EHH)

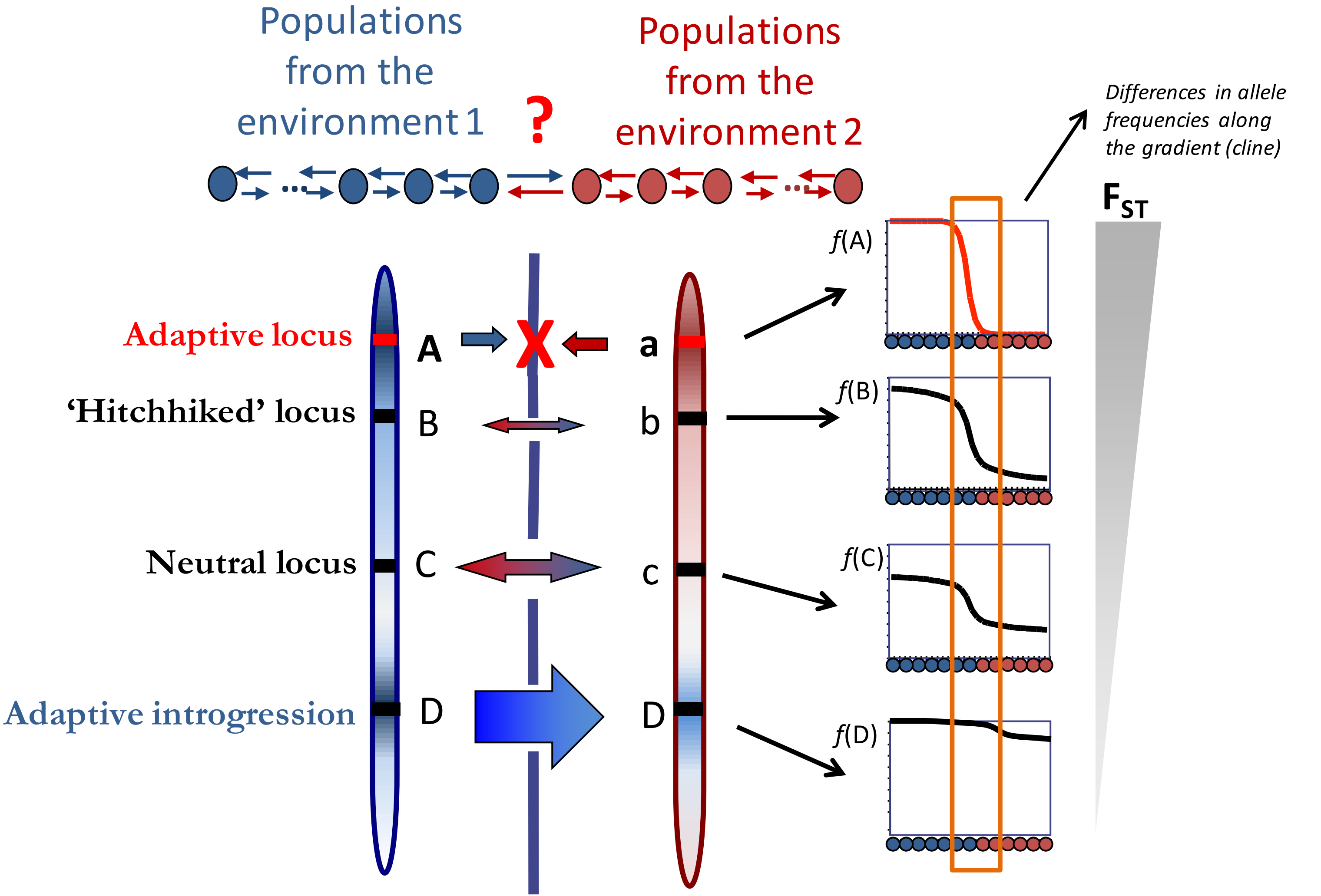
Genetic differentiation
(between populations)


Locus not yet targeted by selection
New adaptive mutation
Adaptive allele will rapidly increase in allele frequency
Allele frequency pop2
Divergence

Extreme allele frequency differencies between the two populations at the selected locus

SNP in close vicinity with the targeted SNPs also exhibit strong differences in allele frequency

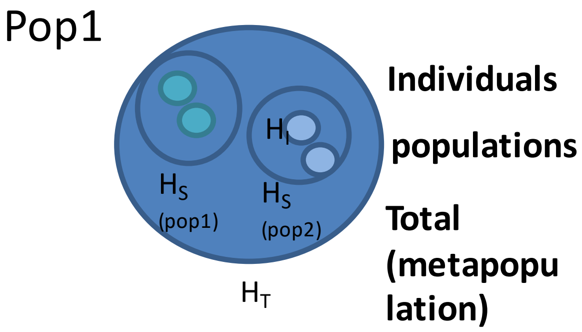
Genetic differentiation


Populations from the environment 1
?
Populations from the environment 2
Differences in allele frequencies along the gradient (cline)
F_{ST}
Adaptive locus
A
a
'Hitchhiked' locus
B
b
Neutral locus
C
c
Adaptive introgression
D
D
f(A)
f(B)
f(C)
f(D)

Modified from Bierne (2001)

Fixation indices (F-statistics, F_{ST} in particular) <-> inbreeding

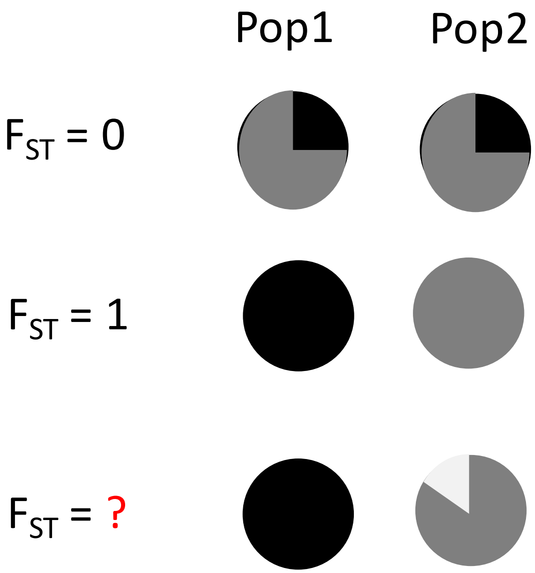
In nature, individuals rarely mate completely at random because of some geographically or ecologically-restricted mating among individuals. Such a non-random population mating drive differentiation among populations over the whole genome (i.e. population structure).



F_{ST} = deviation in allele frequencies among populations relative to the expectation assuming panmixtia (random mating)

$$F_{ST} = (H_T – H_S)/H_T$$
$$= 1 – H_S/H_T \quad \text{with } H_s=2p_{S(pop)}q_{S(pop)} \;\&\; H_T=2p_{Total}q_{Total}$$

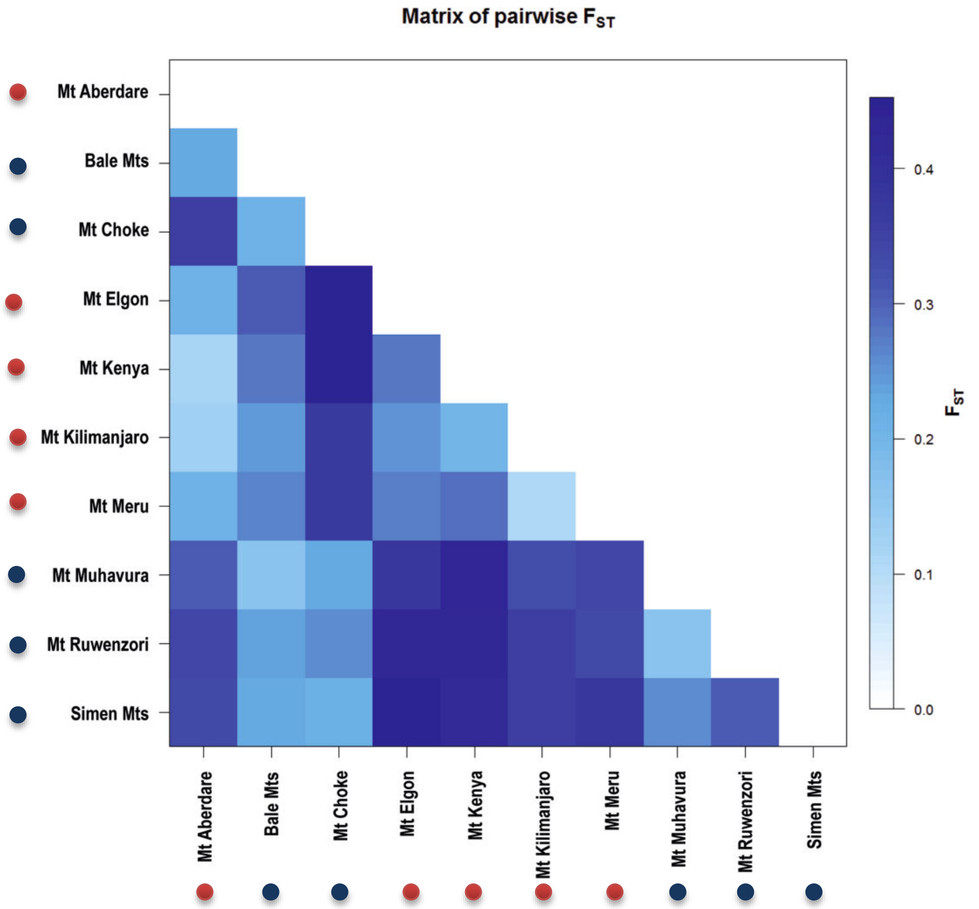
across multiple populations:
average H_S (here average between $H_{S(pop1)}$ & $H_{S(pop2)}$)



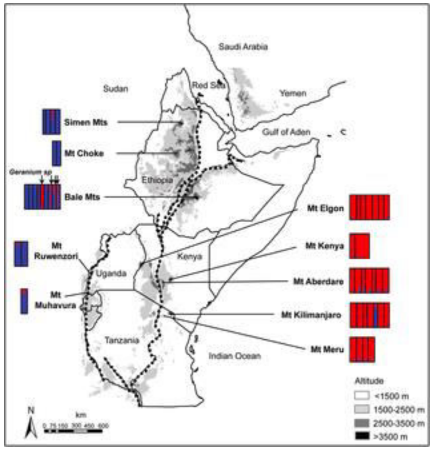
F-statistics are central in population genetics:
$F_{IS} = 1 – H_I/H_S = 1 – f12 / 2p_{S(pop)}q_{S(pop)}$
(deviation from random mating within the subpopulation, i.e. difference between observed and expected heterozygosity)

Among population variation in F_{ST}

Given that the large majority of SNPs in the genome are neutral, the pairwise population differentiations computed over the whole dataset are representative of the population structure (i.e. demographic history contributing to past or present departure from panmixia of a given population)

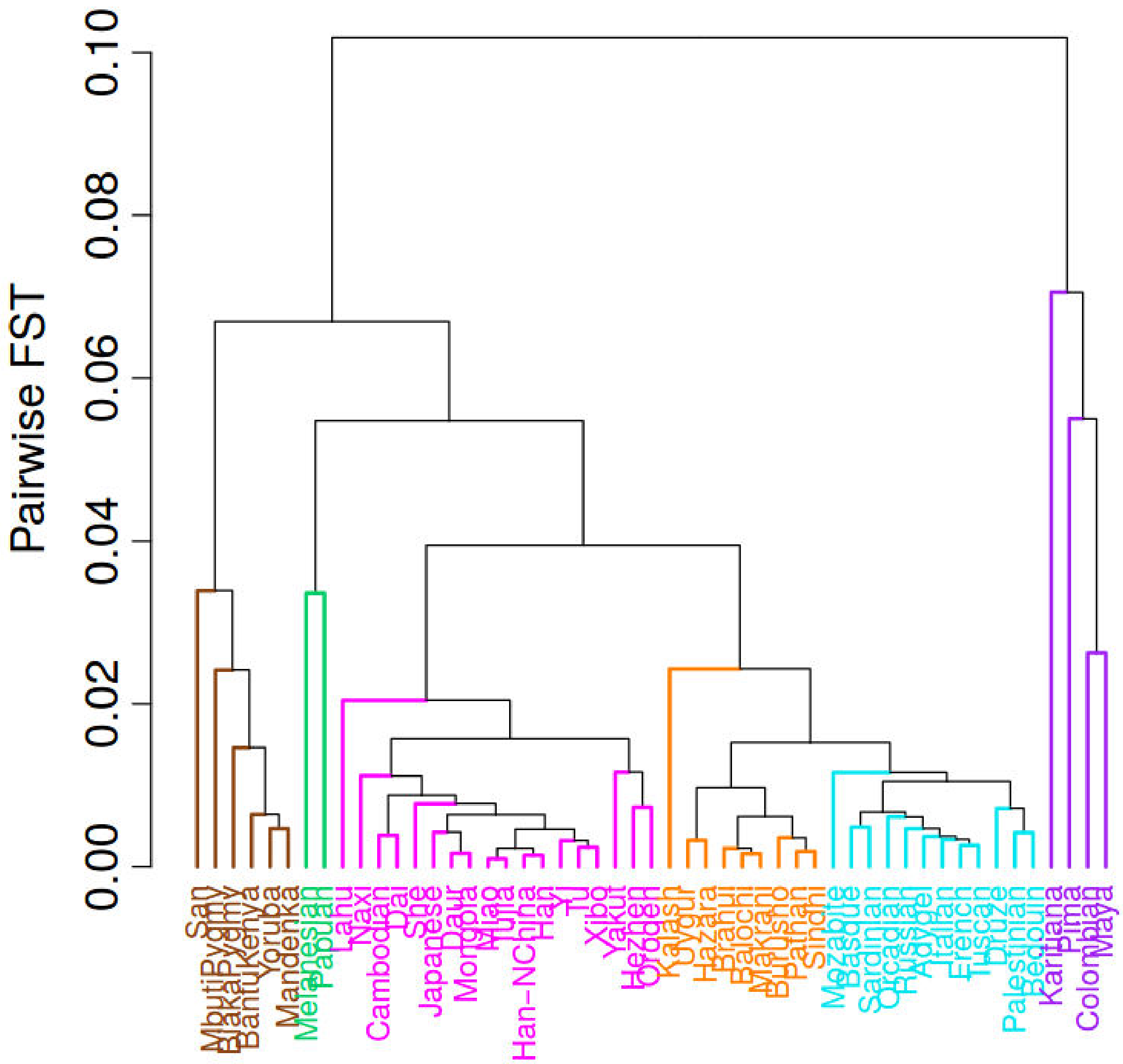


Geranium arabicum/kilimandscharicum



Wondimu et al. 2017 Plos One

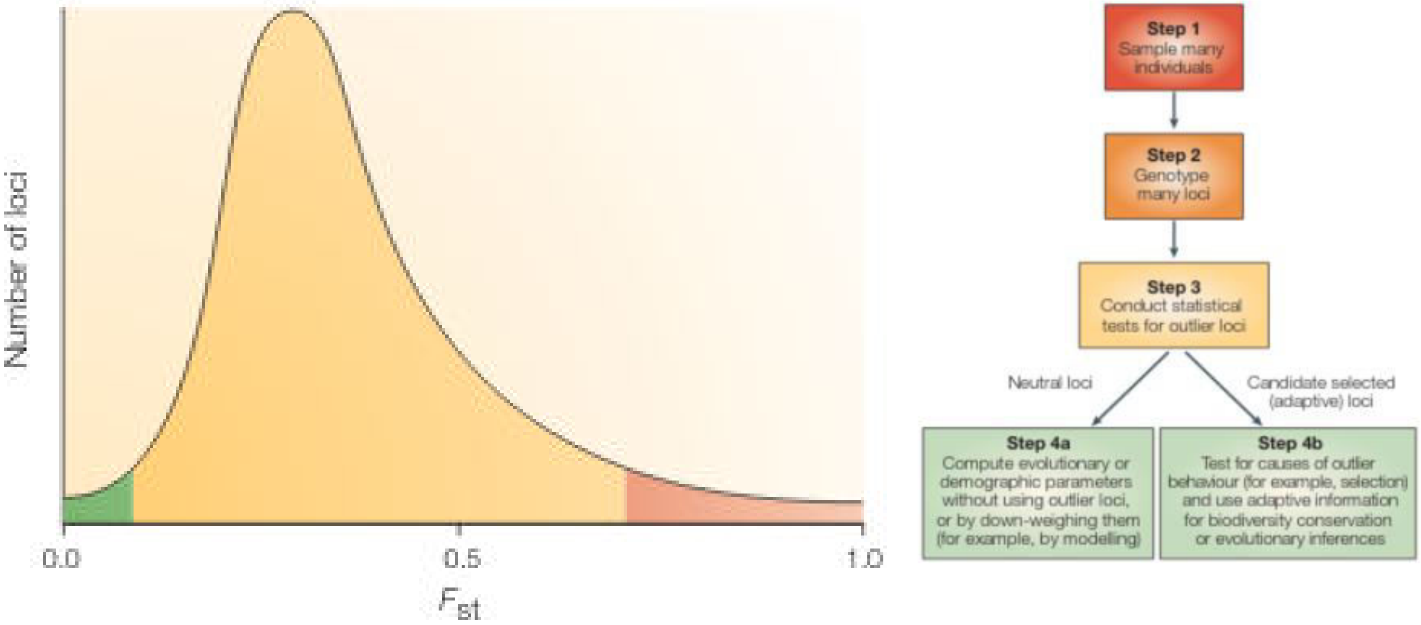
Among population variation in F_{ST}



1,035 individuals
377 SSR
Kitada et al. 2020
bioRxiv

Among locus variation in F_{ST}

Empirical distribution of F_{ST} among all genotyped loci



Step 1 Sample many individuals

Step 2 Genotype many loci

Step 3 Conduct statistical tests for outlier loci

Step 4a Compute evolutionary or demographic parameters without using outlier loci, or by down-weighing them (for example, by modelling)

Step 4b Test for causes of outlier behaviour (for example, selection) and use adaptive information for biodiversity conservation or evolutionary inferences

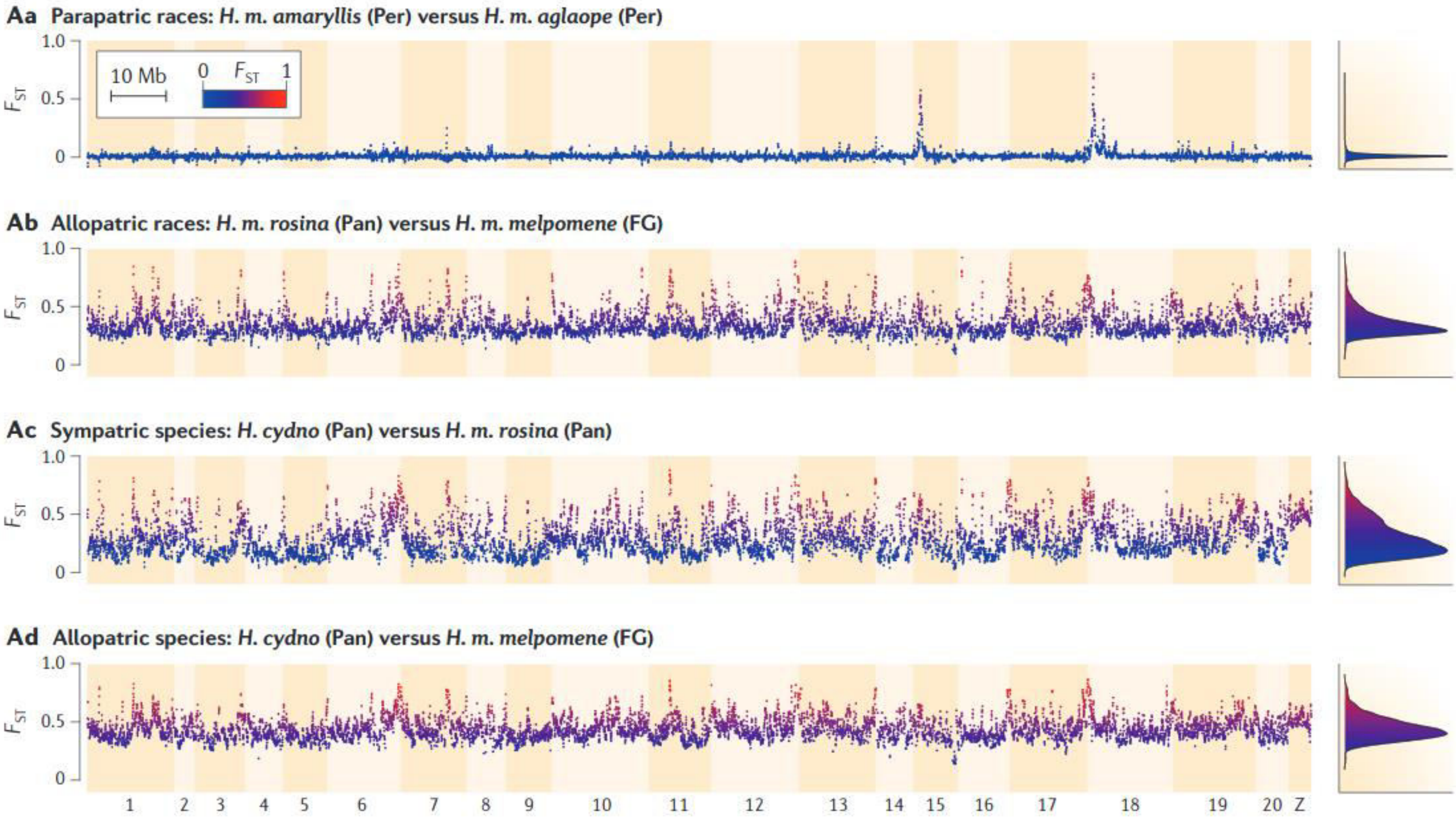
Figure 2 | **Identifying outlier behaviour.** A hypothetical distribution of F_{st} (genetic divergence) and F_{is} (deviation from Hardy–Weinberg proportions) among neutral loci that are sampled from across the genome. Locus-specific effects lead to a few outlier loci with a highly divergent F_{st} or F_{is} value relative to most other loci across the genome. Modified with permission from REF. 1 © (2001) Annual Reviews.

Loci targeted by natural selection can be on both tailed of the distribution ('outlier loci'):

Very low F_{ST} levels = putative loci under balancing selection (less differentiation than expected for a neutral marker)

Very high F_{ST} levels = putative loci under positive selection (more differentiation than expected for a neutral marker)

Among locus variation in Fst

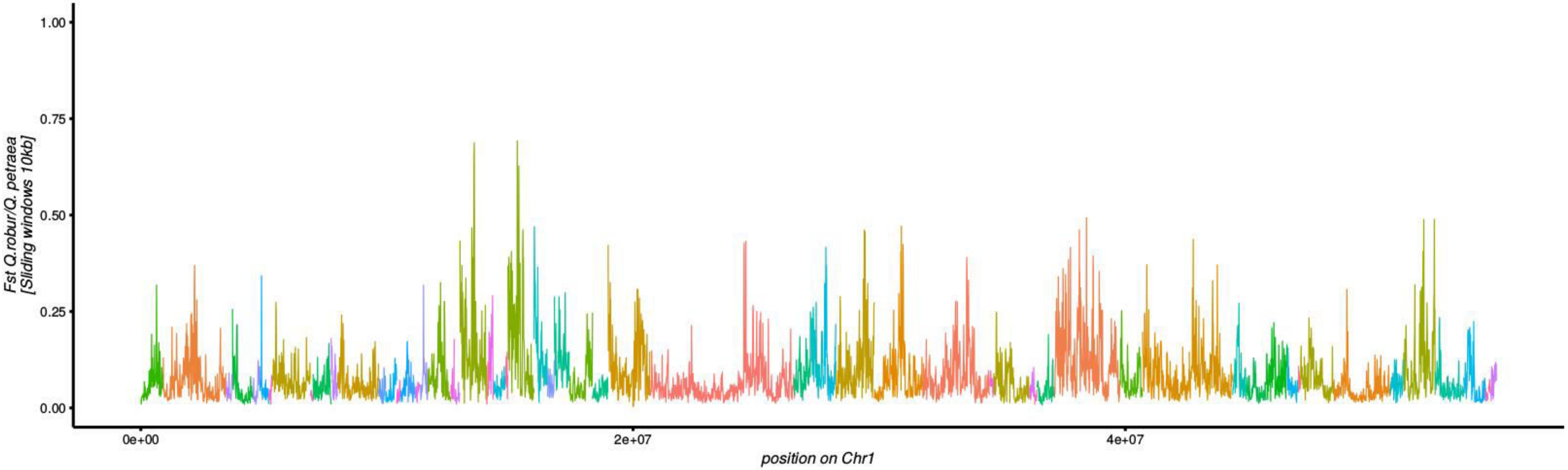


The plot showing the variation of the differentiation along chromosomes are called 'Manhattan plots'

Seehausen et al. 2014 Nature Review Genetics

Among locus variation in F_{ST}

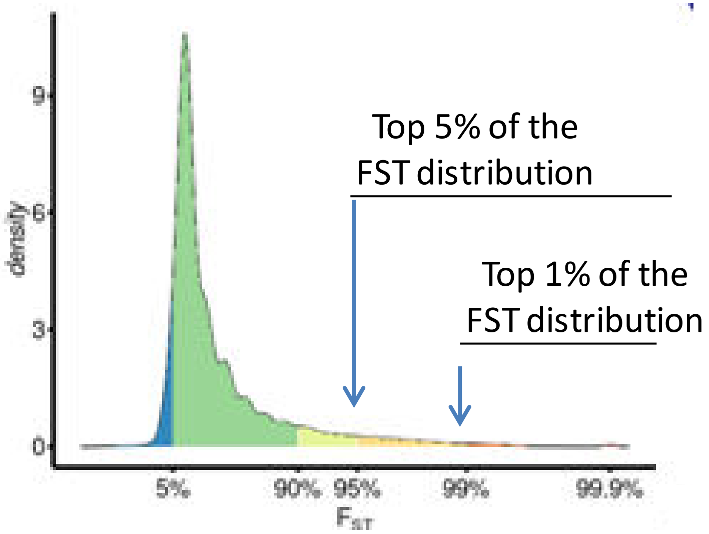
The same approach can be used to compute F_{ST} using a sliding windows approach (F_{ST} is computed based on all variants of the windows)



Leroy et al. 2020a New phytologist

How to detect outliers based on the distribution of F_{ST}?

A widely used way is to consider only variants exhibiting the top e.g. 1% values.

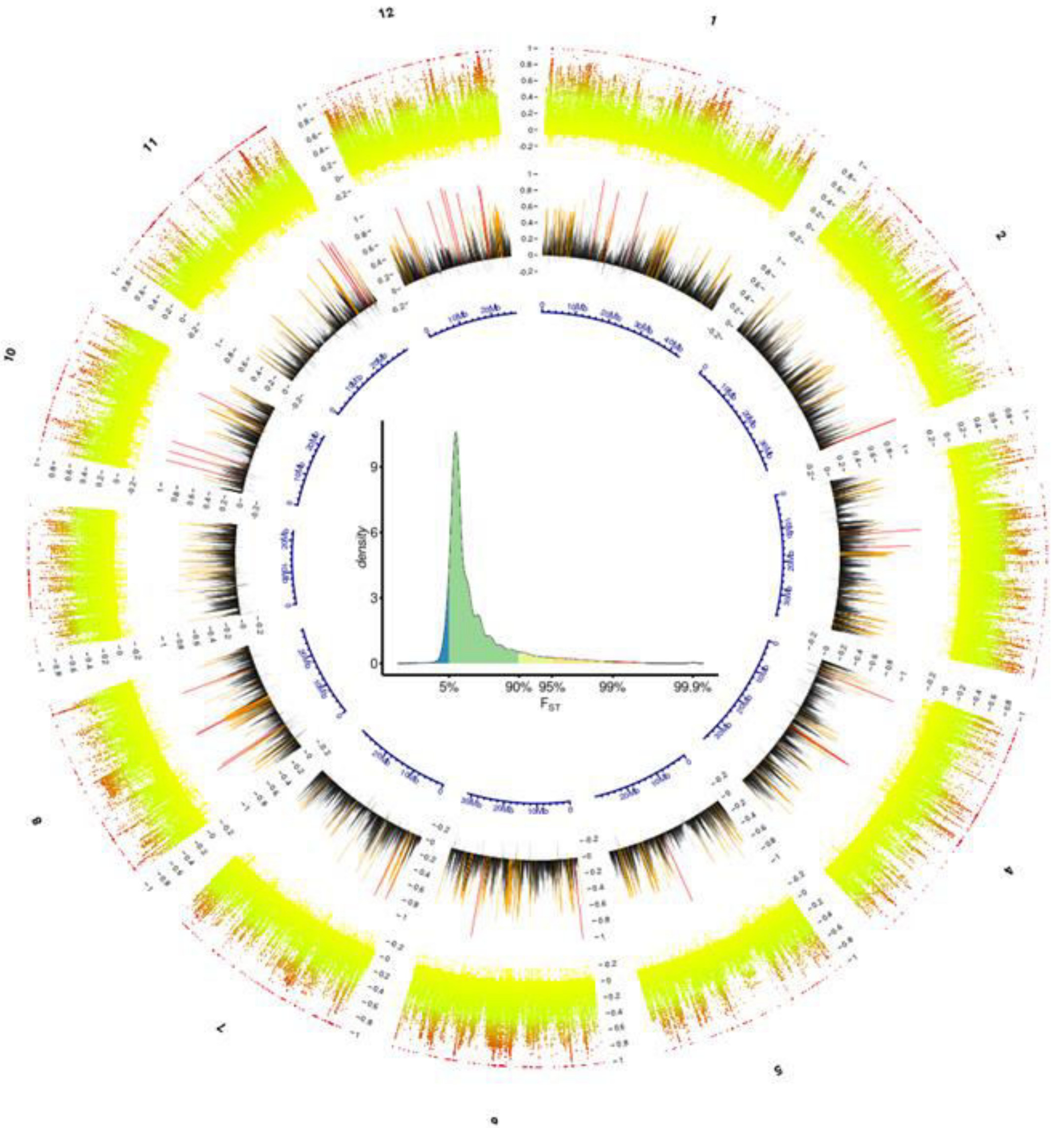


Top 5% of the FST distribution

Top 1% of the FST distribution

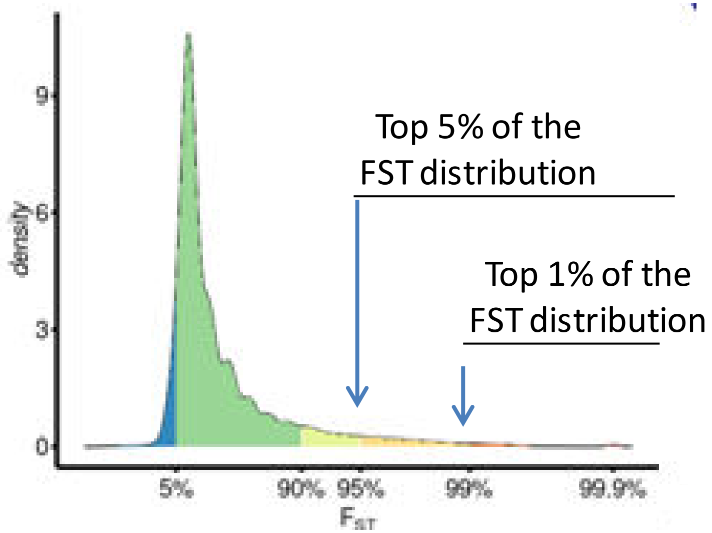
How to detect outliers based on the distribution of F_{ST}?

A widely used way is to consider only variants exhibiting the top 1% values.



How to detect outliers based on the distribution of F_{ST}?

A widely used way is to consider only variants exhibiting the top 1% values.



Top 5% of the FST distribution

Top 1% of the FST distribution

The main problem of such an approach is that you assume that the threshold you use corresponds to the proportion of loci that were targeted by natural selection (or that are in close vicinity with these genes)

Assume two populations evolving under strict neutrality, using this strategy you will always be able to find the top 1% most differentiated loci. How do we resolve this issue?

How to detect outliers based on the distribution of F_{ST}?

Identifying footprints of selection ↔ disentangle locus-specific from demographic effects on allele frequency differences

Difficult to do with the empirical F_{ST} distribution itself.

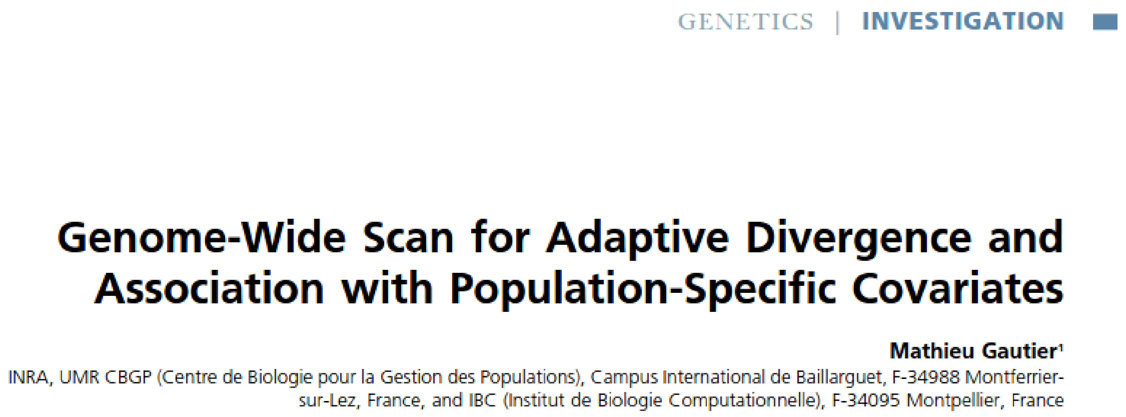
With the notable exception of the strategy developped by Whitlock & Lotterhos, which are based on a trimmed distribution of F_{ST} values to infer the distribution of F_{ST} for neutral markers.

The best to do is to define the neutral distribution

- ✓ Either theoretically (includes model assumptions)
- ✓ Or through neutral simulations (includes demographic assumptions)

The general idea is to generate a neutral expectation and to identify the loci that deviate from this neutral expectation ("outliers")

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

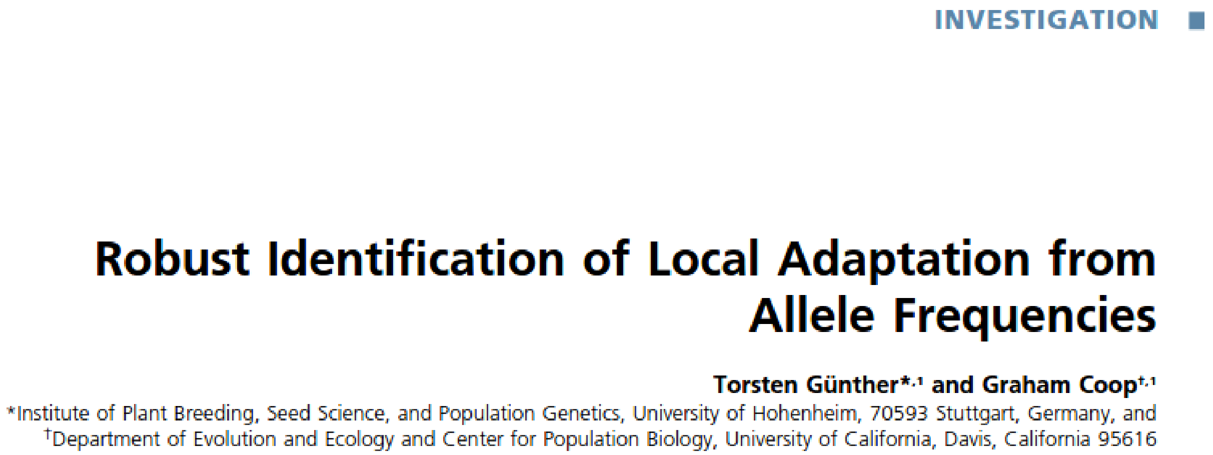


GENETICS | INVESTIGATION

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

Mathieu Gautier[1]

INRA, UMR CBGP (Centre de Biologie pour la Gestion des Populations), Campus International de Baillarguet, F-34988 Montferrier-sur-Lez, France, and IBC (Institut de Biologie Computationnelle), F-34095 Montpellier, France



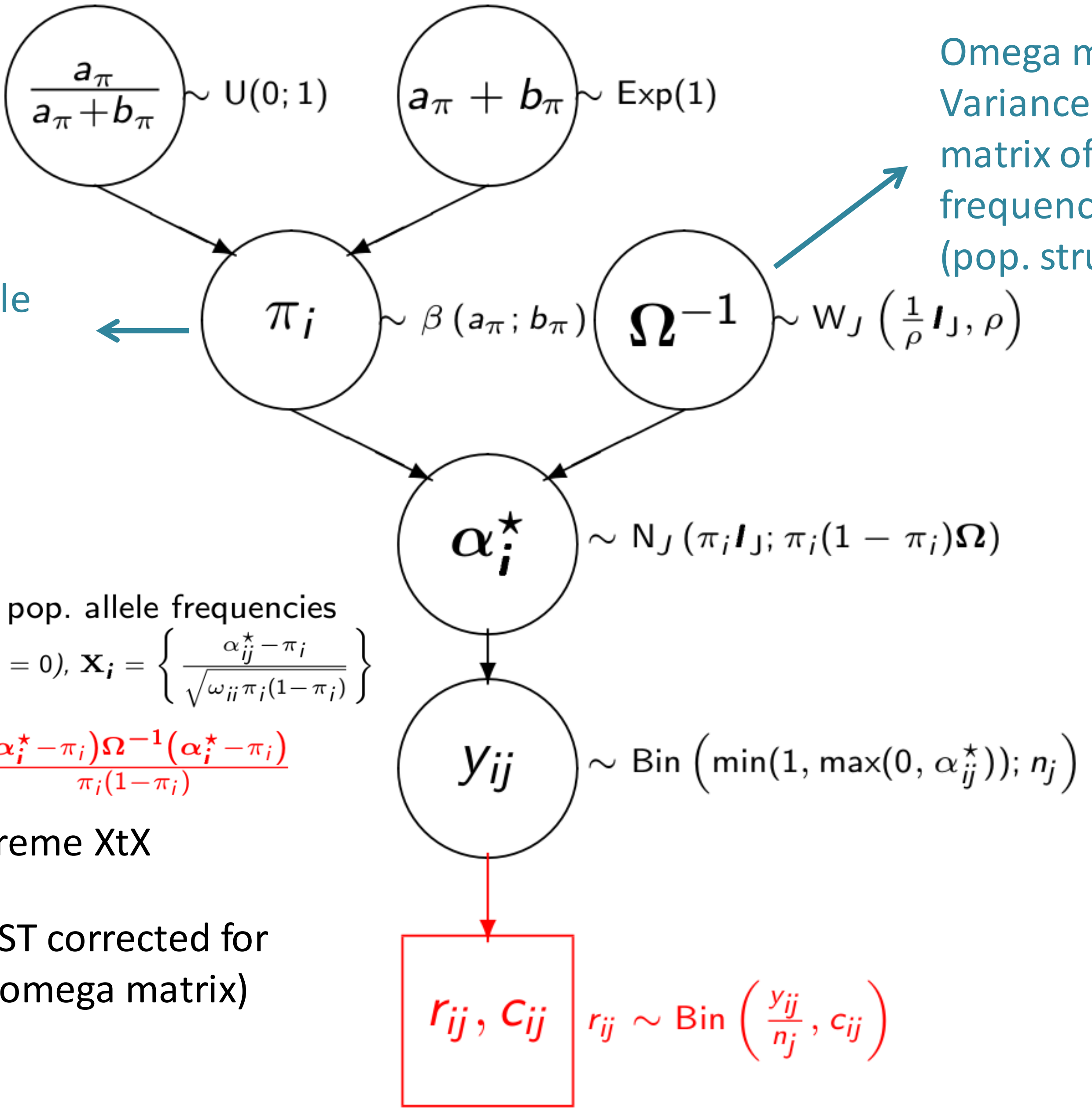
INVESTIGATION

Robust Identification of Local Adaptation from Allele Frequencies

Torsten Günther*,[1] and Graham Coop†,[1]

*Institute of Plant Breeding, Seed Science, and Population Genetics, University of Hohenheim, 70593 Stuttgart, Germany, and
†Department of Evolution and Ecology and Center for Population Biology, University of California, Davis, California 95616

The general strategy developped in BayPass (Mathieu Gautier, 2015) ~ Bayenv2 (Coop)


$\frac{a_\pi}{a_\pi+b_\pi} \sim \mathrm{U}(0;1)$
$a_\pi + b_\pi \sim \mathrm{Exp}(1)$
$\pi_i \sim \beta\left(a_\pi; b_\pi\right)$
$\boldsymbol{\Omega}^{-1} \sim \mathrm{W}_J\left(\frac{1}{\rho}\boldsymbol{I}_J, \rho\right)$
$\boldsymbol{\alpha}_{\boldsymbol{i}}^{\star} \sim \mathrm{N}_J\left(\pi_i\boldsymbol{I}_J; \pi_i(1-\pi_i)\boldsymbol{\Omega}\right)$
$y_{ij} \sim \mathrm{Bin}\left(\min(1, \max(0, \alpha_{ij}^{\star})); n_j\right)$
r_{ij}, c_{ij}
$r_{ij} \sim \mathrm{Bin}\left(\frac{y_{ij}}{n_j}, c_{ij}\right)$

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

Omega matrix: Variance-Covariance matrix of allele frequencies (pop. structure)

Ancestral allele frequency (unobserved)

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

$\mathbf{X}_{\boldsymbol{i}} \simeq$ vector of scaled pop. allele frequencies

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

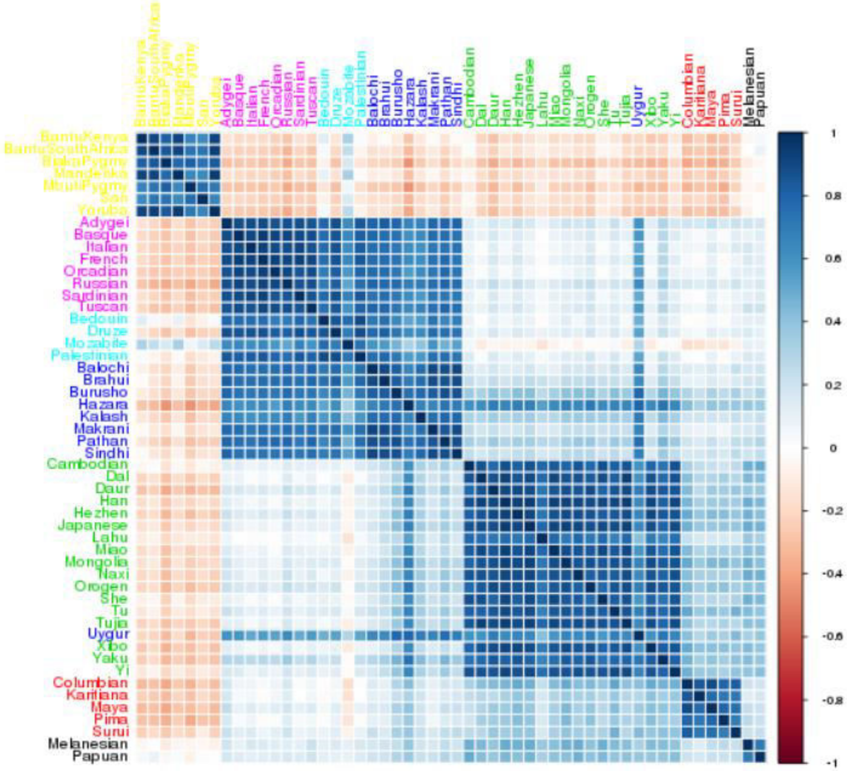
$$\mathrm{X}^{\mathrm{t}}\mathrm{X}_i = \mathrm{Var}\left(\mathbf{X}_{\boldsymbol{i}}\right) = \frac{(\boldsymbol{\alpha}_{\boldsymbol{i}}^{\star} - \pi_i)\boldsymbol{\Omega}^{-1}(\boldsymbol{\alpha}_{\boldsymbol{i}}^{\star} - \pi_i)}{\pi_i(1-\pi_i)}$$

Selected SNPs = Extreme XtX

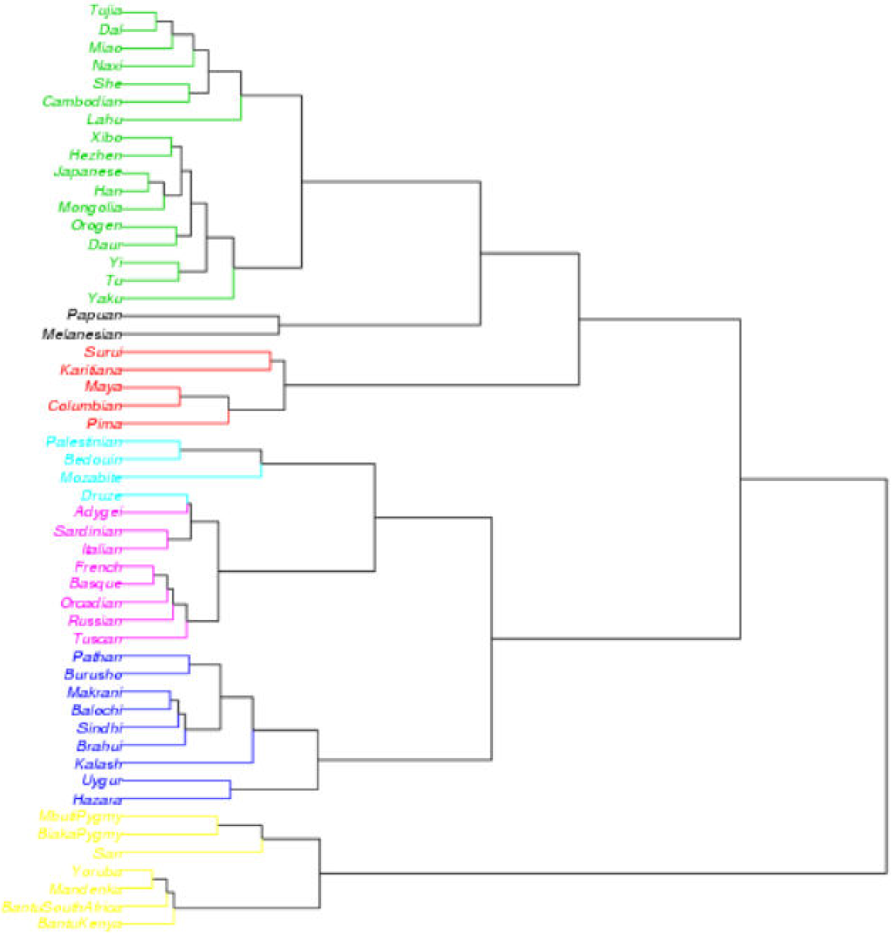
XtX= SNP–specific FST corrected for population history (omega matrix)

Omega matrix: 52 human populations, 2333 autosomal SNPs

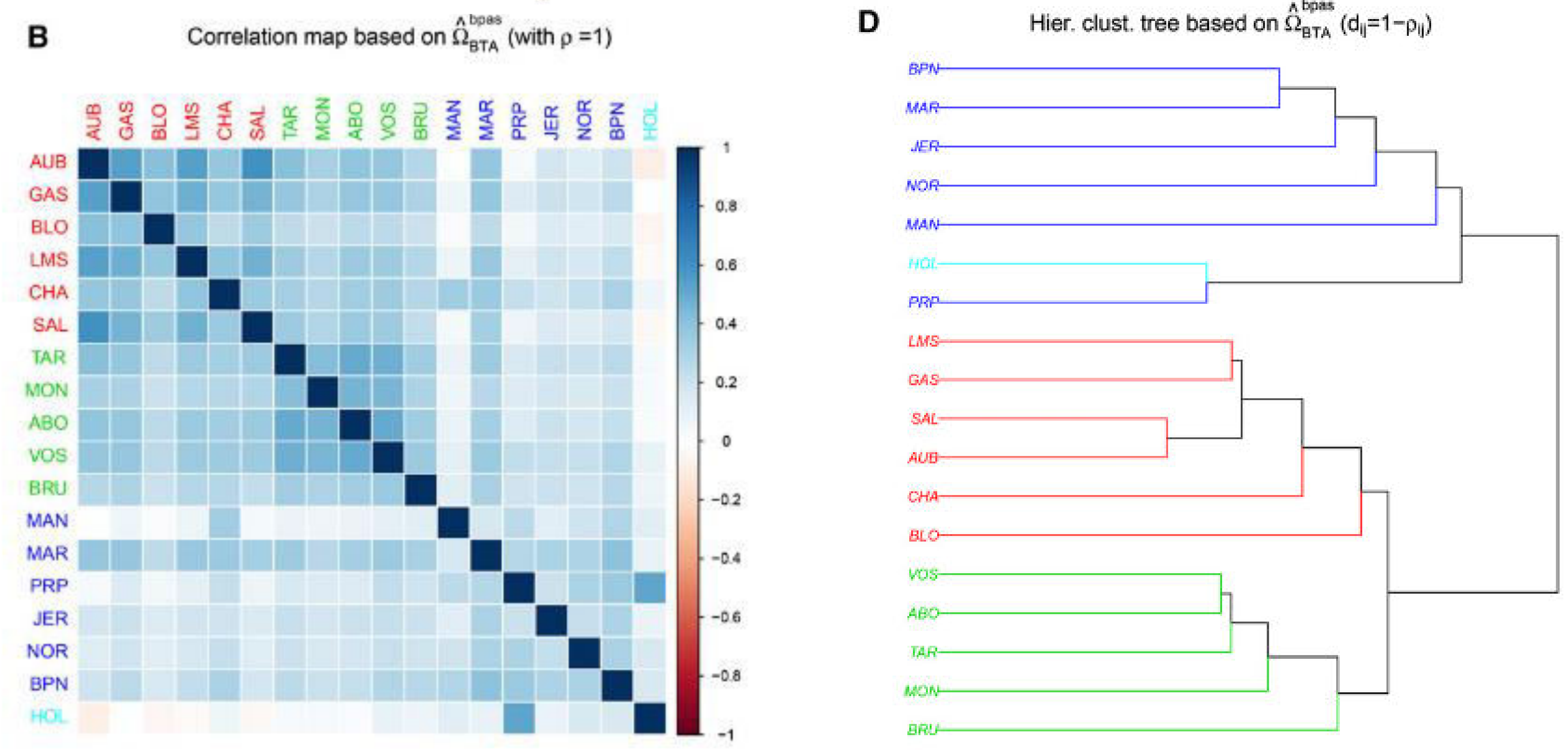
variance-covariance matrix, here shown as after a cov2cor transformation



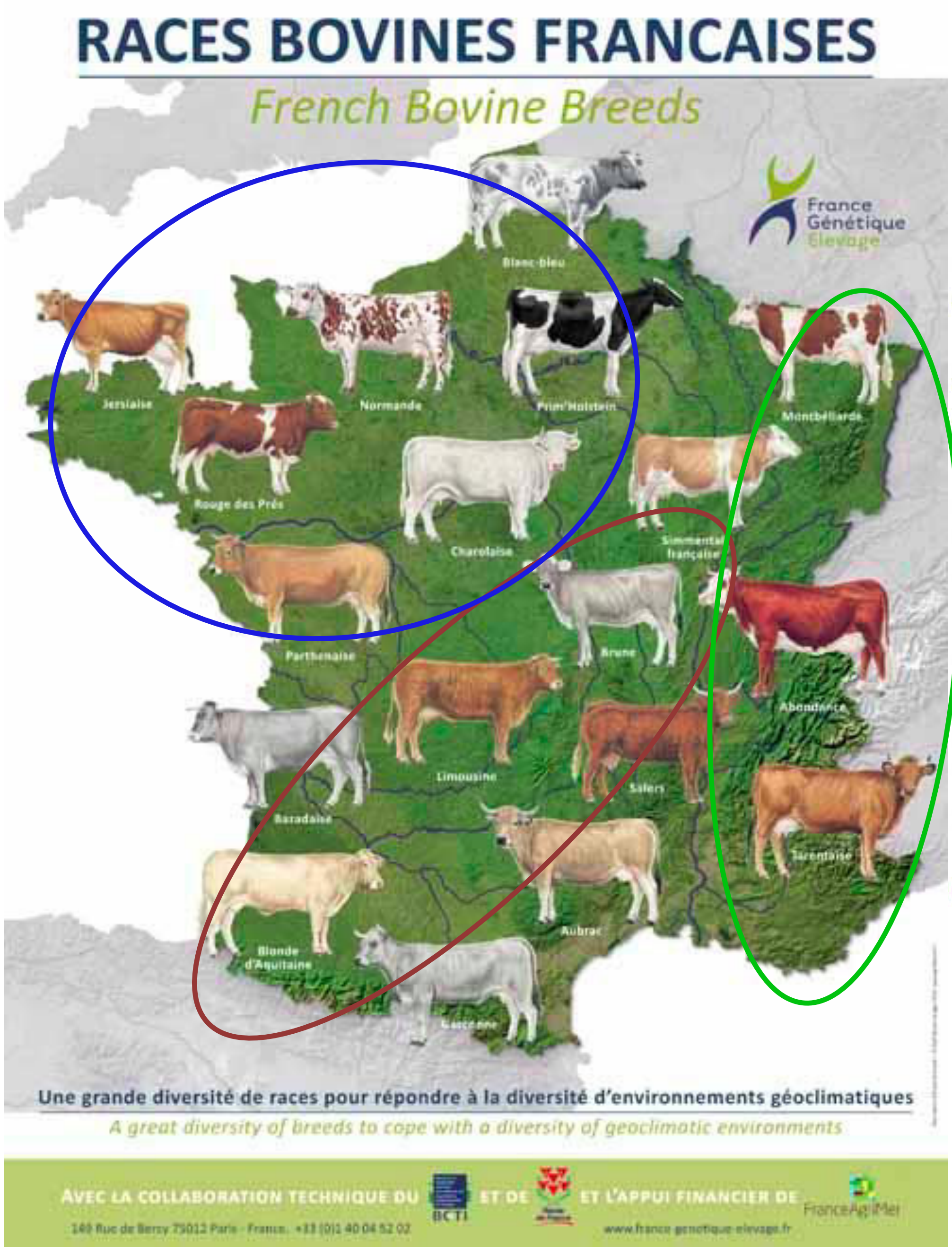
D) Hier. clust. tree based on $\hat{\Omega}^{bpas}_{HSA}$ ($d_{ij}=1-\rho_{ij}$)

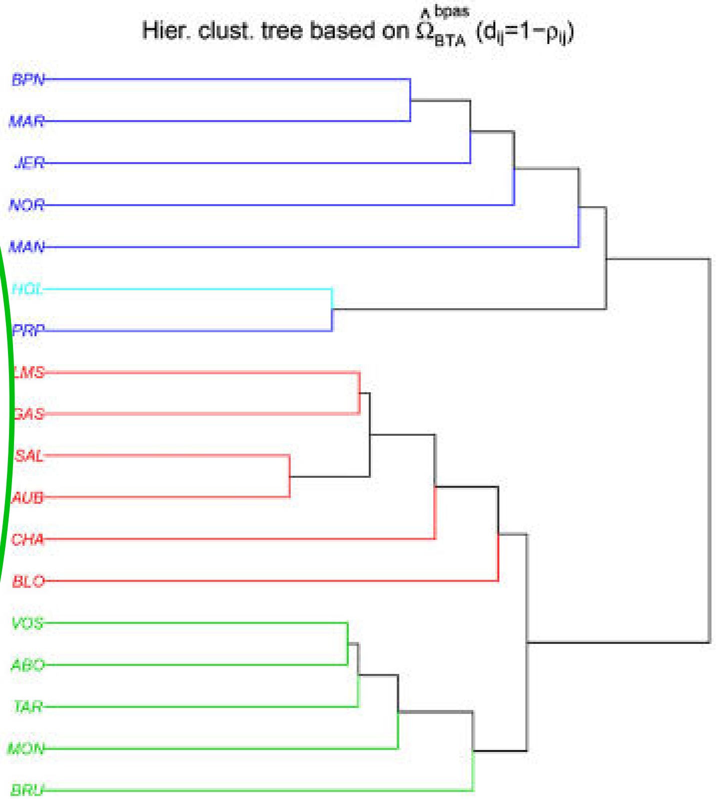


Omega matrix (variance-covariance matrix, here shown as after a cov2cor transformation)

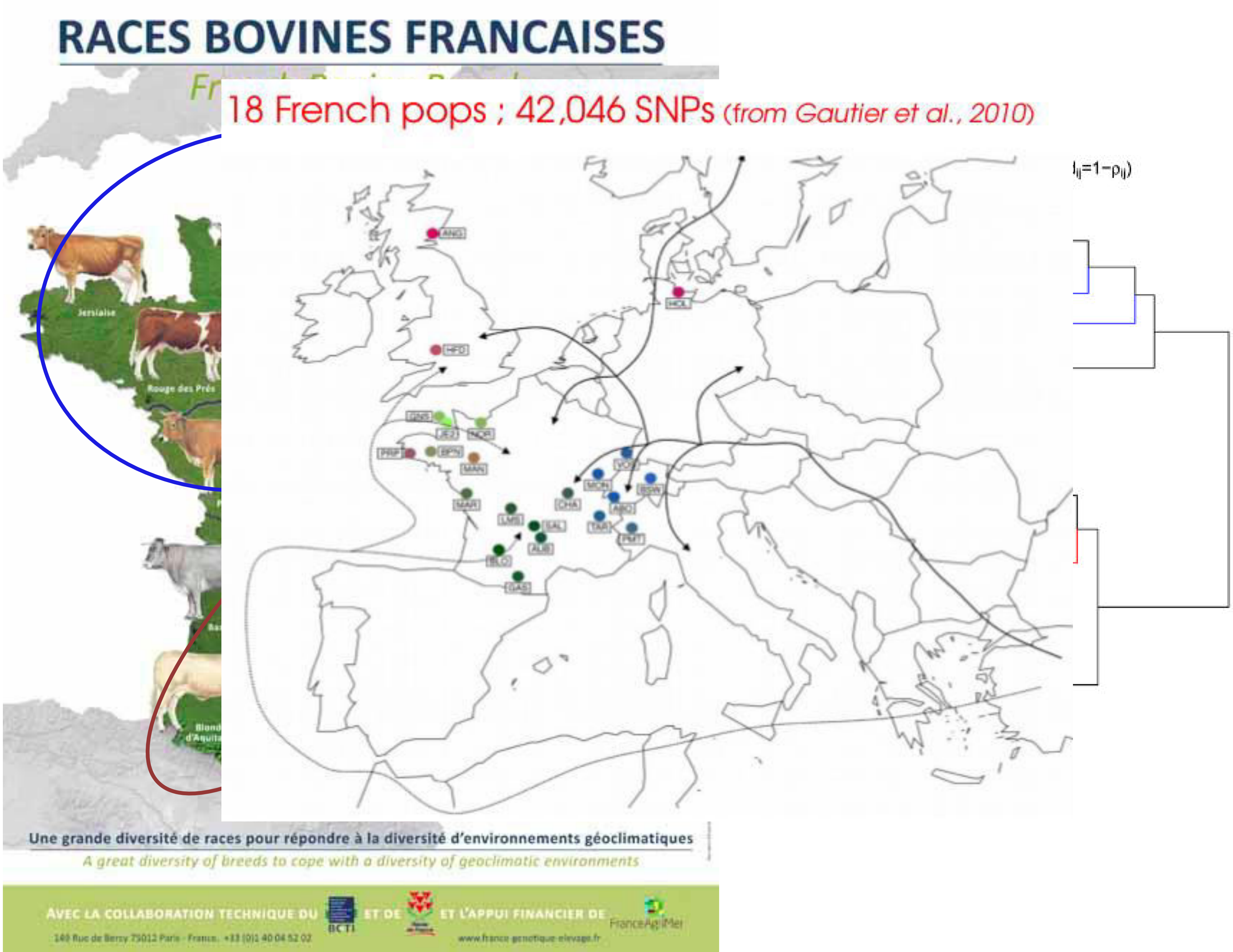

B
Correlation map based on $\hat{\Omega}_{BTA}^{bpas}$ (with ρ =1)
AUB
GAS
BLO
LMS
CHA
SAL
TAR
MON
ABO
VOS
BRU
MAN
MAR
PRP
JER
NOR
BPN
HOL
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1
D
Hier. clust. tree based on $\hat{\Omega}_{BTA}^{bpas}$ ($d_{ij}=1-\rho_{ij}$)
BPN
MAR
JER
NOR
MAN
HOL
PRP
LMS
GAS
SAL
AUB
CHA
BLO
VOS
ABO
TAR
MON
BRU

Omega matrix (variance-covariance matrix, here shown as after a cov2cor transformation)




Hier. clust. tree based on $\hat{\Omega}^{bpas}_{BTA}$ $(d_{ij}=1-\rho_{ij})$

Omega matrix (variance-covariance matrix, here shown as after a cov2cor transformation)

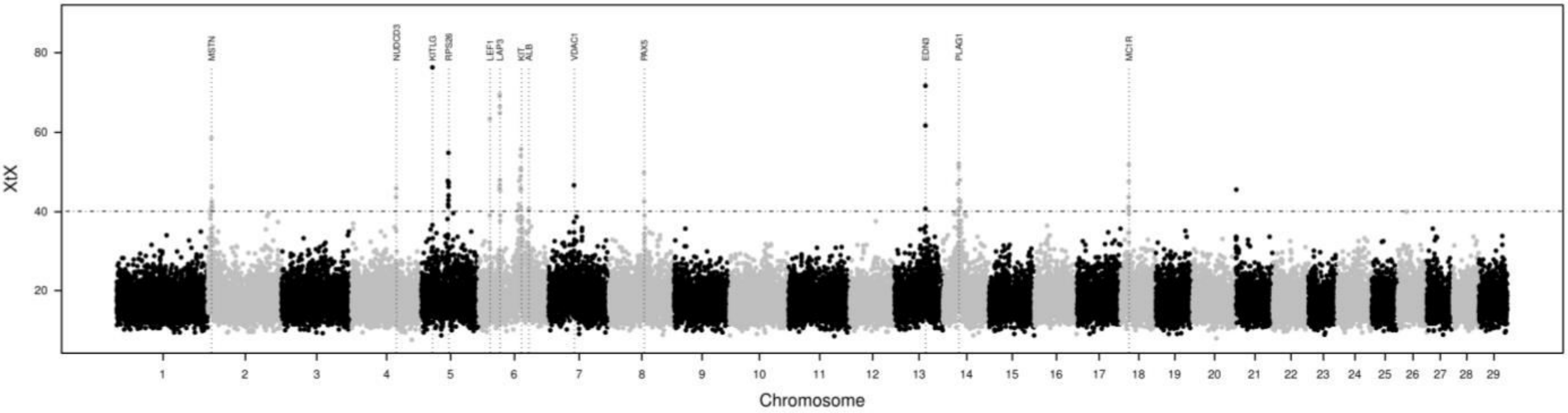


18 French pops ; 42,046 SNPs *(from Gautier et al., 2010)*

RACES BOVINES FRANCAISES
Fr
Jersiaise
Rouge des Prés
Blond d'Aquita
Une grande diversité de races pour répondre à la diversité d'environnements géoclimatiques
A great diversity of breeds to cope with a diversity of geoclimatic environments
AVEC LA COLLABORATION TECHNIQUE DU
BCTI
ET DE
ET L'APPUI FINANCIER DE
FranceAgriMer
149 Rue de Bercy 75012 Paris - France. +33 (0)1 40 04 52 02
www.france-genetique-elevage.fr
ANG
HOL
HFD
GNS
JER
NOR
PRP
BPN
MAN
MAR
LMS
SAL
AUB
BLO
GAS
CHA
MON
VOS
ABO
BSW
TAR
PMT
$_{ij}=1-\rho_{ij})$

Genome scans (XtX) - cattle

XtX ~ F_{ST} accounting for the population structure



XtX is no longer a value between 0 and 1.
SNPs that are more likely under selection are those exhibting the most elevated XtX values.

⟶ **To identify the proportion of outliers, a neutral calibration is still needed**

Now it is easier to do because we already infered the variance-covariance matrix of allele frequencies (omega matrix)

Neutral calibrations – XtX metrics

Generate “Pseudo-Observed Datasets” (PODs) assuming the parameters used by the core model (in particular the omega matrix).

Of course, all simulated SNPs (e.g. 100,000 SNPs) assume strict neutrality (in order to be used as a null model)

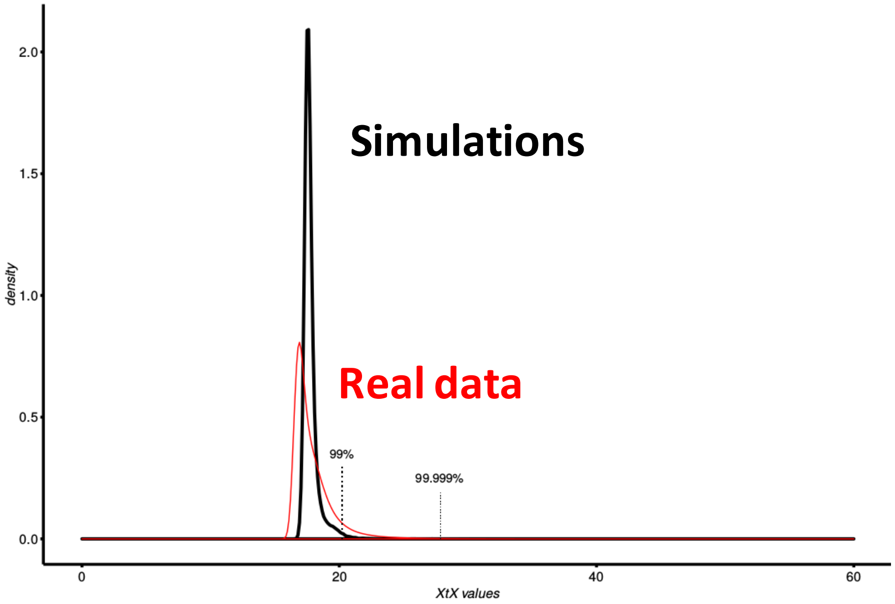
Same analysis under BayPass. But here, we know that all SNPs are neutral, we can therefore compute quantiles values for these neutral SNPs, allowing to have a neutral expectation.

Neutral calibrations – XtX metrics

Generate "Pseudo-Observed Datasets" (PODs) assuming the parameters used by the core model (in particular the omega matrix).

Of course, all simulated SNPs (e.g. 100,000 SNPs) assume strict neutrality (in order to be used as a null model)

Same analysis under BayPass. But here, we know that all SNPs are neutral, we can therefore compute quantiles values for these neutral SNPs, allowing to have a neutral expectation.

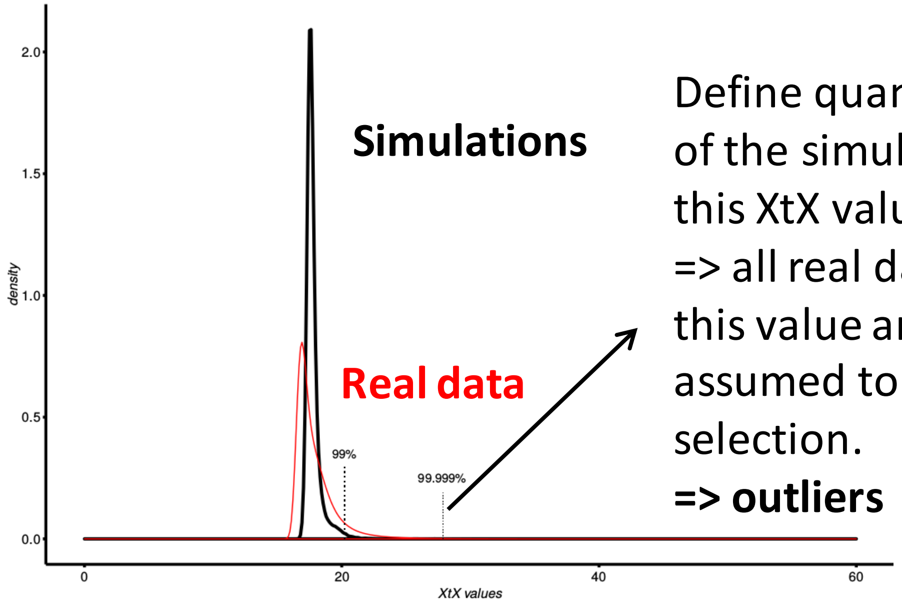


Neutral calibrations – XtX metrics

Generate “Pseudo-Observed Datasets” (PODs) assuming the parameters used by the core model (in particular the omega matrix).

Of course, all simulated SNPs (e.g. 100,000 SNPs) assume strict neutrality (in order to be used as a null model)

Same analysis under BayPass. But here, we know that all SNPs are neutral, we can therefore compute quantiles values for these neutral SNPs, allowing to have a neutral expectation.

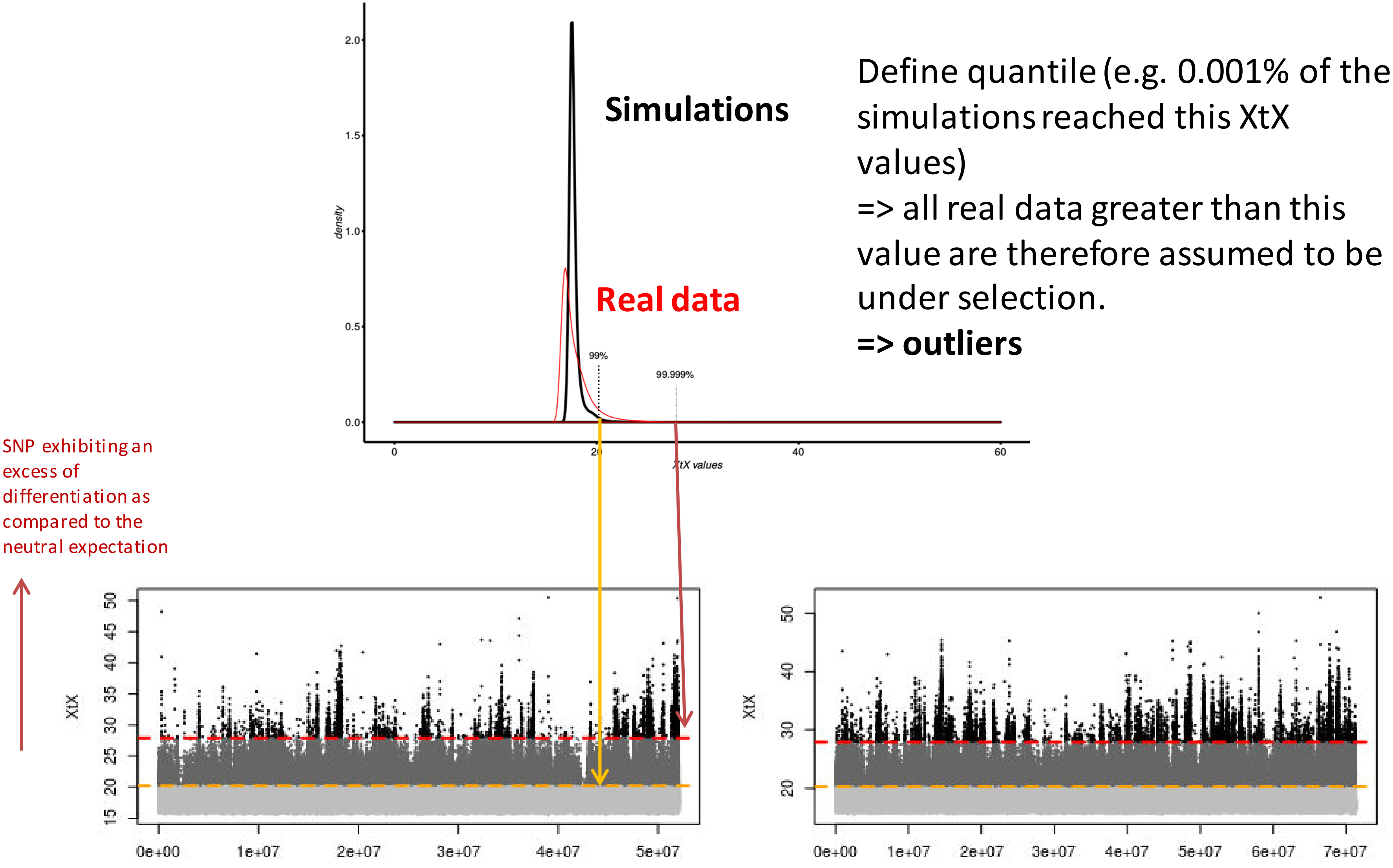


Simulations

Real data

Define quantile (e.g. 0.001% of the simulations reached this XtX values)
=> all real data greater than this value are therefore assumed to be under selection.
=> outliers

Use neutral calibrations for the genome scan



Simulations

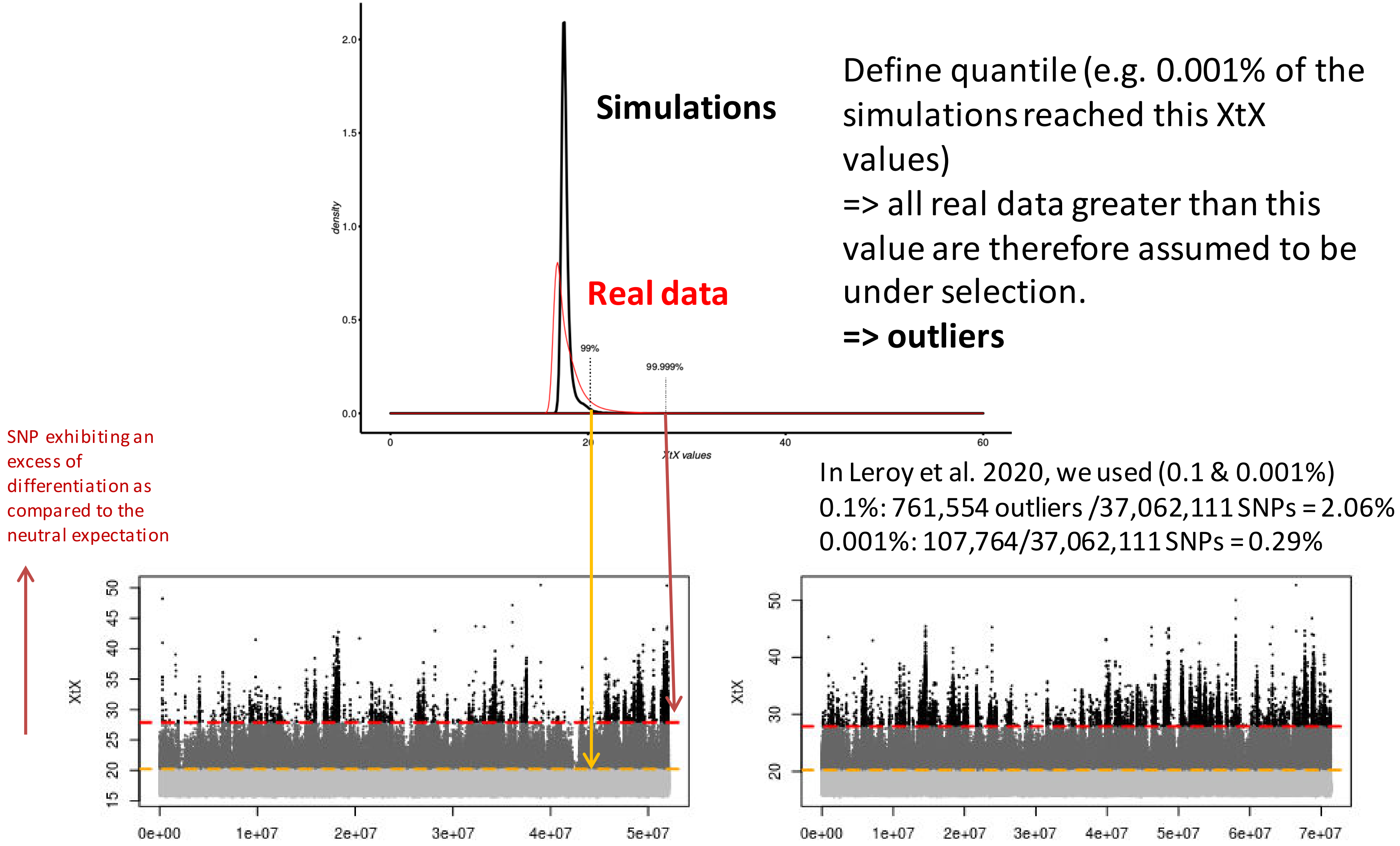
Real data

SNP exhibiting an excess of differentiation as compared to the neutral expectation

Define quantile (e.g. 0.001% of the simulations reached this XtX values)
=> all real data greater than this value are therefore assumed to be under selection.
=> outliers

Leroy et al. 2020b New Phytologist; Leroy & Rougemont, in press

Use neutral calibrations for the genome scan



Define quantile (e.g. 0.001% of the simulations reached this XtX values)
=> all real data greater than this value are therefore assumed to be under selection.
=> outliers

SNP exhibiting an excess of differentiation as compared to the neutral expectation

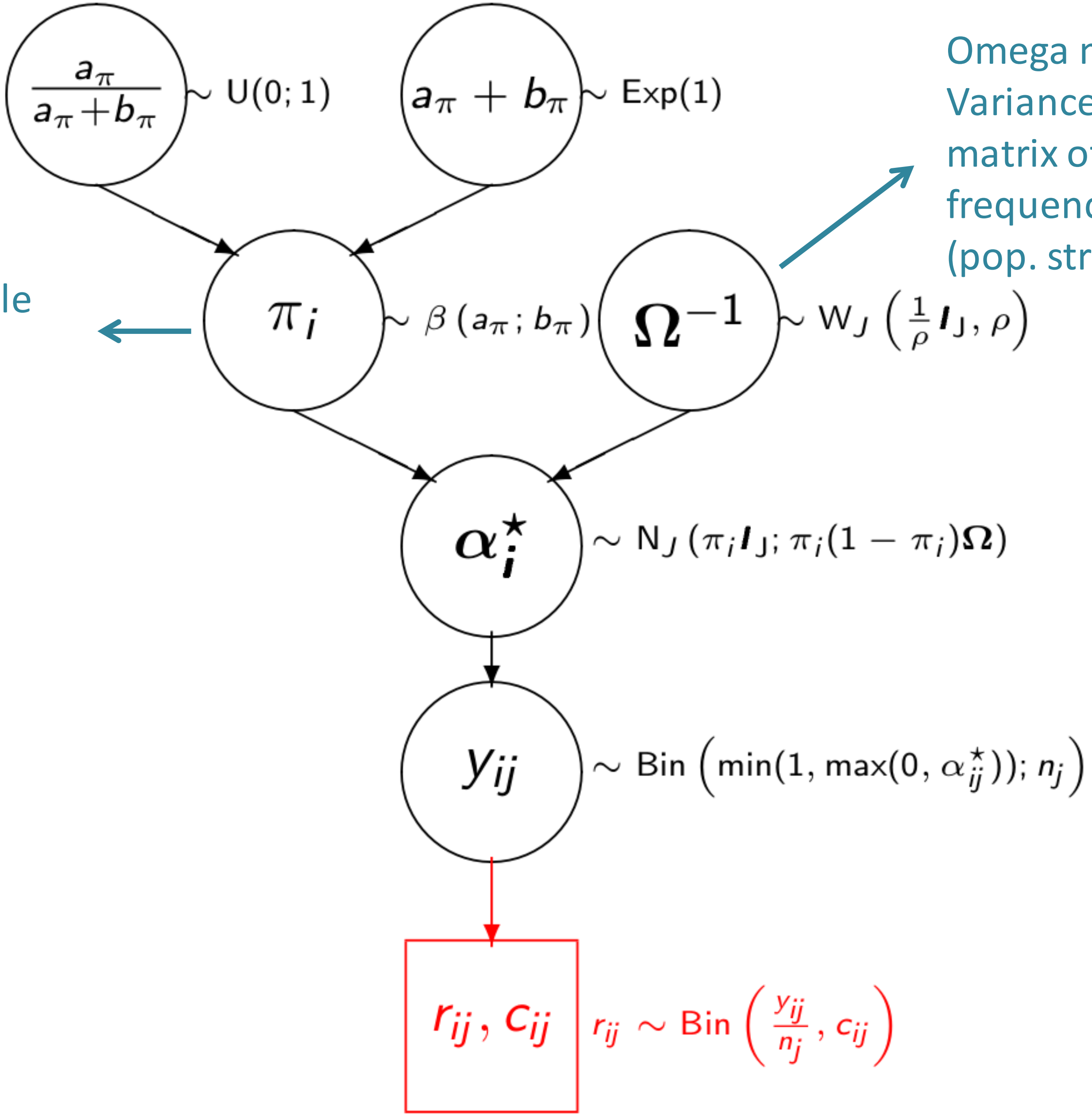
In Leroy et al. 2020, we used (0.1 & 0.001%)
0.1%: 761,554 outliers /37,062,111 SNPs = 2.06%
0.001%: 107,764/37,062,111 SNPs = 0.29%

XtX outliers are not randomly distributed along the genome, but rather cluster in several genomic regions (=> interesting)

Leroy et al. 2020b New Phytologist; Leroy & Rougemont, in press

The general strategy developped in BayPass (Mathieu Gautier, 2015) ~ Bayenv2 (Coop)

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

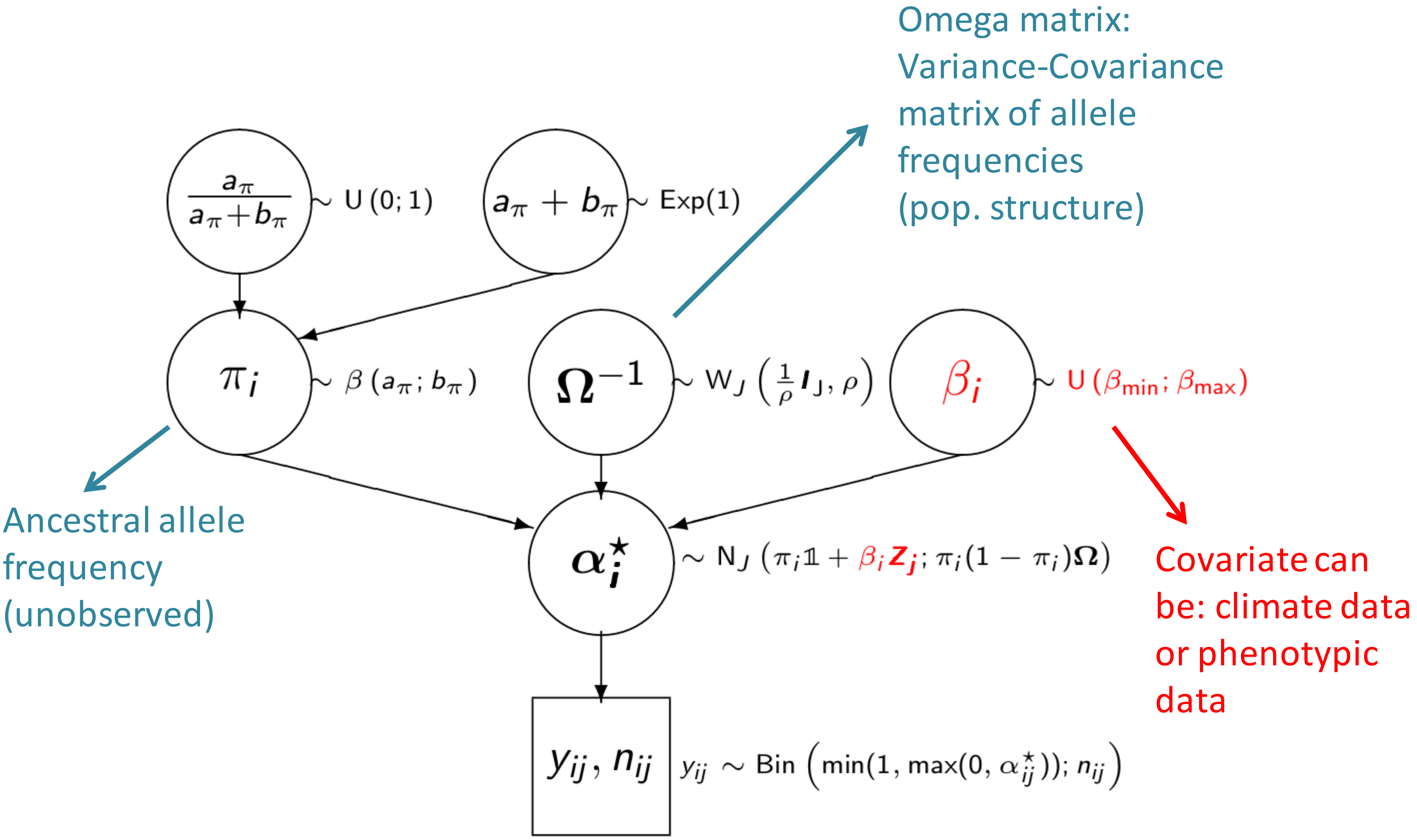

$\frac{a_\pi}{a_\pi + b_\pi} \sim U(0;1)$
$a_\pi + b_\pi \sim Exp(1)$
$\pi_i \sim \beta(a_\pi; b_\pi)$
$\Omega^{-1} \sim W_J\left(\frac{1}{\rho}\boldsymbol{I}_J, \rho\right)$
$\alpha_i^\star \sim N_J(\pi_i \boldsymbol{I}_J; \pi_i(1-\pi_i)\Omega)$
$y_{ij} \sim Bin\left(\min(1, \max(0, \alpha_{ij}^\star)); n_j\right)$
r_{ij}, c_{ij}
$r_{ij} \sim Bin\left(\frac{y_{ij}}{n_j}, c_{ij}\right)$

Omega matrix: Variance-Covariance matrix of allele frequencies (pop. structure)

Ancestral allele frequency (unobserved)

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

Genotype-environment association: Model with a covariate


$\frac{a_\pi}{a_\pi+b_\pi} \sim \mathrm{U}(0;1)$
$a_\pi + b_\pi \sim \mathrm{Exp}(1)$
$\pi_i \sim \beta(a_\pi; b_\pi)$
$\boldsymbol{\Omega}^{-1} \sim \mathrm{W}_J\left(\frac{1}{\rho}\boldsymbol{I}_J, \rho\right)$
$\beta_i \sim \mathrm{U}(\beta_{\min}; \beta_{\max})$
$\boldsymbol{\alpha}_i^\star \sim \mathrm{N}_J\left(\pi_i \mathbb{1} + \beta_i \boldsymbol{Z}_j; \pi_i(1-\pi_i)\boldsymbol{\Omega}\right)$
y_{ij}, n_{ij}
$y_{ij} \sim \mathrm{Bin}\left(\min(1, \max(0, \alpha_{ij}^\star)); n_{ij}\right)$

Omega matrix: Variance-Covariance matrix of allele frequencies (pop. structure)

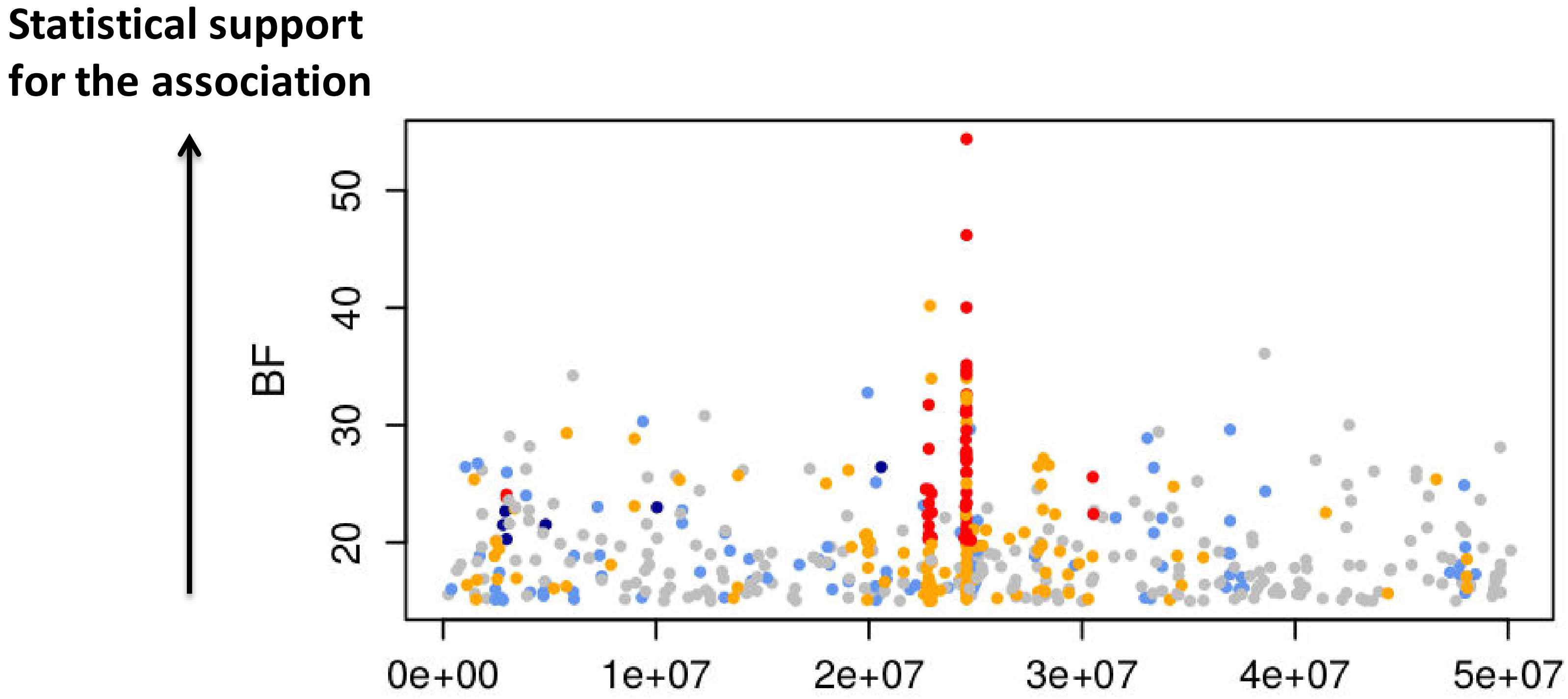
Ancestral allele frequency (unobserved)

Covariate can be: climate data or phenotypic data

Comparison of a model with no gradient (β=0, i.e. previous model) and a model with an association of the allele frequencies along the environmental gradient (β≠0).

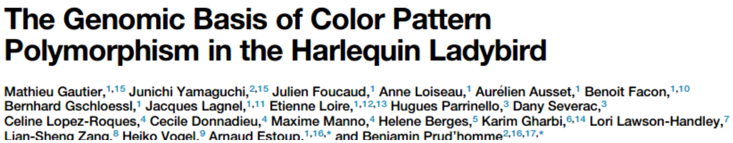
Bayes Factor captures the support for the association (higher =more supported)

Genotype-environment association:



e.g. SNPs associated with temperature & rainfall

- BF > 15 are considered as strong evidences, BF > 20 are considered as decisive evidences
(- It is also possible to calibrate a neutral expectation using quantiles estimed from neutral simulations (PODS), in a similar way than done for the XtX metrics)

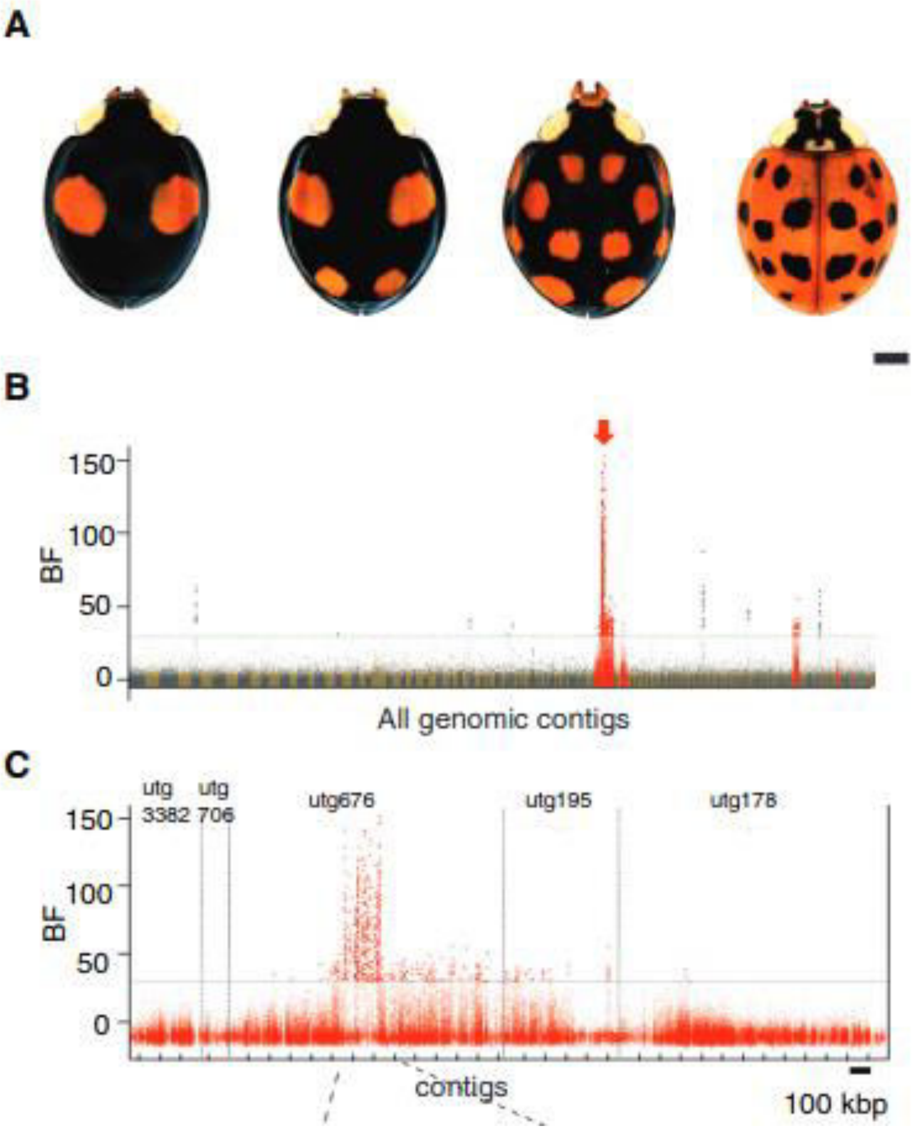


The Genomic Basis of Color Pattern Polymorphism in the Harlequin Ladybird

Mathieu Gautier,[1,15] Junichi Yamaguchi,[2,15] Julien Foucaud,[1] Anne Loiseau,[1] Aurélien Ausset,[1] Benoit Facon,[1,10] Bernhard Gschloessl,[1] Jacques Lagnel,[1,11] Etienne Loire,[1,12,13] Hugues Parrinello,[3] Dany Severac,[3] Celine Lopez-Roques,[4] Cecile Donnadieu,[4] Maxime Manno,[4] Helene Berges,[5] Karim Gharbi,[6,14] Lori Lawson-Handley,[7] Lian-Sheng Zang,[8] Heiko Vogel,[9] Arnaud Estoup,[1,16,*] and Benjamin Prud'homme[2,16,17,*]

Correlations with phenotype data (« population GWAS »)

(association for the proportion of Red-nSpots individuals)



Summary:

- F_{ST} computed over all SNPs are informative about the levels of population structure (pairwise F_{ST} matrix)

- Empirical distribution of F_{ST} among all genotyped SNP are highly informative, but remains descriptive, because it is difficult to use this distribution to properly define the proportion of loci under selection

- Observed variance of F_{ST} is due to the demographic history of a population, so a strategy can be to use a statistic like F_{ST} but expliclty accounting for the population structure (-> XtX)

- Based on the observed levels of the structure, it is therefore possible to perform neutral simulations to generate an expectation for the distribution of this statistic

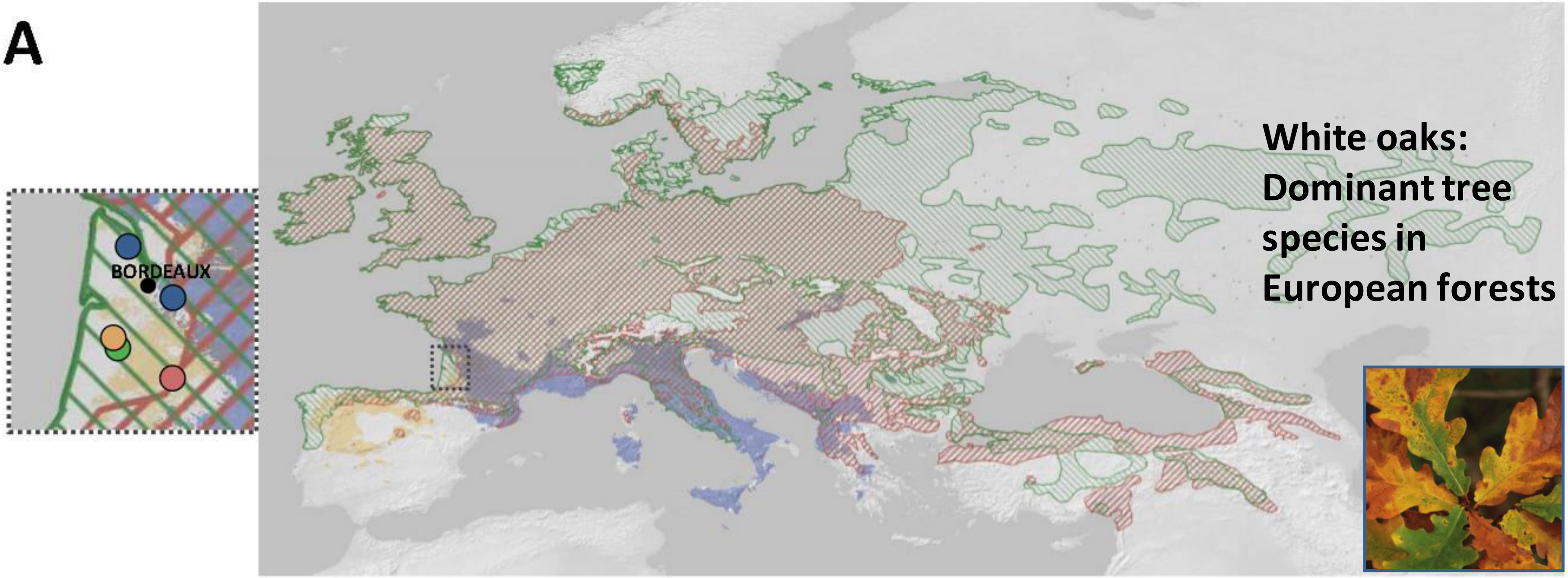
- (Slightly) more complex models with covariables allow to identify SNPs with allele frequency changes along this covariate (*i.e.* cline of allele frequencies)

- These covariables can be climate data (e.g. temperature, precipitations, altitude, latitude, …), phenotypic data, … -> Genotype-Environment associations and ‘population Genome-Wide Association Study’ (pGWAS)

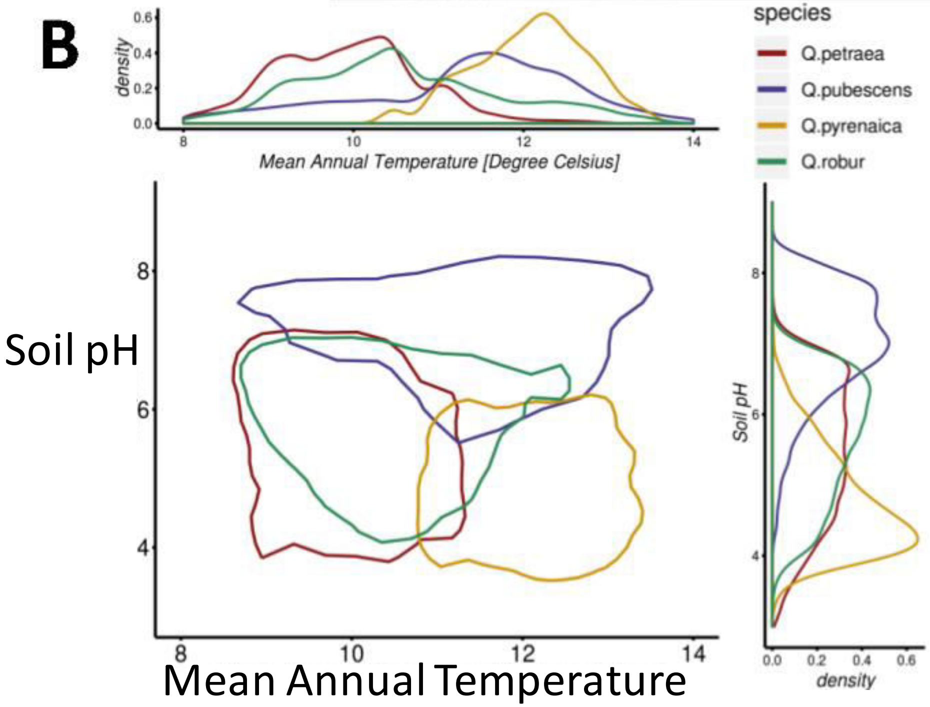
Practical:

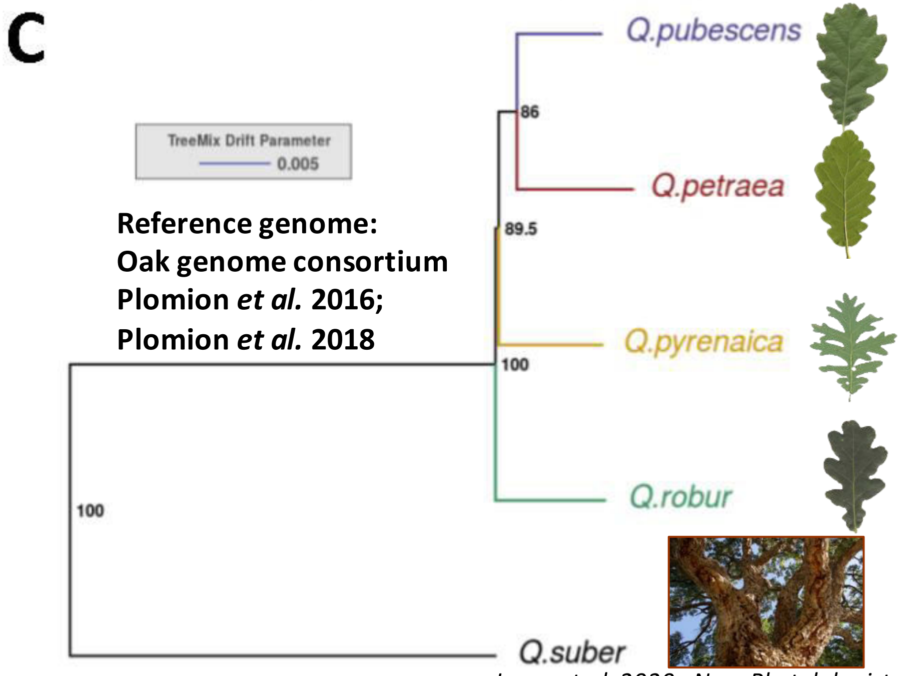
Local adaptation to climate in sessile oak populations

European white oaks



White oaks: Dominant tree species in European forests





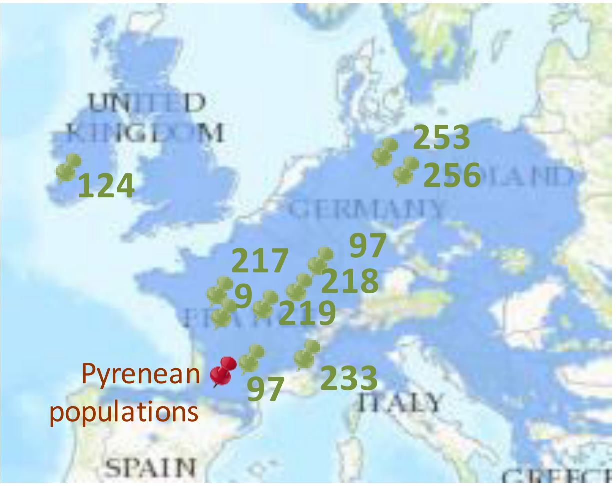
Reference genome:
Oak genome consortium
Plomion *et al.* 2016;
Plomion *et al.* 2018

Leroy et al. 2020a New Phytolologist

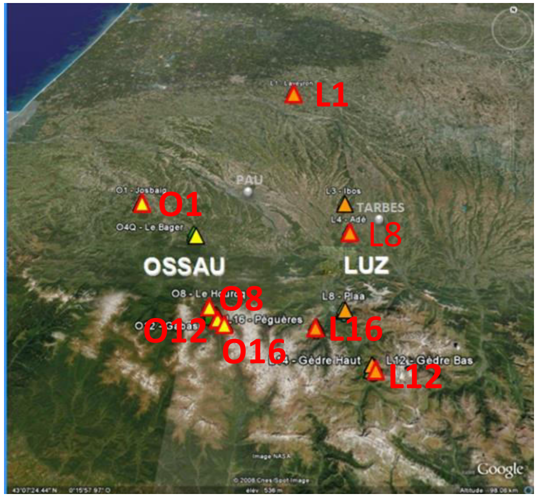
Local adaptation in sessile oak populations

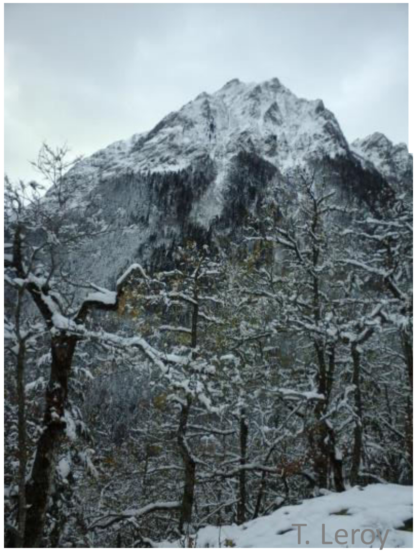
- Genomic data (Pool-seq):

10 populations at low elevation (25 ind/pool)



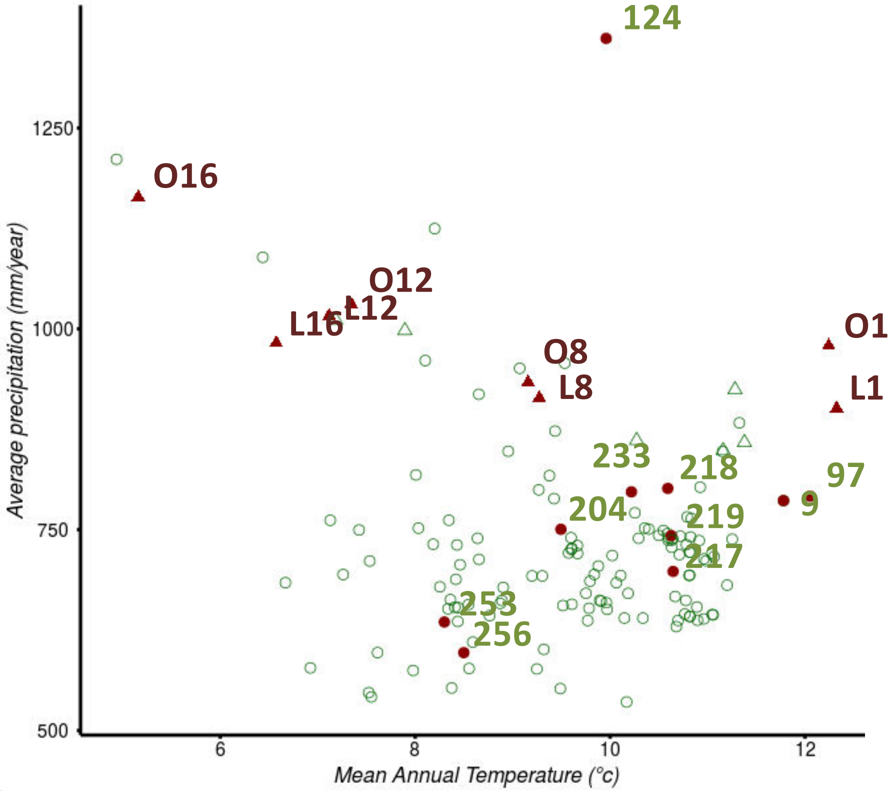
8 Pyrenean populations from low to quite high elevation (up to 1630m; 10-20 ind/pool)





- Climate data (1950-2000):

Mean annual temperature & precipitation sums



- Phenotypic data:

leaf unfolding in common gardens

Leroy et al. 2020b New Phytologist

Local adaptation in sessile oak populations

- Climate data (1950-2000):

Mean annual temperature & precipitation sums

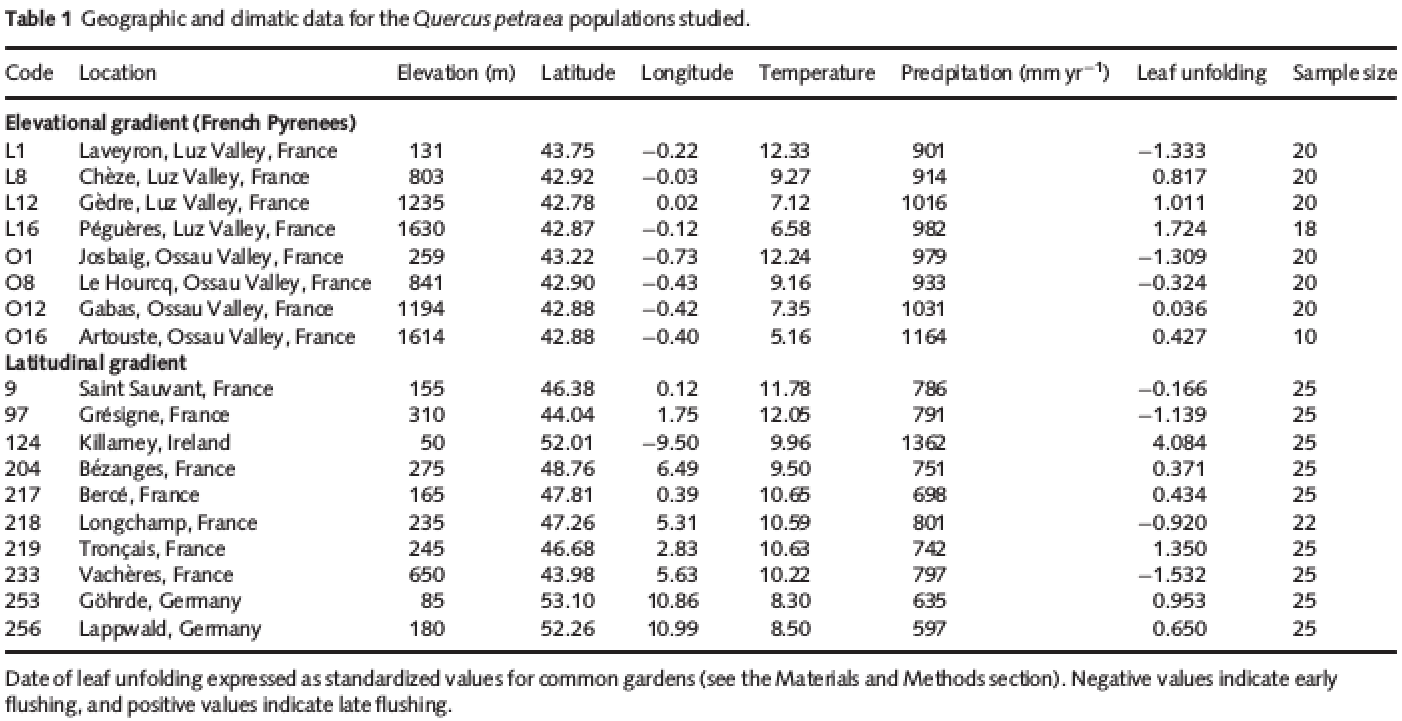


Table 1 Geographic and climatic data for the *Quercus petraea* populations studied.

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

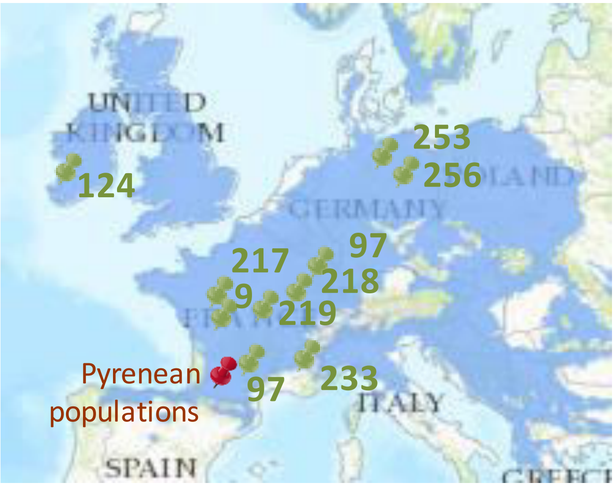
Date of leaf unfolding expressed as standardized values for common gardens (see the Materials and Methods section). Negative values indicate early flushing, and positive values indicate late flushing.

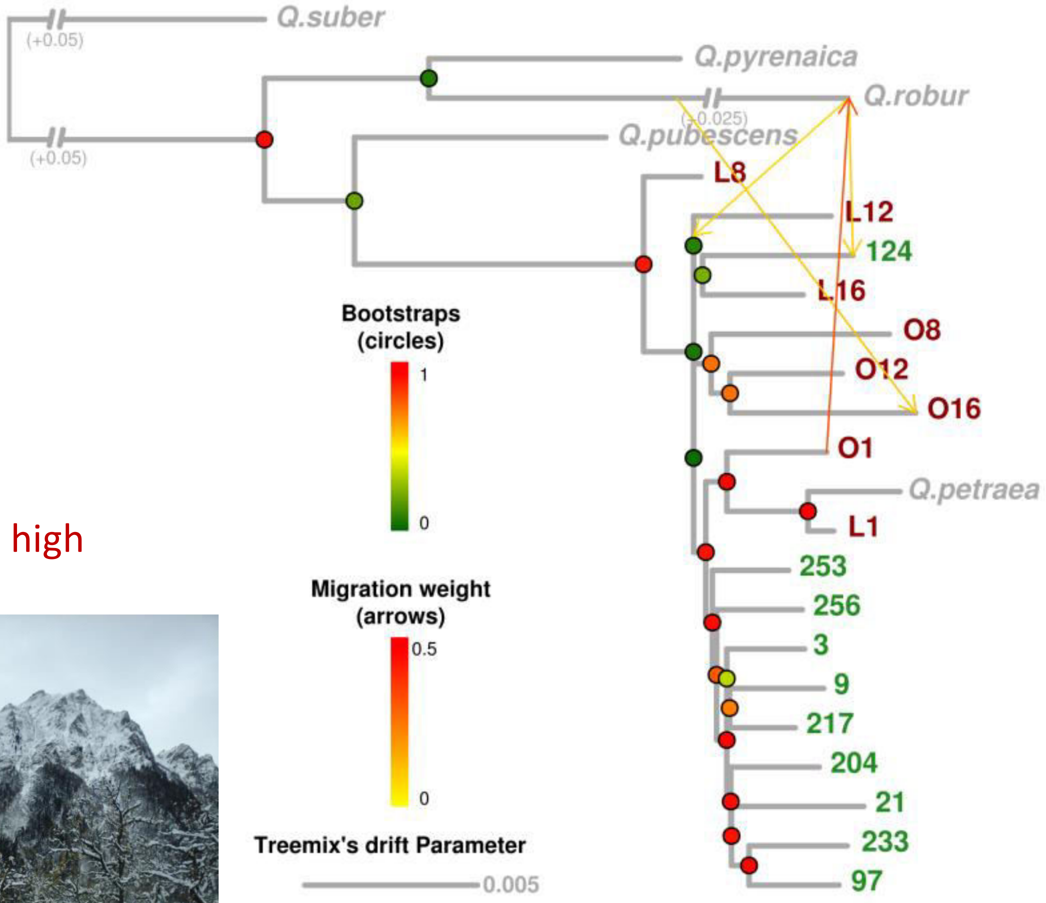
Leroy et al. 2020b New Phytologist

Local adaptation in sessile oak populations

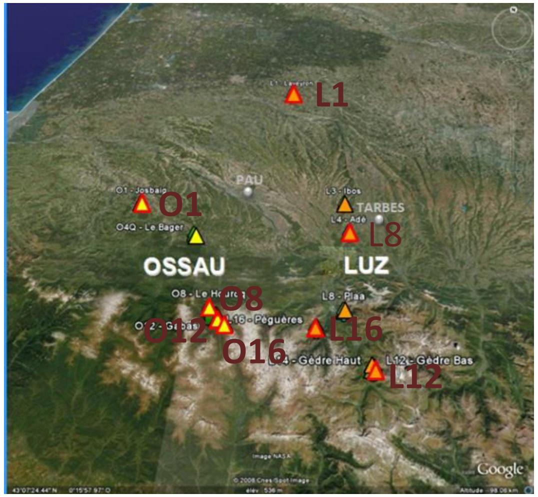
- Genomic data (Pool-seq):

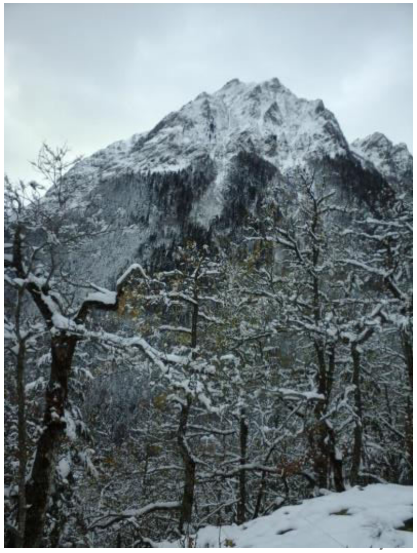
10 populations at low elevation (25 ind/pool)





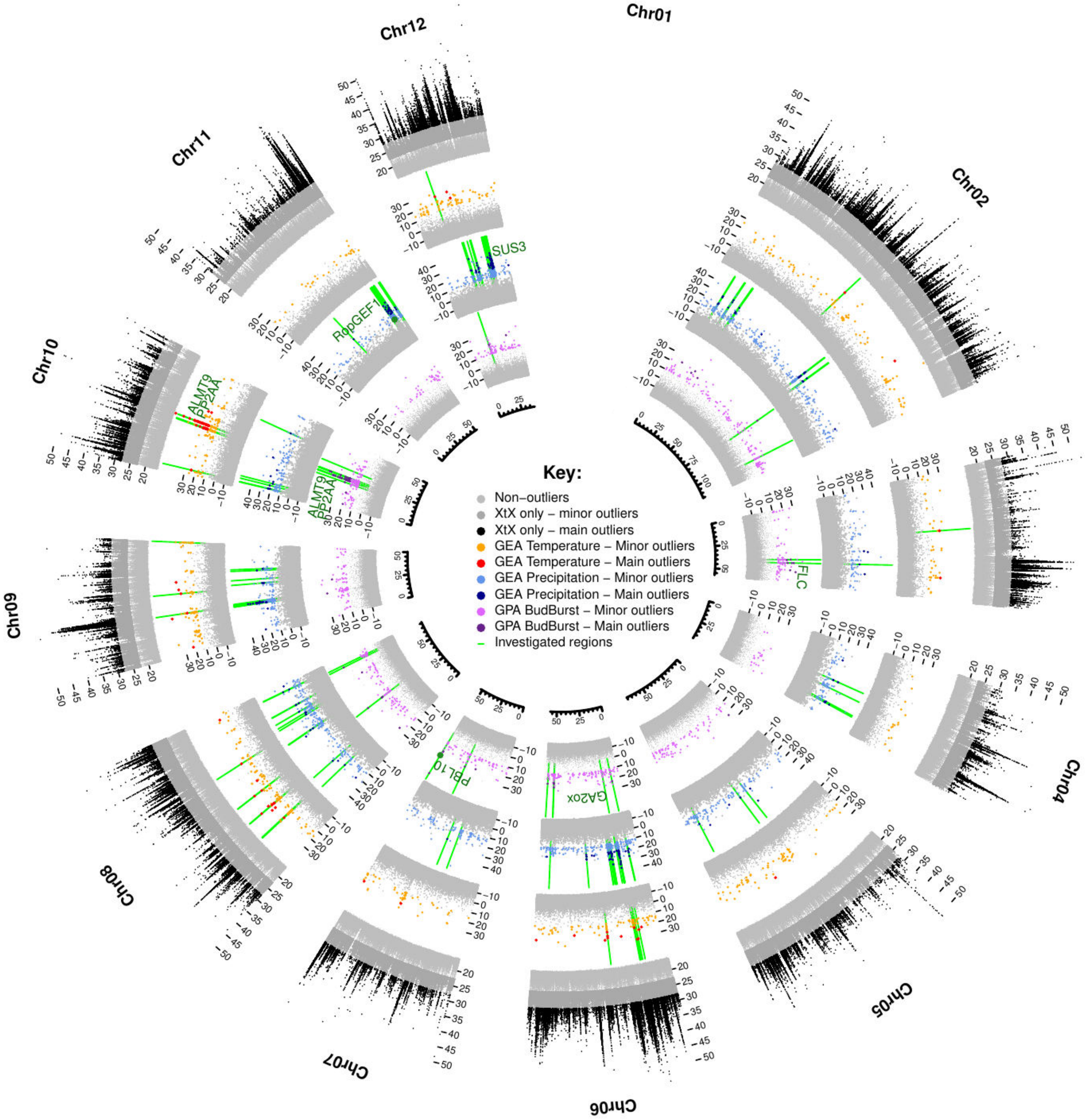
8 Pyrenean populations from low to quite high elevation (up to 1630m; 10-20 ind/pool)



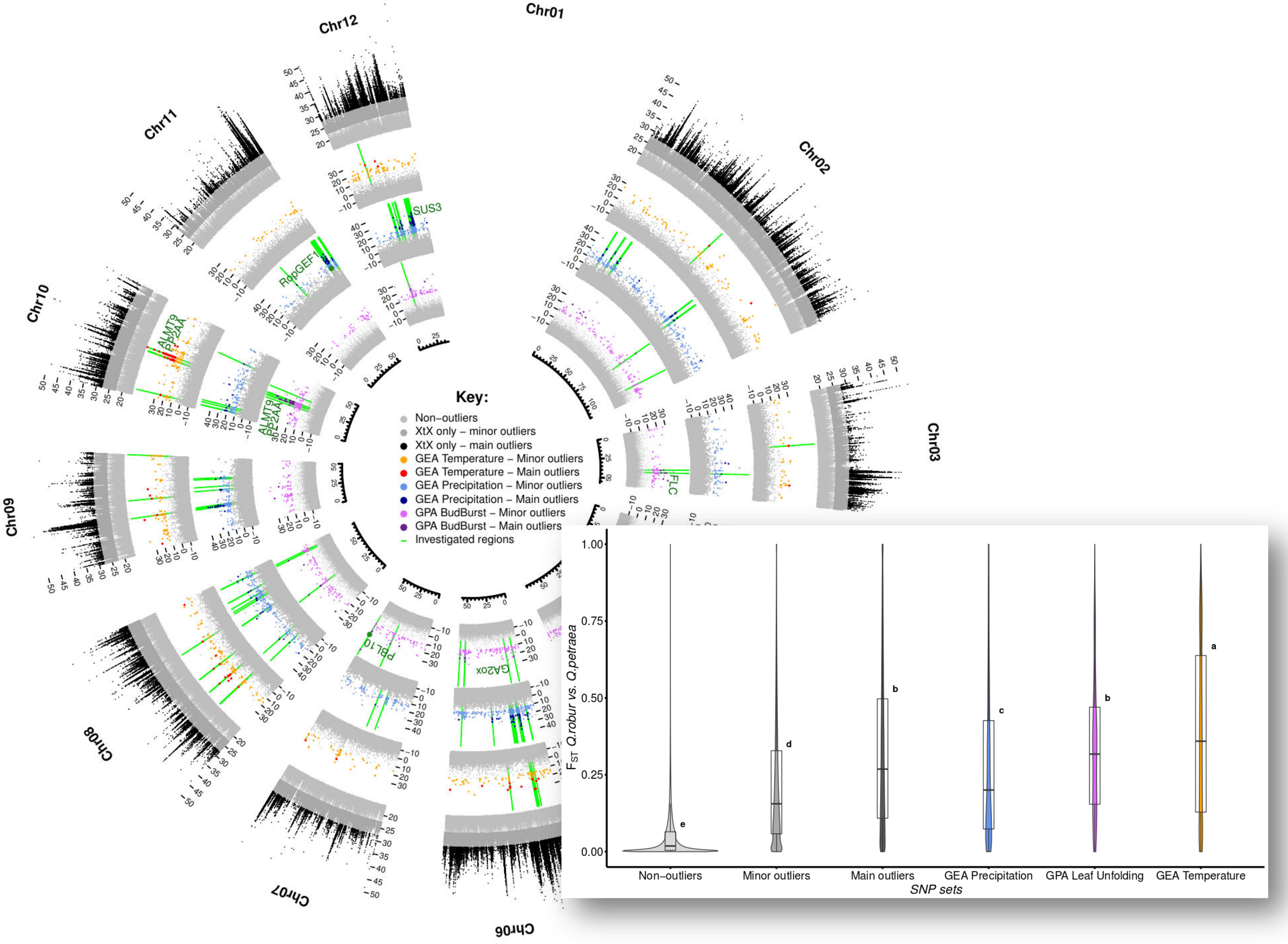


Leroy et al. 2020b New Phytologist

Local adaptation in sessile oak populations



Local adaptation in sessile oak populations

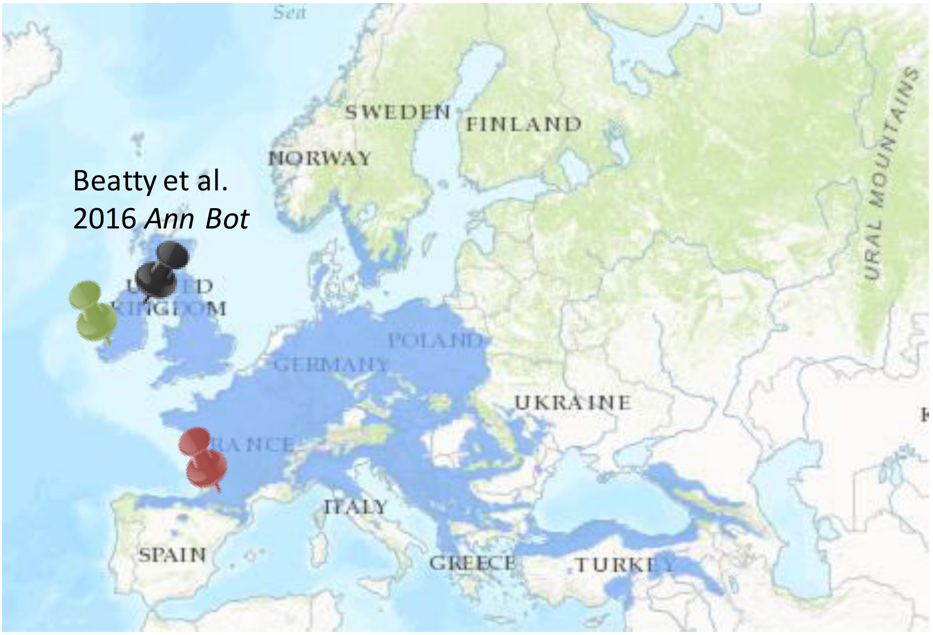


Stability/Cohesiveness

Local adaptation in sessile oak populations

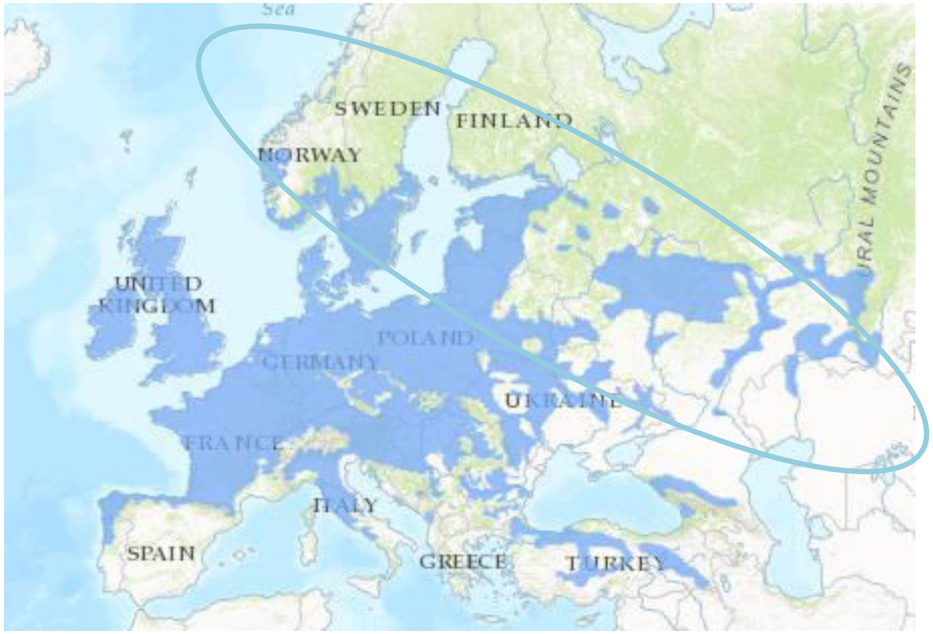
Adaptive introgression from *Q.robur* to *Q. petraea* in cold marginal habitats
(northern range, high elevation)

Quercus petraea



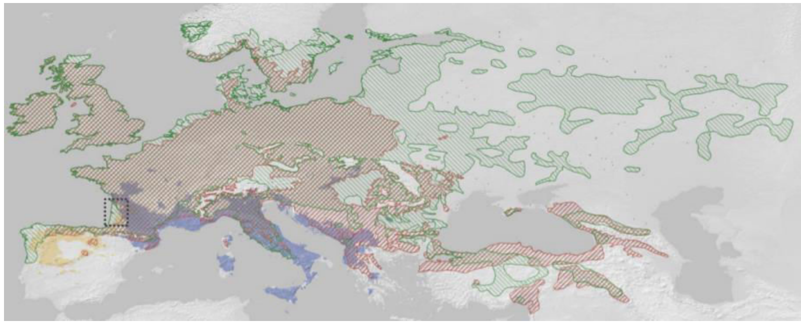
Beatty et al. 2016 *Ann Bot*

Quercus robur



Euforgen

Adaptive introgression from *Q. pubescens* and *Q. pyrenaica* in the south of their range?



Dataset to be used today

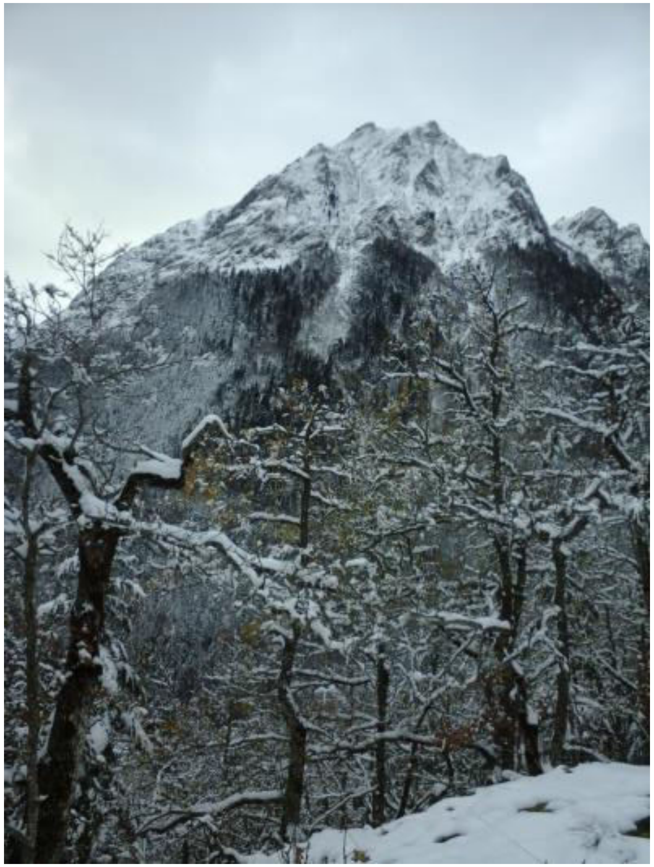
Pool-seq data of 18 populations (all from Q. petraea, the sessile oak).

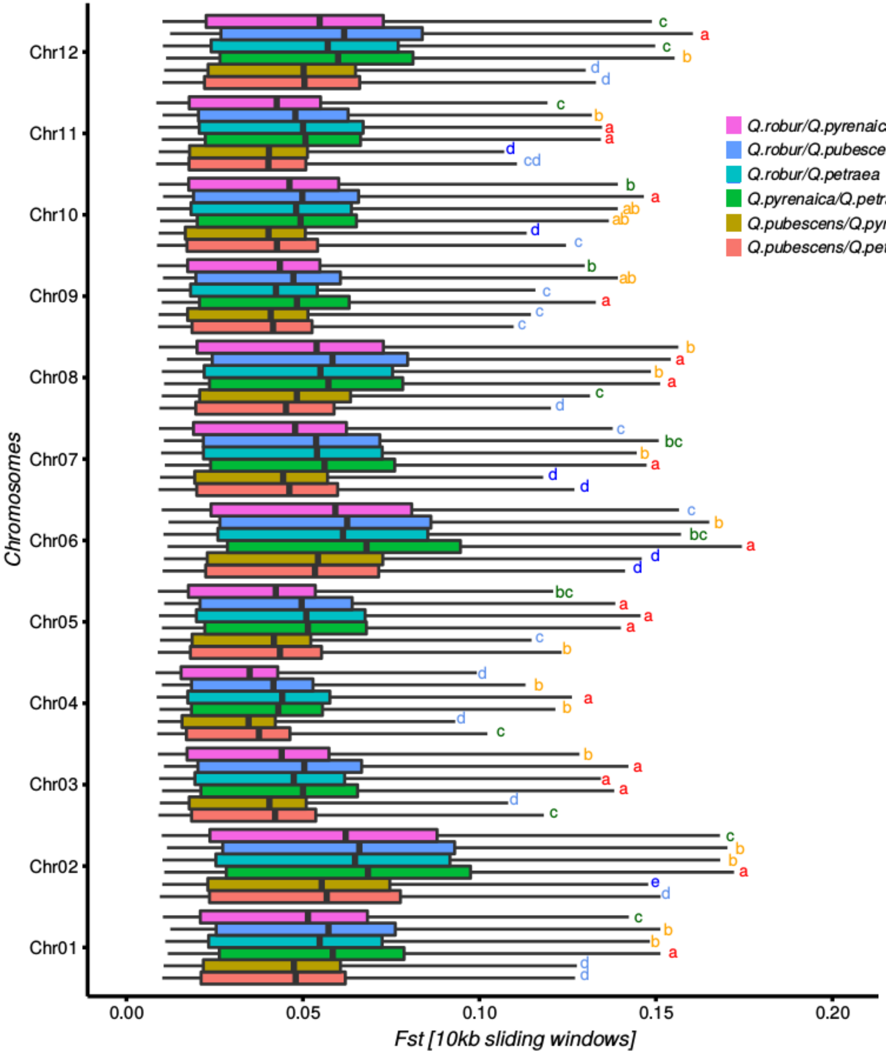
You will directly use observed allele counts (using two different files, one corresponding to a random sampling of 200,000 positions along the genome, and another one corresponding to a special focus on the 30 first Mb of the chromosome 1).

Now you just need to try to do the practical, and to get some information from your analyses.

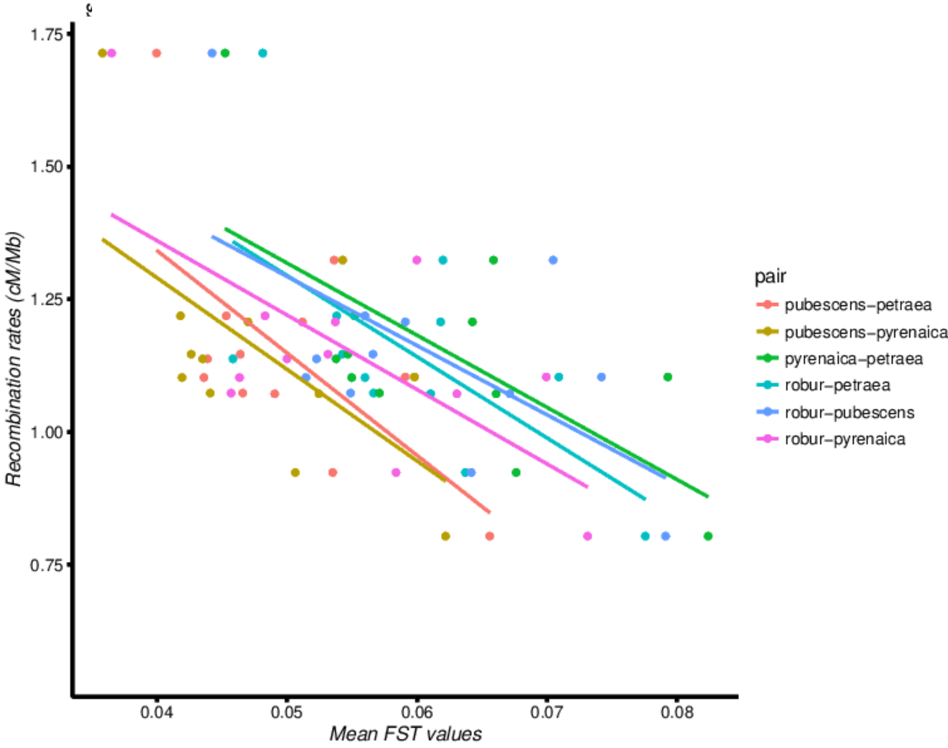
Remember to save your work frequently (Rcode, plots, …)!

Lost? Stuck? You can send me an email throughout the day! (thibault.leroy@univie.ac.at)





Some other source of variation (interchromosomal difference in recombination rates, effective population sizes variation…), but generally not taken into account (it is better to perform a specific analysis for sex chromosomes, independent from autosomes)



Leroy et al. 2020a New Phytologist