Chapter 13

Introduction to population genomics methods

Thibault Leroy & Quentin Rougemont

Abstract

High-throughput sequencing technologies have provided an unprecedented opportunity to study the different evolutionary forces that have shaped present-day patterns of genetic diversity, with important implications for many directions in plant biology research. To manage such massive quantities of sequencing data, biologists however need new additional skills in informatics and statistics. In this chapter, our objective is to introduce population genomics methods to beginners following a learning-by-doing strategy in order to help the reader to analyze the sequencing data by themselves. Conducted analyses cover several main area of evolutionary biology, such as an initial description of the evolutionary history of a given species or the identification of genes targeted by natural or artificial selection. In addition to the practical advices, we performed re-analyses of two cases studies with different kind of data: a domesticated cereal (African rice) and a non-domesticated tree species (sessile oak). All the code needed to replicate this work is publicly available on github (https://github.com/ThibaultLeroyFr/Intro2PopGenomics/).

Keywords:
Whole-genome sequencing, Pool-seq, Nucleotide diversity, Molecular evolution, genome scans, population structure, individual, artificial and natural selection, bioinformatics, perseverance

1. Introduction

Population genetics is an increasingly important discipline at the interface between genetics and evolutionary biology focusing on the analysis of DNA variation and evolution across different loci and populations. Population genetic concepts help

to understand the contribution of key evolutionary forces (mutation, migration, genetic drift and natural selection) to the observable present-day distribution of genetic diversity. Prior to describe how various important and long-standing questions in plant biology can be addressed using population genetic concepts (for plant breeding, plant conservation biology, plant ecology for example), it is important to notice that a major shift occurred in this discipline. Over the last decade, cost-effective and high-throughput sequencing methods have accelerated and amplified the interest for population genetics by taking advantage of large-scale comparisons of DNA sequences or large sets of Single Nucleotide Polymorphisms (SNPs) to better understand the contribution of the different evolutionary forces to the present-day DNA variation, leading to the emergence of a closely-related field, called **population genomics** ([1] for a historical retrospective).

Biologists have now access to very large amounts of sequencing data. This change makes new investigations possible, but also induces a considerable shift in the professional skills needed to generate (wet lab) or analyse the data (dry lab). Indeed, large-scale sequencing projects with several hundred or thousands of samples sequenced have considerably shifted the limits in plant research (e.g. 3,000 Rice Genome project, [2]; *Arabidopsis thaliana* 1001 Genomes Project, [3]). These new investigations require additional skills in biology, especially regarding the bioinformatic analysis of the sequencing data (e.g. a strong experience in using command-line versions and high-performance computing clusters, a proficiency in scripting or programming, a solid competence in statistical methods) to be able to handle such big genomic data projects. This greater transdisciplinarity between genetics, informatics and statistics, can make access to population genetics more difficult. In this chapter, our main objective is to tackle this issue by providing a simple and step-by-step guide. Unlike many great academic writings in the field (e.g. [4]), this chapter is not interested at covering the basis of the theory of evolution, but rather at **introducing population genomics methods to beginners** following a "learning-by-doing" strategy. All the genomic data we used are publicly available, as well as our scripts (see Materials below).

Population genetics is a broad discipline and we do not claim to be exhaustive. Our objective is rather to introduce population genomics by focusing on some key analyses: the analysis of population structure, the inference of population splits and exchanges, and the detection of footprints of natural or artificial selection. We hope that some plant biologists, including students, will discover the benefits of population genomics analyses, including its applications for breeding and conservation, despite the fact that this discipline is, rightly or wrongly, reputed to be particularly difficult and demanding.

2. Materials

This tutorial requires the use of command-line software (preferentially on high-performance computing clusters) and some basic knowledge about Linux and bash commands (*e.g.* cd, mkdir, cp, paste, awk, grep). There are plenty of good tutorials available on Internet to learn these aspects in a couple of hours, such as the Ryan Chadwick's website (https://ryanstutorials.net).

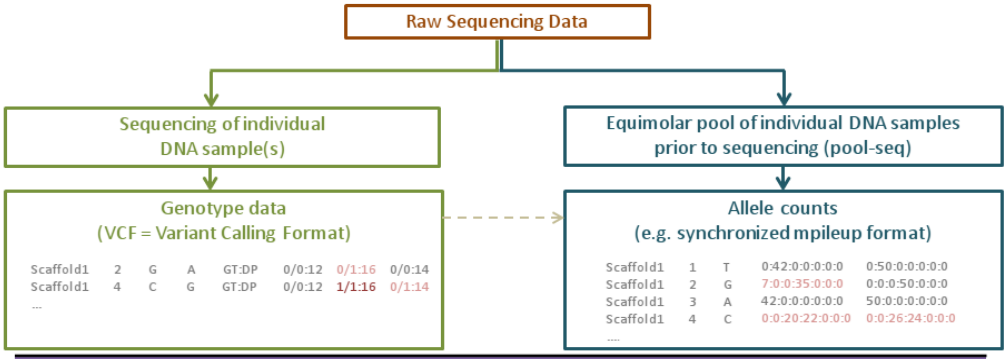
Due to space constrains, the code and commands are not described in this book chapter. However, all our scripts (bash, python and R) are freely available on github: https://github.com/ThibaultLeroyFr/Intro2PopGenomics/

This code repository is therefore an essential and complementary part of this chapter.

These scripts require different softwares:

1. BayPass: http://www1.montpellier.inra.fr/CBGP/software/baypass/download.html
2. BWA mem: http://bio-bwa.sourceforge.net/
3. Blast+: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download
4. Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml
5. FastStructure: https://rajanil.github.io/fastStructure/
6. GATK: https://software.broadinstitute.org/gatk/download/
7. Plink: https://www.cog-genomics.org/plink2/
8. Picard: https://broadinstitute.github.io/picard/
9. R https://cran.r-project.org/
   (Rstudio is not mandatory but can be useful: https://www.rstudio.com/products/rstudio/download/)
   including R packages:
   - ape: https://cran.r-project.org/web/packages/ape/index.html
   - circlize: https://cran.r-project.org/web/packages/circlize/index.html
   - ggplot2: https://cran.r-project.org/web/packages/ggplot2/index.html
   - pcadapt: https://cran.r-project.org/web/packages/pcadapt/index.html
   - poolfstat: https://cran.r-project.org/web/packages/poolfstat/index.html
   - reshape2: https://cran.r-project.org/web/packages/reshape2/index.html
10. SNPRelate: https://bioconductor.org/packages/release/bioc/html/SNPRelate.html/
11. SAMtools: http://samtools.sourceforge.net/

12. Seq_stat to compute nucleotide diversity and Tajima's D: https://tinyurl.com/yxurjgdx
13. TreeMix: https://bitbucket.org/nygcresearch/treemix/downloads/
14. Trimmomatic: https://github.com/timflutre/trimmomatic
15. VCFtools: http://vcftools.sourceforge.net/
16. wget: https://www.gnu.org/software/wget/



MAIN GOALS	TASKS	GENOTYPES (INDIVIDUAL-BASED)	ALLELE FREQUENCIES (POOL-SEQ BASED)
Reconstruct the evolutionary history of a given species	Population structure	PCA, fastSTRUCTURE	Covariance Matrix
	History of Gene Flow	TreeMix dadi/ABC	TreeMix dadi/ABC
	Historical changes in population sizes	PSMC/MSMC/SMC+ dadi/ABC	dadi/ABC
Detect key genes involved in: - adaptation - domestication - speciation	Detection of local adaptation	PCAdapt	BayPass
	Selective sweeps	Tajima's D/SweeD/RAiSD	Tajima's D (Popoolation suite)
	Genotype-Phenotype Association (GPA)	LFMM2/BAYPASS GWAS	BayPass

Fig. 1*: Data format and analyses using individual vs. pooled samples (i.e. DNA of several individuals mixed prior to sequencing, hereafter pool-seq). All analyses can be performed with individual data (dotted arrow), but the pool-seq data have limitations (see also #3.3.1 for the advantages and disadvantages of pool-seq). Methods or programs shown in green are those used in the following sections.*

3. Methods

After introducing notions related to the handling of large sequencing data, we will provide guidelines to perform population genomic analyses based on two publicly available data from two different species: African rices from [5] and sessile oaks from [6]. These two examples were selected to cover broad plant biology related issues, with both crop- and wild flora-associated topics. In addition, these two studies used different kind of sequencing data: individual-based genotypes vs. pooled DNA samples (a mixture of the DNA from several individuals prior to sequencing, hereafter

pool-seq). As shown in **Fig. 1**, all analyses described in the analyses of the pool-seq data are based on the allele frequencies can also be performed for individual-based data, at least when a minimum of 12-15 individuals were sequenced per population. In other words, analyses based on pool-seq data are far more limited than individual-based sequencing data, but pool-seq represents a cheaper strategy than the sequencing of individuals (see #3.3.1). Our analyses focus on plant species but it has to be noted that such analyses can also be used to analyze various non-plant datasets, at least for diploid eukaryotic species.

3.1 From raw DNA data to genetic variants

1. Reads: All genomic projects start from the sequencing of very small pieces of DNA generated by a DNA sequencer, called **reads**. Despite recent advances in sequencing technologies (hereafter NGS, "for Next-generation sequencing") to generate long fragments (up to 100,000 bases or more, e.g. Oxford Nanopore or PacBio technologies), these technologies remain, at the time of writing, too expensive to sequence multiple individuals of a given population in order to describe the genetic variation observed within this population. Such new technologies therefore remain little used in population genomics projects. Most population genomicists rather use huge quantities of very short - but affordable - sequencing reads (e.g. Illumina sequencing of both ends of a short DNA fragment, so called **paired-end reads**, generating 100-300 bases of known sequence for each ends).
2. FASTQ file structure: High-throughput sequencing instruments generally output sequences under a FASTQ format. A FASTQ file is a text file with *n* repeats of 4 lines, with *n* depending of the total number of generated reads. The first line begins with a "@" (equivalent of a ">" for a FASTA sequence) which indicates a new sequence. This line then contains a unique sequence identifier. The second line corresponds to the sequencing read itself, *i.e.* the succession of the different DNA bases read by the sequencer instrument. The third line generally only contains a "+" character. The fourth line corresponds to the quality values for the corresponding bases in second line, in the exact same order. In other words, the DNA sequencer provides a confidence score in the assignment of the corresponding base call. The very first step of a population genomic project is therefore to exclude low quality reads and bases from these raw FASTQ files, in order to eliminate the majority of sequencing errors, a process commonly referred as **read trimming**.
3. Read mapping to reference genome: All along this chapter, we assume that a reference genome is already available for the species you are interested in (or at least a closely related one). If not, the best solution is to start by generating a

high-quality de novo genome assembly (this step ideally requires to establish a close collaboration with an experienced bioinformatician). If so, trimmed reads are then "mapped" against a reference genome in order to find the most likely genomic location for a read sequence, a process hereafter referred as **read mapping**. A read mapper is not strictly speaking a read alignment software. The read mapper tries to find the best location(s) for a given read, but without establishing the base-to-base correspondence with the reference sequence. It might seem surprising, but can be explained by a complex time-sensitivity trade-off. Any increase in the sensitivity of the mapping heavily slows down the speed of execution. To remain computationally efficient, particularly with extremely high volumes of sequence data, the two most commonly used read mappers BWA [7-8] and Bowtie2 [9] identify the potential loci of origin of a sequencing read, but without performing precise local alignments. For short read data, these softwares remain fast and accurate methods, but it remains important to bear this limit in mind, especially in the future when reads will increase in length.

4. Variant calling: The identification of genetic variants from NGS data, hereafter **variant calling**, requires the accumulation of several reads at the same location, to increase the confidence in the identification of polymorphisms. Such methods generally predict the likelihood of variation at each locus to take into account some sequencing or mapping biases. Current population genomic studies are generally based on short polymorphisms, either SNP or short indels (insertions and deletions). Large structural variations (*e.g.* large indels, translocations, duplications) represent a non-negligible part of the genetic variation, but remain quite difficult to access with the commonly used short-read data. This specific genetic diversity is therefore not addressed in the following sections.

3.2/ Case study 1: individual-based genotyping

3.2.1 African rice

Plant domestication might appear at first sight to be a simple and abrupt transition from a wild ancestor to a domesticated species. Following this view, it is generally assumed that only a part of the phenotypic (and genetic) diversity of the ancestral species has been used by the early farmers and therefore has contributed to the newly domesticated one, generating a so-called **domestication bottleneck**. As a consequence, theoretical work predicts that domestication is associated with a reduction of the genetic variation and a higher **mutation load**, *i.e.* an increase in the number of deleterious alleles. This prediction is empirically supported in several plant or animal species [10]. For most domesticated species, domestication can be viewed as a long transitional process over millennia rather than a sudden event. This induces several other layers of complexity (reviewed in [11]), such as the possibility for (i) past and/or contemporary gene flow between wild and domesticated species, (ii)

several wild contributors, (iii) several centers of domestication, (iv) massive changes in census and **effective population sizes** (*Ne*) of either the wild, the domesticated or both species. All these situations are expected to have substantial impacts on the neutral diversity and can generate confounding patterns leading to inappropriate conclusions.

In this section, we decided to use huge NGS data from the domesticated African rice (*Oryza glaberrima*). This species is characterized by a small genome (<350 Mb) and a simple organization (diploid), at least for a plant species. In addition, Cubry et al. (2018) recently investigated the evolutionary history of this species through a large sequencing projects of 83 wild (*O. barthii*) and 163 domesticated individuals. This study represents an excellent and detailed piece of work. To speed up computations and help the reader to replicate this work, we have focused on a subset of 23 wild and 25 domesticated individuals from the centre of domestication (as identified by Cubry et al. [5], corresponding to present-day Mali, Ghana, Niger, Nigeria, Benin, and Togo).

3.2.2 Variant discovery from publicly available data

1. Databases: Before downloading publicly available sequence from the Sequence Read Archive (SRA) or the European Bioinformatics Institute (EMBL-EBI), a close reading of the webpage associated to the project can provide considerable useful information about the data. Both the SRA and the EMBL-EBI website give relatively similar information, but from our perspective, the EMBL-EBI website is more user friendly (**Fig. 2**). In the search bar, enter the ID of a project (e.g. ERP023549 for the African rice). To have an overview of the data, click on the associated project (for the African rice project: IRIGIN for International RIce Genome INitiative). The webpage containing a table with several fields by default: sample accession ID, the species name, some information relative to the sequencing instrument or the library protocol or different URL to download the data (**Fig. 2**). By selecting some additional columns, further information is available such as the number of reads or the sizes of the gzipped FASTQ files.

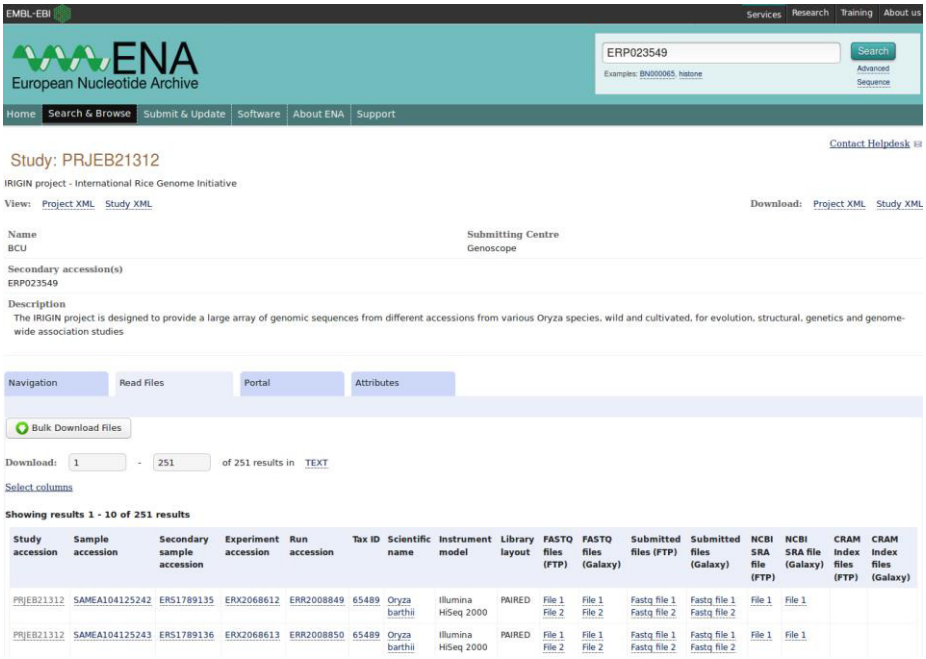


Fig. 2: *screenshot of the* EMBL-EBI *webpage for the African rice sequencing project described in Cubry et al. [5].*

2. Data downloading to SNP dataset: To download the data, the best solution is to use a shell File Transfer Protocol (FTP) client such as wget. For example the accession ERR2008855 can be downloaded from SRA servers using the following command in a terminal emulator:

wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR200/005/ERR2008855/ERR2008855_1.fastq.gz
wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR200/005/ERR2008855/ERR2008855_2.fastq.gz

And so on, for all individuals you want to download the read data.

All our scripts are available to download and replicate all steps (including trimming, read mapping, variant calling, see https://github.com/ThibaultLeroyFr/Intro2PopGenomics/tree/master/3.2.2/). In a nutshell, we use Trimmomatic to remove low qualities bases using a window computing average quality and sliding along the read, excluding all remaining bases of the read, if the average quality over 4 successive bases drops below 15. After excluding low quality bases, reads with less than 50 remaining nucleotides are discarded. Then, we map all the remaining reads using BWA, remove duplicates with Picard and perform the variant calling under GATK. We use the GATK haplotypeCaller to first generate individual VCF file (gVCF for genome Variant Call Format) and then perform the joint genotyping of the 48 individuals (joint VCF in **Fig. 1**). Low quality SNPs are excluded, generating a set of 6,150,642 filtered SNPs (i.e. with a “PASS” label in the final VCF file).

3.2.3 Population structure

Genetic differences between populations can be investigated by examining population structure - sometimes referred to as population stratification - which represents genome-wide differences in allele frequencies. Such a difference in genetic ancestry among individuals is possible because the samples can be derived from several populations that have experienced different demographic histories. As a consequence, all population genomics project first assess population structure in order to take it into account in the downstream analyses. Aside from biological reasons, analyses of population structure allow to identify errors such as the accidental misidentification of some individuals arising during sample preparation, sequencing or bioinformatics phases.

Given that this population structure represents a systematic shift in allele frequencies, a very large set of SNPs is unnecessary to investigate population structure patterns. A limited number of unlinked SNPs randomly selected across the entire genome (e.g. few thousands SNPs with a low proportion of missing data) is sufficient to get an accurate picture of the population structure. Such genome complexity reduction is also more computationally-efficient and reduces the number of variants in strong linkage disequilibrium (LD). LD represents a deviation from the hypothesis of random association of alleles within a genome and may impact the inferred population structure (*see* **Note 1**). Indeed, most popular population genetic tools use models assuming no or weak linkage disequilibrium within populations, including the most widely used model-based population genetics program STRUCTURE [12-14].

1. Principal Components Analysis (PCA): PCA is a commonly used exploratory analysis to infer population structure among individuals [15]. PCA helps to visualize genetic distance and relatedness between individuals by calculating principal components, with the top components explaining most of the differences among samples. In practice, PCA is sensitive to missing data. As a consequence, depending of the proportion of missing data in the VCF file (*i.e.* individuals with an unknown genotypes: "./."), population geneticists either exclude all SNPs with missing data or replace missing values by the mean of the values based on the individuals with known calls. As a general rule, it is better to investigate population structure with few highly-informative SNPs than using large proportion of poorly-genotyped SNPs. This warning is especially important for SNP set derived from Restriction site-Associated DNA data (RAD-seq data, [16]) which generally contain a large proportion of missing data.

   The African rice project is based on massive Illumina sequencing data, leading to a VCF contain very little missing data. As a consequence, we have chosen to remove all SNPs with missing data before performing PCA (*i.e. grep -v* "\./\."

[VCFfile]). An example of PCA based on the 48 African rice samples is shown in **Fig. 3**. The first axis of the PCA accounts for 14.5% of the total variance and separates four wild individuals from present-day Mali and three wild from Nigeria from all other samples. The second axis separates wild Nigerian samples from all other Malian samples. The third axis mostly separates 12 *O. barthii* samples from Mali. In summary, the PCA indicates different outcomes in the two species, with distinct population clusters observed in the wild species, while the domesticated species forms a single, relatively homogeneous, genetic group.

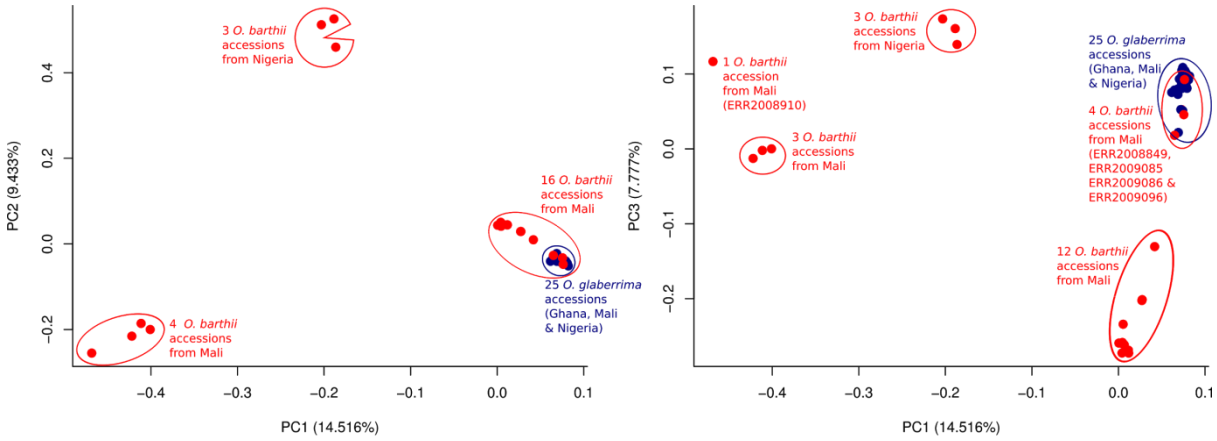


Fig. 3: *Principal component analysis of the 48 investigated samples represented by dots (left: PC1 & PC2, right: PC1 & PC3). Geographical location and species labels are based on the information provided in the Table S1 of Cubry et al. [5].*

2. Bayesian clustering: In addition to PCA, **Bayesian clustering programs** assigning individuals to ancestral populations such as Structure [12-14], TESS 2 [17], BAPS [18] are very popular tools to infer population structure. Some more recent methods used roughly similar method approach but are more adapted to large set of SNPs, e.g. FastStructure [19], LEA [20-21] or TESS3 [22]. These methods **infer the admixture proportion** of each individual (Q-value) for a given number of ancestral populations ("K"). After a Plink transformation of the input file, we use the method implemented in FastStructure to provide an example based on the African rice data (**Fig. 4**). Assuming two ancestral populations (K=2), FastStructure partially excludes 7 wild *O. barthii* samples, including 4 from present-day Mali and 3 from Nigeria, from all other samples. The individual assignment of these 7 samples suggests that these samples are admixed between the genetic cluster observed in all investigated cultivated samples (maroon) and an unknown genetic cluster (yellow). At K=3, FastStructure infers a third group containing 12 samples from present-day Mali. PCA and FastStructure have generated very concordant results concerning these

48 African rice samples. Both analyses already suggest some complexity in the evolutionary history of the African rice.

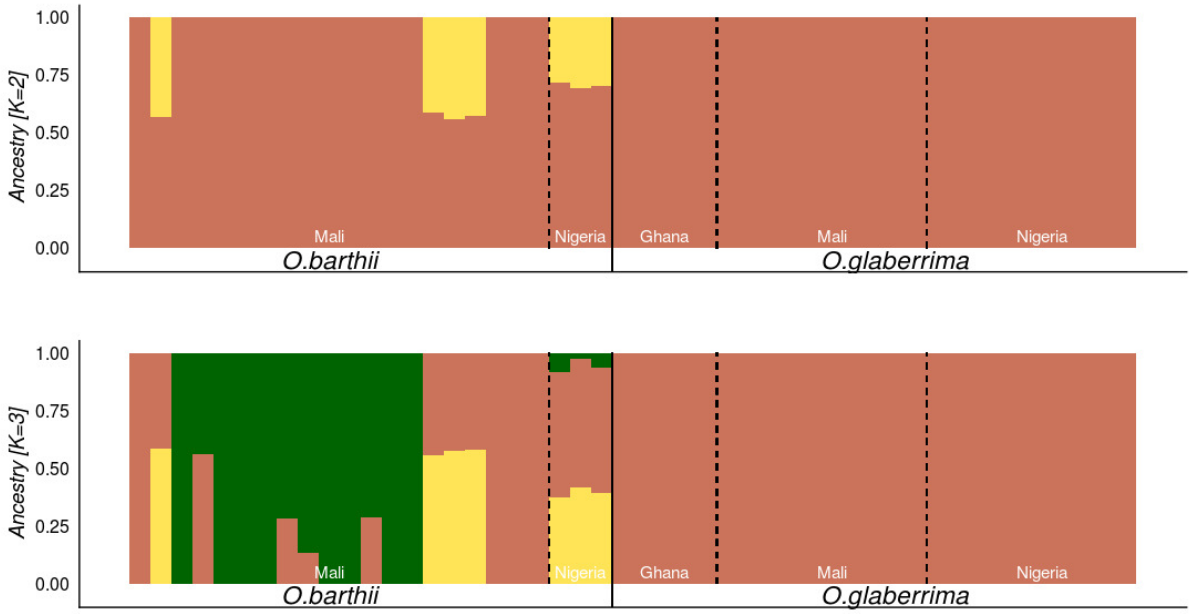


Fig. 4: *Individual assignment to two (top) or three (below) genetic clusters by FastStructure. Each bar represents a single individual, with portions of the bar colored depending on the ancestry proportions estimated assuming K=2 or K=3. The number of subpopulations that maximize the marginal likelihood is 2 (see FastStructure manual for details). Geographical location and species labels are based on the information provided in the Table S1 of Cubry et al. [5].*

3.2.4 Diversity

Nucleotide diversity greatly varies along the genome, with more genetic variation in intergenic regions than in genes. This general pattern reflects varying degrees of natural selection acting on genome, from neutral regions that do not positively or negatively affect the organism's ability to survive and reproduce (i.e. fitness), to genes under strong negative or positive selection. Negative selection refers to the purging of **deleterious alleles** at functionally constrained genes, because individuals with deleterious alleles are selected against and therefore contribute less to the next generation than the average of the population. Reciprocally, positive selection refers to the rapid fixation of advantageous mutation because individuals carrying this advantageous allele are expected to contribute more to the next generation. In both cases, it is important to keep in mind that the footprints of natural selection can extend to the vicinity of these regions because of linkage disequilibrium, over relatively long distance in regions of low recombination (generating so-called linked selection [23]).

Two important measures of nucleotide diversity are generally used in population genomics, the number of polymorphic sites (Ɵ) and the mean proportion of nucleotide differences between different pairs of sequences randomly sampled in a population (π). By comparing the diversity of different groups of individuals, a Reduction of Diversity (ROD) can be estimated by computing: $1 - \frac{\pi Group1(e.g.domesticated)}{\pi Group2(e.g.wild)}$. Such ratios are particularly meaningful for different research questions associated to plant conservation or plant breeding. For example, the total genetic diversity loss since the onset of plant domestication (or along a plant breeding program) can be investigated by comparing wild and domesticated species (e.g. wheat [24]). Based on a comparison of the 23 *Q. barthii* and 25 *Q. glaberrima* samples, an overall ROD of 0.327 is estimated, indicating that 32.7% of the *Q. barthii* diversity is lost. Genomic heterogeneity in ROD is also informative, particularly regions with very high ROD estimates (ROD exceeding 0.8 in red, **Fig. 5**). Remarkably reduced levels of nucleotide diversity in the domesticated species as compared to the wild progenitor species can be informative about candidate genomic regions (including genes) that have been subjected to strong artificial selection during domestication or breeding.

A great statistical property of π and Ɵ (to be strictly accurate, π and $\frac{\theta}{a1}$, where $a_1 = a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$) is that these two statistics are equals in values assuming mutation-drift equilibrium and constant population size ($d = \pi - \frac{\theta}{a1} = 0$, see [25]). Any excess or lack of rare alleles in the population however creates deviations from zero because π tends to underestimate the number of mutations that are rare in the population. As a consequence, the difference between the two estimators is a commonly used measure to evaluate non-equilibrium demographic situation such as population expansion (generating an excess of rare alleles, overall negative Tajima's D value) or population contraction (generating a lack of rare alleles, overall positive D value).

By observing the genome heterogeneity in Tajima's D, the footprints of natural and artificial selection can also be revealed in some specific regions of the genome. Positive values can be observed if selection maintains variation in some specific regions (balancing selection). Strongly negative values are informative about recent selection that has removed neutral variation surrounding a selected site (*i.e.* a selective sweep). Negative Tajima's D values found in a domesticated species can therefore be informative about footprints of domestication and human selection (e.g. **Fig. 5** for the case study on African rice).

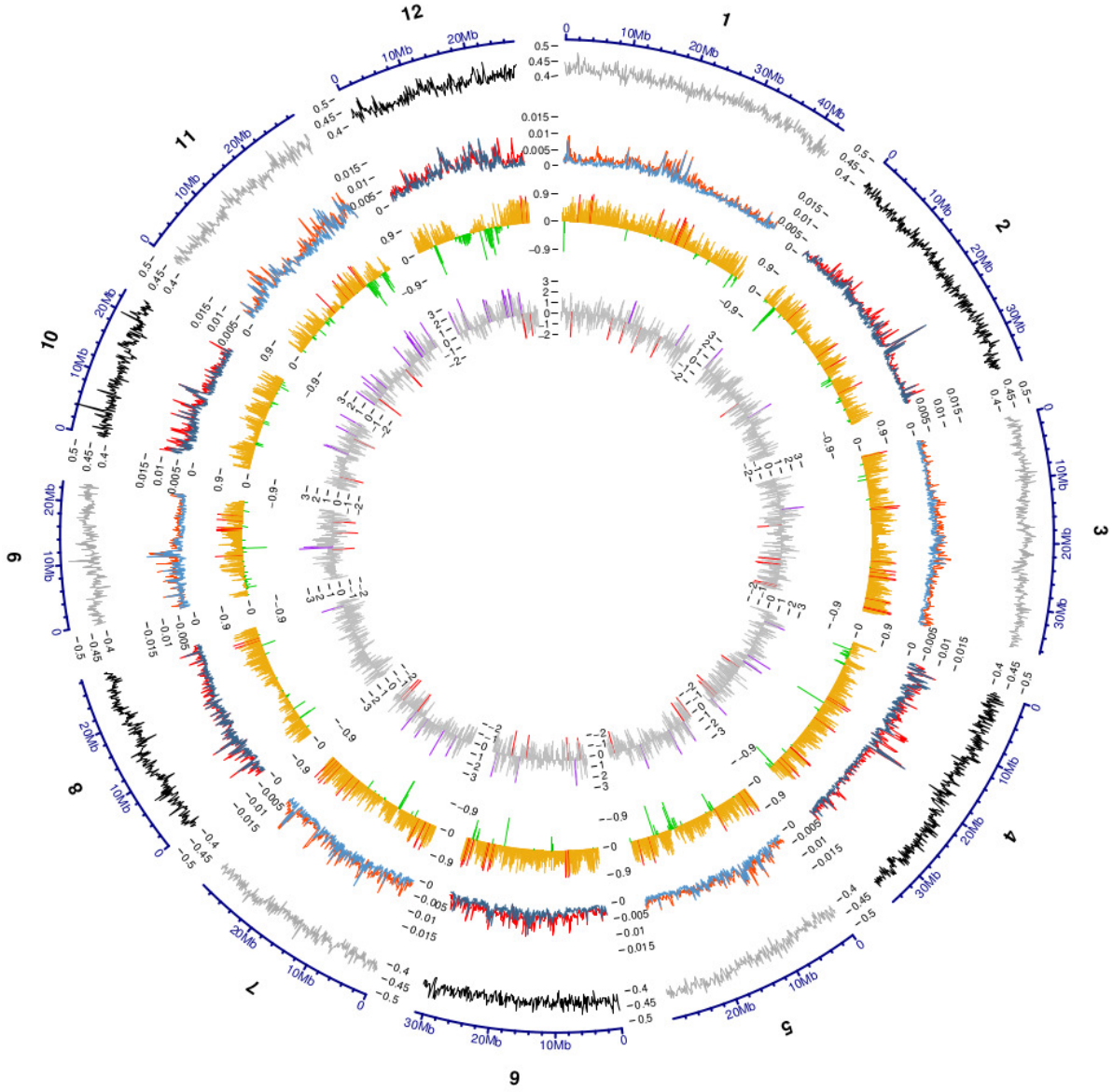


Fig. 5: *Circular diagram showing different nucleotide diversity estimates for the two African rice species along the 12 chromosomes. From external to internal: GC content; π estimates (red = O. barthii, blue = O. glaberrima); Reduction of Diversity (ROD) to evaluate the difference in the domesticated O. glaberrima species as compared to the wild O. barthii (green = negative ROD values, orange = positive ROD values, red = positive ROD values exceeding 0.8); observed Tajima's D values for O. glaberrima. Tajima's D values are represented as a deviation from the median Tajima's D values observed over the all sliding windows (D=0.171). Values lower than -1.83 or greater than 2.17 are shown in red and purple, respectively. These threshold values correspond to the -2/+2 decision rule, which is a simple rule of thumb, but remain commonly used in practice to find some candidate regions under selection. All estimates are based on non-overlapping 100-Kb sliding windows.*

3.2.5 Inferring population size history

Whole-genome sequence data are increasingly used to infer the history of a population, such as the historical changes in effective population sizes (*Ne*). *Ne* represents the number of breeding individuals in an idealized Wright-Fisher population that experiences similar amount of genetic drift than the real population (see [26] for a review). It may seem like an abstract concept, but the study of the evolution of *Ne* is particularly important in population genomics because *Ne* variation explains the dynamic of genetic diversity within a population (loss or gain) or the fixation of deleterious alleles (#3.2.6). Following the nearly neutral theory, the genetic diversity Θ equals 4 x *Ne* x μ (for a diploid species, where μ is the per-generation mutation rate). Assuming that μ remains constant over quite long periods of time, recent variation of Θ only depends of the effective population size (*Ne*) - which captures the effect of genetic drift -, with more chance for variants to be fixed by drift in small *Ne* as compared to large *Ne* populations.

To investigate this variation, many methods based on the coalescent theory are now available. Without going into details, a coalescence event occurs when two alleles merged into a single ancestral copy (*i.e.* the most recent common ancestor), when looking backwards in time starting from the present. In other words, the coalescent theory models how genetic variants sampled from a given population may have originated from a common ancestor (See [27] for an introduction). By estimating the rate of coalescence during any period of time, it is therefore possible to infer population size changes. Over the last decade, these new methods have rapidly become popular to provide information about the factors driving genetic diversity of a given species, which is especially crucial for conservation-related issues. Major shifts in the evolutionary trajectories can be identified and potentially be correlated with the major climate change periods, or with geological and anthropogenic disturbances.

The coalescent-based method implemented in SMC++ [28] is one of the best methods currently available to reconstruct the history of *Ne*. This method is fast, easy-to-use and efficient, even for analyzing tens or hundreds of unphased whole-genome sequences. We therefore performed a simple test based on the African rice dataset and observed considerable changes in past effective population size (**Fig. 6**).

As a limited number of individuals of the progenitor species had presumably been used by the early farmers and therefore contributed to the domesticated species, a drastic reduction in effective population size (*Ne*) at the onset of the domestication is generally assumed, which is commonly referred to as the domestication bottleneck. Similarly to the study of Cubry et al. [5], we inferred substantial changes in effective population size of the African rice over the last 100,000 years. Surprisingly, we were however unable to infer the expected reduction of *Ne* at the onset of the African rice domestication, but rather we inferred an expansion between 2,000 and 10,000 years ago. This lack of support for the domestication bottleneck can be due to a series of

factors such as the reduced number of genomes used or the existence of long runs of homozygosity (masked in Cubry et al. [5]). As a consequence, the pattern we have recovered over the last 10,000 years should be interpreted with caution. This result is illustrative of the importance of remaining prudent when interpreting such inferences. Violations of some assumptions can substantially distort the inference of effective population size changes. SMC++, as well as similar methods (e.g. PSMC [29]; MSMC [30]), relies on the assumption of no external gene flow (originating from another population or species). This assumption is one of the most frequently violated. Some more advanced methods available to decipher more complex evolutionary histories (*see* **Note 2**), including several closely-related species that have experienced different periods of gene flow, can also be helpful to provide additional statistical support for historical changes in *Ne* [31-33].

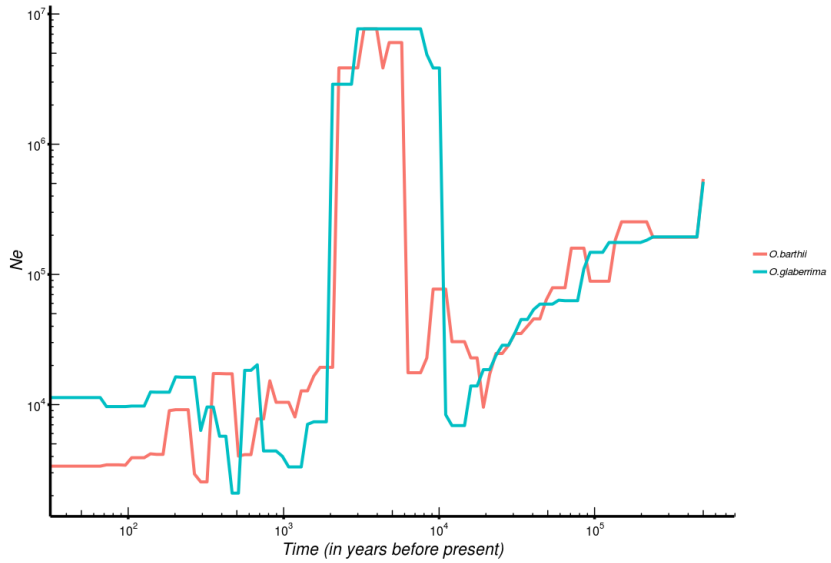


Fig. 6: *Estimated changes in past effective population sizes (Ne) for O.barthii (red) and O. glaberrima (blue) inferred using the coalescent-based method smc++.*

3.2.6 Deleterious mutation load

A downstream consequence of the domestication bottleneck is the higher load of deleterious mutations in the domesticated species as compared to the wild counterpart. Following the nearly neutral theory, neutral nucleotide diversity is expected to be reduced proportionally to the reduction in *Ne* because neutral variants have more chance to be fixed by drift in small *Ne* as compared to large *Ne* populations (#3.2.5 above). For non-neutral variants (i.e. s≠1), fixation probabilities depend on the strength of selection and effective population size ($N_e s$, *e.g.* [34]). A domestication bottleneck is therefore expected to induce a shift in the balance between selection and drift, with drift playing a greater role after the bottleneck. This also holds true for deleterious mutations, particularly slightly deleterious mutations, which are therefore

expected to accumulate more easily. In other words, the domestication bottleneck reduces the efficacy of purifying selection, the force which tends to remove harmful mutations. Domesticated plants are therefore expected to have an increased mutation load as compared to their wild progenitor species. This hypothesis is often referred as the ‘cost of domestication’ [35]. Some recent studies have provided considerable empirical support for this hypothesis, e.g. in maize [36], Asian rice [37], cassava [38] or wine [39].

In addition to the 23 and 25 WGS of *O. barthii* and *O. glaberrima*, we use sequencing data of three *O. meridionalis* (Australian wild rice individuals from [40]) and three *O. sativa* individuals (domesticated Asian rice individuals from [2]) to infer the ancestral allele of each SNP (the original non-mutated allele). In short, the recent phylogeny of the *Oryza* species based on the WGS data suggests that *O. meridionalis* had diverged from the common ancestor of African and Asian rices 2.4 million years ago. The divergence of African and Asian rices is more recent (<1 million years ago, see [40] for details). Australian and Asian rices are used to infer the ancestral allele in order to count the number of derived alleles in the wild and the domesticated African rice. Based on the SNPs for which the ancestral allele was unambiguously determined, we identify more fixed derived alleles in *O. glaberrima*, as compared to *O.barthii* (1,050,545 and 825,826, respectively), which can be considered as another piece of proof supporting the hypothesis of a domestication bottleneck.

To look into more details the burden of deleterious genetic mutations, various methods are available. Simple methods such as the comparisons of ratios of the nucleotide diversity (or heterozygosity) at non-synonymous as compared to synonymous polymorphisms can be very relevant (*e.g.* between a wild progenitor and a domesticated species [41]). Indeed, most within-gene mutations changing the amino acid sequence are expected to be slightly or strongly disadvantageous (i.e. deleterious). Higher ratios of non-synonymous to synonymous polymorphisms are therefore informative of higher deleterious loads. *In silico* methods predicting the potential deleterious effects of mutations are more and more popular (e.g. SIFT [42]). Subsequently, we use the software PROVEAN (PROtein Variation Effect Analyzer [43]) which performs local alignments (BLAST) against a protein database to predict whether an amino acid change in a given protein affects its function. A score is then computed based on the 30 best cluster hits. A negative PROVEAN score is indicative of a deleterious mutation.

This analysis requires different steps, which are detailed on github (https://github.com/ThibaultLeroyFr/Intro2PopGenomics/tree/master/3.2.6/Scripts_provean/). Before running PROVEAN, we have built a NCBI 'non-redundant' (nr) database containing only proteins corresponding to monocots species. By limiting to

monocotyledons species, our objective is to avoid spurious BLAST alignments against evolutionary distant species. Among a total of 120,324 candidate non-synonymous mutations passing PROVEAN filtering criteria, 18,369 mutations are predicted to be putatively deleterious mutations (score < -2.5). Among these 18,369 SNPs, the ancestral state is unambiguously determined for 11,829 variants (see above). Deleterious allele frequency spectra at these 11,829 putatively derived deleterious SNPs are generated for both the wild and the domesticated species (**Fig. 7**). Interestingly, a higher mutation load in *O. glaberrima* as compared to *O. barthii* is identified, but this difference is relatively small. Our analyses are rather consistent with substantial deleterious mutation load *O. barthii* and a slight increase in *O. glaberrima*, which can be compatible with the African rice domestication.

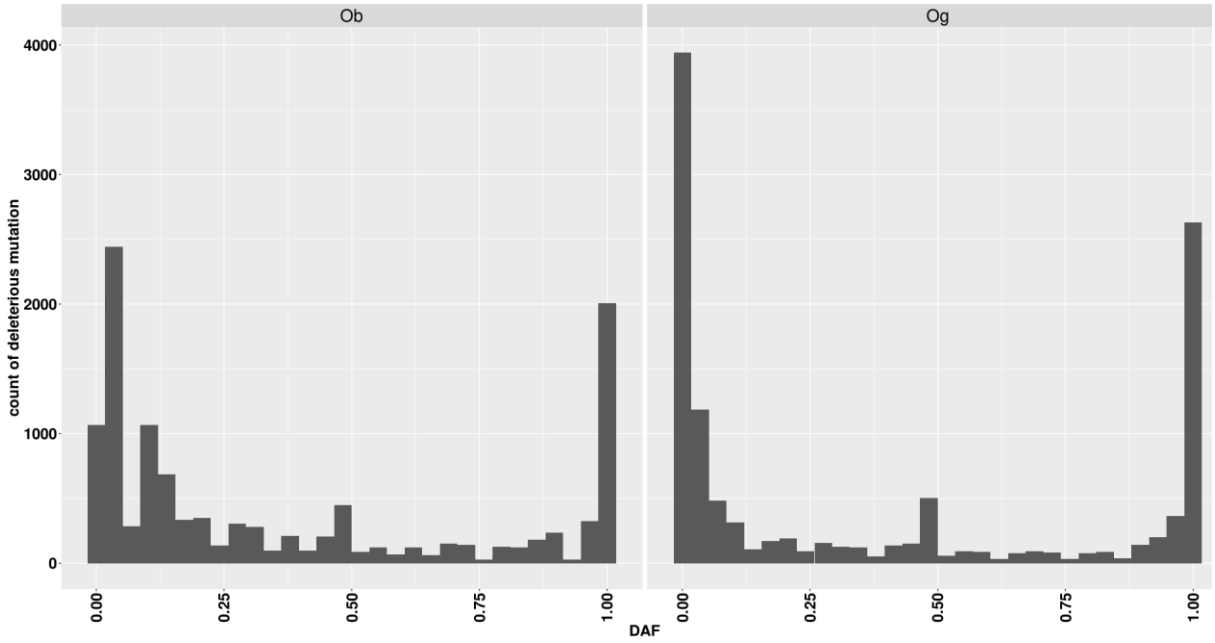


Fig. 7: *Deleterious mutation loads in the wild O. barthii and the domesticated O. glaberrima species, as estimated using proteins of the African rice. DAF = Deleterious Alleles Frequencies.*

Looking at this difference more carefully, the number of deleterious alleles per individual is slightly higher in *O. glaberrima* (**Fig. 8**), but this difference seems to be more explained by a difference in heterozygous sites than by a strong difference in the number of homozygous deleterious variants. Because deleterious mutations tend to be recessive [44], such a limited difference in the number of homozygous variants therefore suggests that this higher mutation load may only induce a marginal fitness difference between the two species. This first investigation already gives an overview

of the accumulation of deleterious variants, but some analyses are available to conduct more precise measurements [45-47].

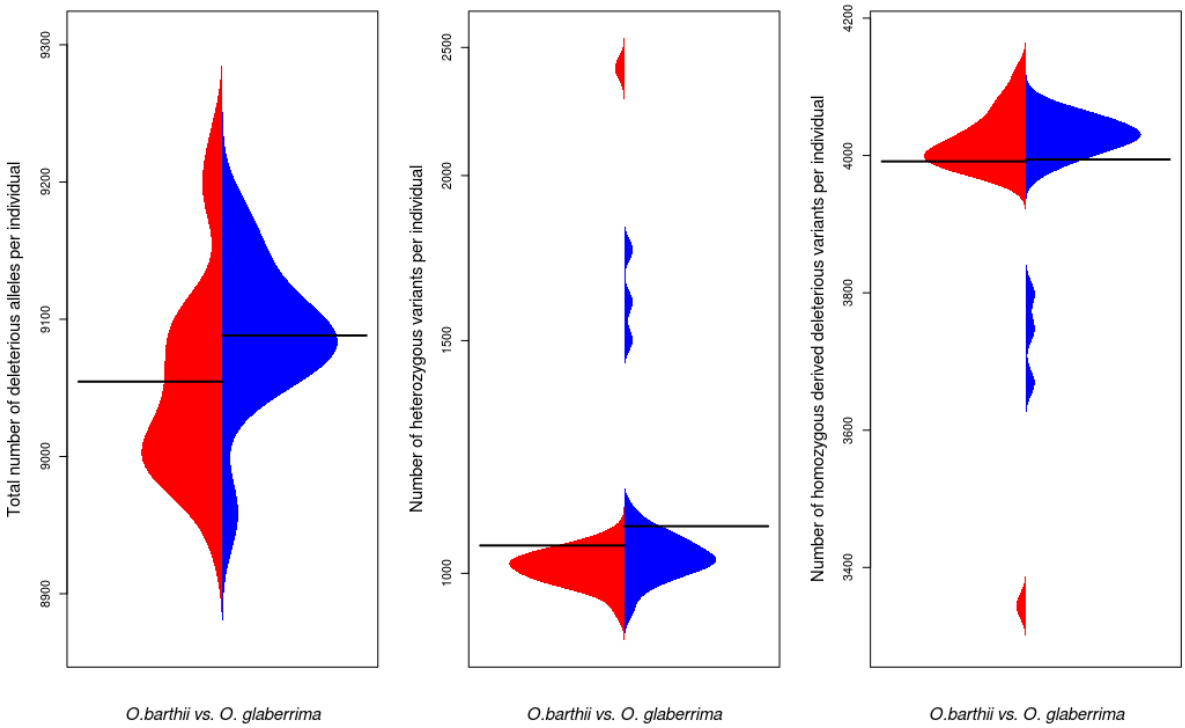


Fig. 8: Total numbers of deleterious alleles (left), heterozygous calls (centre) and homozygous derived alleles per individus for *O. barthii* (red) and *O. glaberrima* (blue). The black bar indicates average per species.

3.2.7 F_{ST} & genome-scans for selection

The fixation index F_{ST} is probably the most widely used population genetic statistics. F_{ST} measures the differentiation between populations and ranges from 0 to 1. However, some slightly negative values can be observed in the case of uneven sample sizes and should be interpreted as a zero value. A value of zero indicates complete panmixia, *i.e.* free interbreeding between the two assumed populations resulting in no population structure or subdivision. On the contrary, a value of 1 indicates that the two populations are homozygous for two different alleles (*e.g.* a SNP with genotypes A/A observed in all individuals of the first population and genotypes C/C for all individuals of the second population). In other words, the higher the F_{ST} value is, the more different the allele frequencies in the two or more populations are.

To give a better idea of how useful report of F_{ST} values can be, we compute F_{ST} between samples of *O. barthii* and *O. glaberrima*, at two different genomic scales: on a SNP-by-SNP basis or using 10-Kb sliding windows (**Fig. 9**).

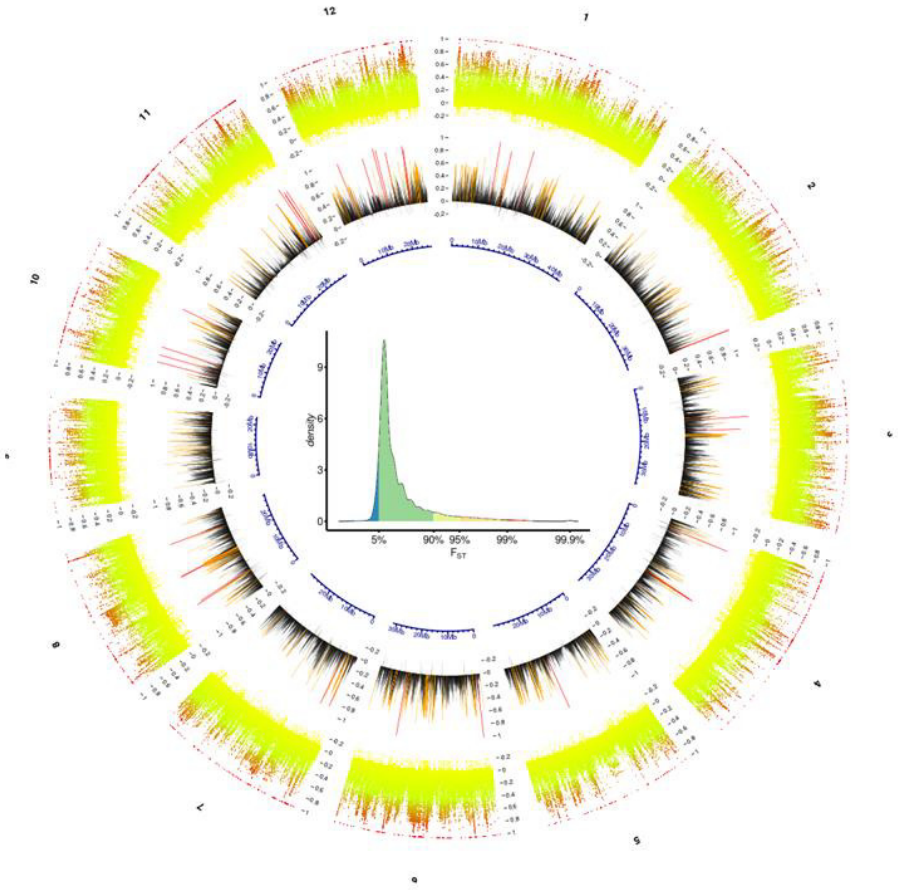


Fig. 9: *Fixation index (F_{ST}) values as computed using vcftools and estimated for each SNP (external circle) or for non-overlapping 10-kb sliding windows (internal circle) over the 12 rice chromosomes. A color scale from yellow (F_{ST} =0) to red (F_{ST} =1) is used for the SNP-by-SNP F_{ST} estimates to illustrate the continuous of variation in F_{ST} values. Empirical distribution of the observed F_{ST} values across all SNPs is shown in the center of this circular graph (corresponding F_{ST} values for the different quantiles: 5% = -0.03; 90% = 0.27; 95% = 0.41; 99% = 0.66 & 99.9% = 1.00)*

The use of the empirical distribution of the among-locus variation in F_{ST} (**Fig. 9**) to identify loci that deviate from neutral expectations - and therefore representing candidate footprints for natural or artificial selection - is inspired by the seminal study of Lewontin & Krakauer [48]. Indeed, loci under balancing selection in the two populations are expected to exhibit lower F_{ST} values, while regions under diversifying selection are expected to exhibit larger differences in F_{ST} as compared to selectively neutral loci. Diversifying selection indeed triggers allele frequency changes over time in such a way of generating and maintaining high genetic differences in the two populations. In practice, identifying loci under balancing selection is a near-impossible task to achieve. Identifying **diversifying selection** remains a complex issue. The difficulty comes from the fact that the among-loci variation in F_{ST} is highly dependent on the demography of the investigated populations [49-53]. Over the last

20 years, considerable attention has been devoted to develop statistical approaches that partially address this challenge (e.g. [54-55]; hereafter referred to as genome scans for selection). In this section, we introduce the use of PCAdapt [56], a R package that is well suited to identify variants with large differences in allele frequencies between clusters of individuals. This package has several advantages. From the user's perspective, this solution is easy to use under an *R* environment, especially with the detailed tutorial available for this package. From a more computational and biological perspective, PCAdapt is computationally-efficient and the analyses do not require to group individuals into populations - i.e. no prior information about the two or more populations, which can be a difficult task to achieve (e.g. #2.2.3 for the African rice). In addition, PCAdapt can handle very large datasets and reports summary statistics in a reasonable computational time, offering an alternative to the genome scans methods based on a Bayesian framework, which are several orders of magnitude longer (see #3.3.6 for the use of a Bayesian framework).

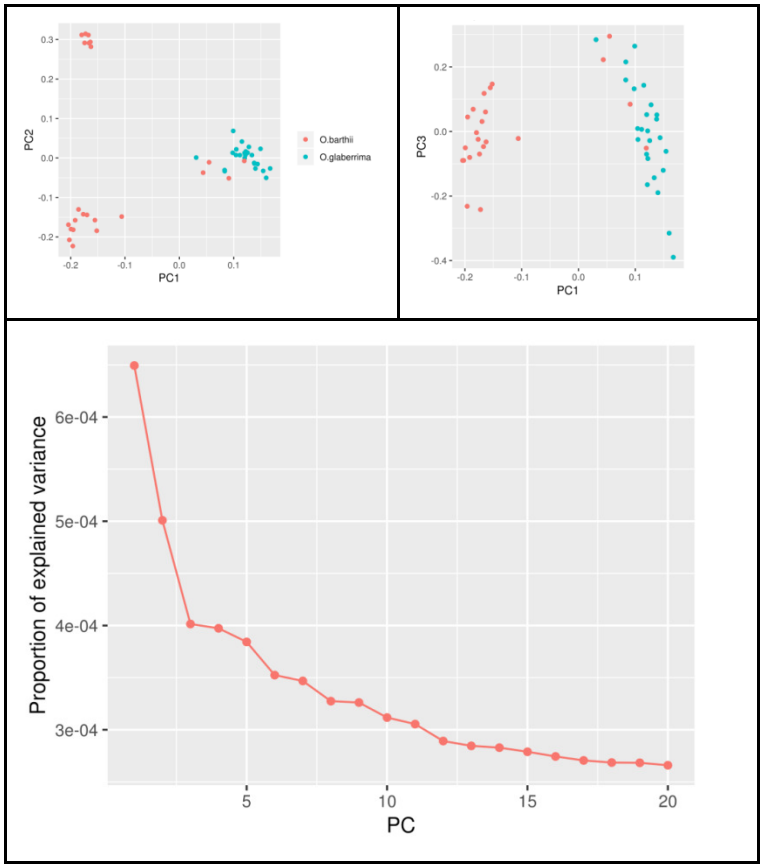


Fig. 10: *Individual PCA and scree plot after LD thinning. Top left) Coordinates of individuals on the two principal components. Top Right) Coordinates of individuals on the PC1 and PC3. Below) Scree plot (proportion of explained variance) for the 20 first PCs after LD thinning. Based on this screeplot, K=3 was preferred.*

After a preliminary analysis revealing some regions of strong linkage disequilibrium (LD) in the African rice dataset, the African rice dataset is pruned to remove SNPs in strong LD. Indeed, such an extent of LD is expected to have a considerable impact on the analysis (*see* **Note 1**). As a consequence, the dataset is first "pruned" to remove SNPs in strong LD, before computing the principal components and performing the outlier detection. Coordinates of individuals on the two principal components (PC) (**Fig. 10**, as compared to **Fig. 3**) are different after SNP pruning. This reduction of the LD likely improves the ability of the PCs to capture the genome-wide patterns reflecting ancestry differences, as commonly assumed [57]. The first PC mostly isolates samples from *O. barthii* and *O. glaberrima*, with the notable exception of 4 *O. barthii* samples from *Mali.* Visual evaluation of the so-called scree plot [58] for PC1 to PC20 suggests that the 3 first components explain a substantial fraction of the total variance in the data, as compared to the 17 additional components that were also investigated (Fig. 10). As a consequence, we use the implemented method in PCAdapt to scan genomes assuming these 3 components.

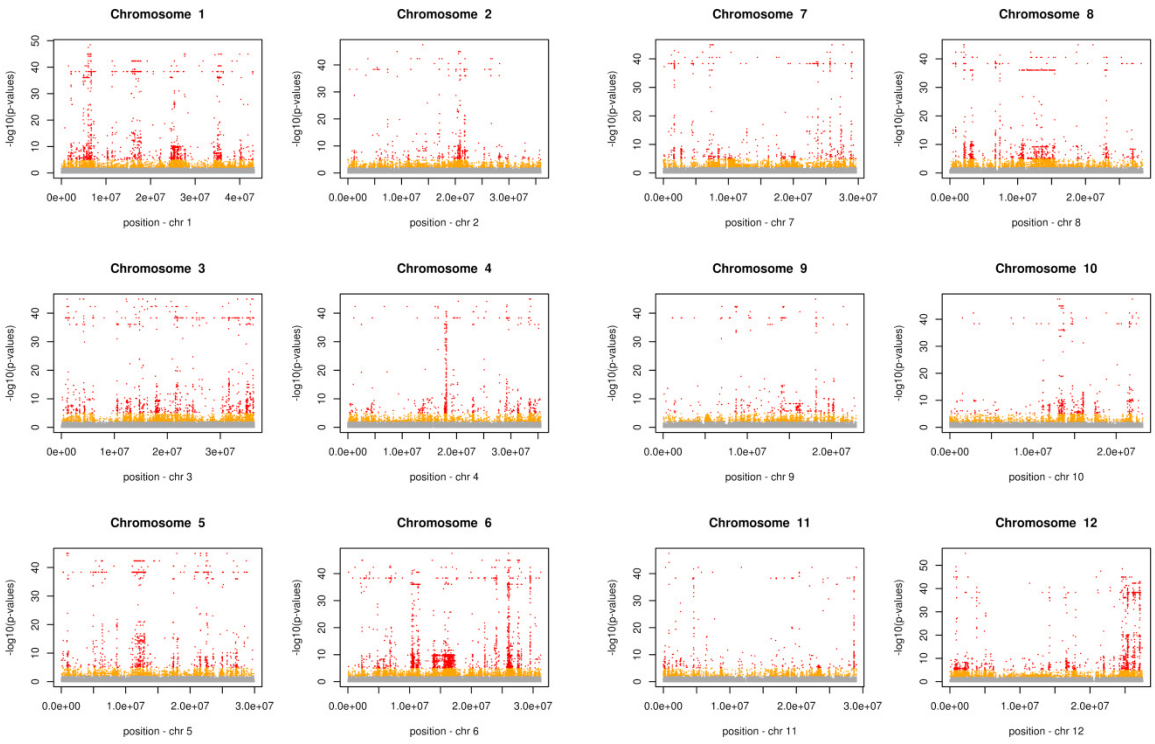


Fig. 11: *Manhattan plots showing the chromosome position of each outlier detected using PCAdapt and assuming K=3. Score is expressed as -log$_{10}$(p-values). SNPs with pvalue < 0.01 (*i.e. *-log$_{10}$(p-values)=2) are shown in orange and pvalue < 0.00001 (-log$_{10}$(p-values)=5) are shown in red.*

The genome positions of all outliers as shown in the so-called Manhattan plots (**Fig. 11**) reveal that they are distributed throughout the genome. SNPs deviating from

neutral expectation and therefore potentially under selection are unexpected to have this distribution, since selection is unlikely to impact all the genome. These outputs are more consistent with a substantial background noise generating an excess of outliers. However, some genomic regions exhibiting hundreds of variants in several narrow genomic regions, e.g. on chromosomes 4 or 6 (**Fig. 11**) are more convincing. These regions therefore represent excellent candidate regions to identify the African rice domestication genes.

3.3 Second case-study: sessile oak populations

3.3.1 Pool-seq as a cost-efficient method

For many plant species, the sequencing of hundreds or more individuals using an individual-based strategy represents a too expensive option. Considering for example, the sequencing of 50 diploid individuals at reasonable sequencing coverage (20X) - the total sequencing effort would be around 1,000X - in order to ensure accurate individual calls for all individuals. For some biological questions, the genotypes of all individuals are not truly necessary. Instead, accurate population estimates of the frequency of each allele along the genome can be sufficient [59]. In this case, a cost-effective alternative remains possible. The strategy is to first equimolarly mix the DNA of these 50 individuals prior to sequencing in order to sequence the pool at a lower coverage. Assuming that the pool is sequenced at 100X (so resulting in a 10-fold drop in the sequencing cost), each chromosome is therefore expected to be sequenced only once, on average, which is low. But, given the total number of chromosome sequenced in the pool, the allele frequency estimated for the whole population is expected to be accurate. Based on mathematical derivations, Gautier et al. [60] provided theoretical support for this accuracy. These authors showed that the sequencing of DNA pools remains an efficient strategy under various realistic experimental designs. They also provide an easy-to-use tool (PIFs [60]) to optimize the experimental pool-seq design considering several parameters or experimental errors (*e.g.* pipetting biases).

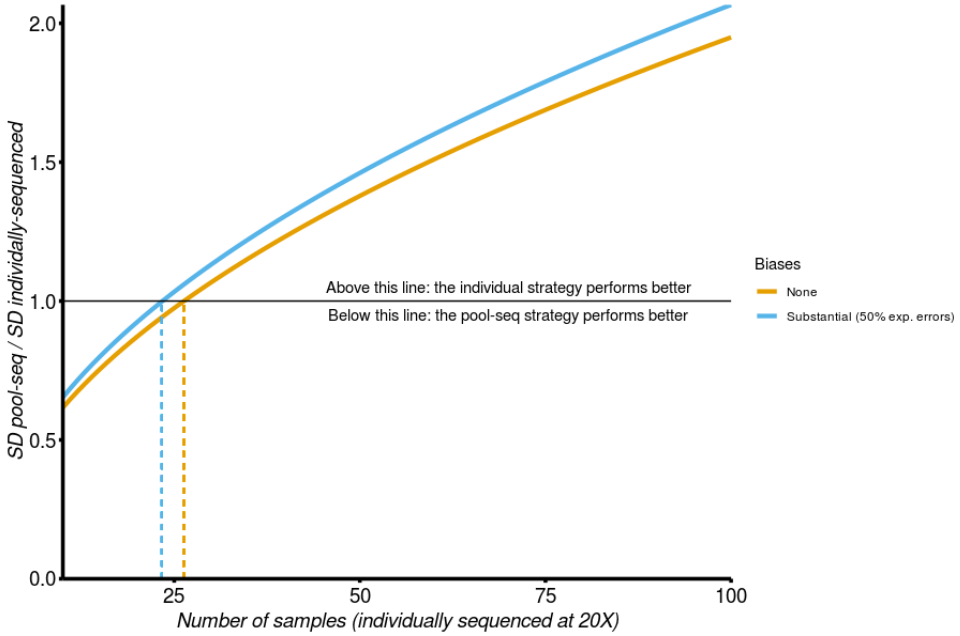


Fig. 12: *Comparison of the accuracy in the allele frequency estimation between two strategies, as performed using PIF [60]: a pool-seq strategy of 50 individuals sequenced at a mean pool coverage of 100X and an individual-based genotyping strategy with a growing number of individually sequenced at 20X. The tipping point is 26 individuals assuming no experimental biases. Even after considering some experimental biases, a pool-seq strategy of 50 individuals sequenced at a pool coverage of 100X is expected to outperform a design with 20 individuals sequenced at 20X (the equivalent of 400X of sequencing data; for details, see [60]).*

Based on a rapid simulation using this tool and the number of individuals previously assumed (**Fig. 12**), it indicates that the sequencing of a pool of 50 individuals with a mean pool coverage of 100X is expected to generate as accurate allele frequency estimates as 26 individuals separately sequenced with a depth of coverage of 20X (the pool-seq strategy therefore reduces by 5 the sequencing costs). Even assuming substantial experimental error (50%) generating departure from equimolarity (i.e. a dispersion of individual contributions around the expected mean value assuming equal DNA quantities), the allele frequency estimates are expected to be roughly similar to 23 individuals separately sequenced, each with a depth of coverage of 20X (**Fig. 12**).

3.3.2 Population genomics in wild sessile oaks

The sessile oak (*Q. petraea*), a species belonging to the European white oaks complex, is an example of plant species with an impressive amount of genomic resources, including huge pool-seq data ([6;61-62]). Sessile oaks extend from

Northern Spain to Southern Scandinavia, thus representing a large diversity of climatic conditions (**Fig. 13**). In South-West French Pyrenees, some sessile oak populations occur from lowlands to middle elevations (up to 1,600 meters, **Fig. 13**), with substantial differences in mean annual temperature (up to 7 degrees Celsius) or in precipitation sums (a difference of up to 250 mm/year, [6] for details). In the subsequent sections, we perform a step-by-step reanalysis of the data used in Leroy et al. [6] to illustrate the possibilities of the pool-seq data. In this study, 18 pools were sequenced: 10 sessile oak populations collected on a latitudinal gradient in Europe (including 7 populations from France, 2 from Germany and a population from Ireland) and 8 sessile oak populations from an altitudinal gradient in the French Pyrenees (collected along two close valleys, with 4 populations per valley (100m, 800m, 1200m and 1600m). The DNA of 20-25 individuals were equimolarly mixed prior to sequencing, except for the two populations at 1600 meters for which only 10 to 18 individuals were used (for details, see [6]). Analyses performed in this section are basically performed following the same strategy than in the original paper, but the analyses are simplified.

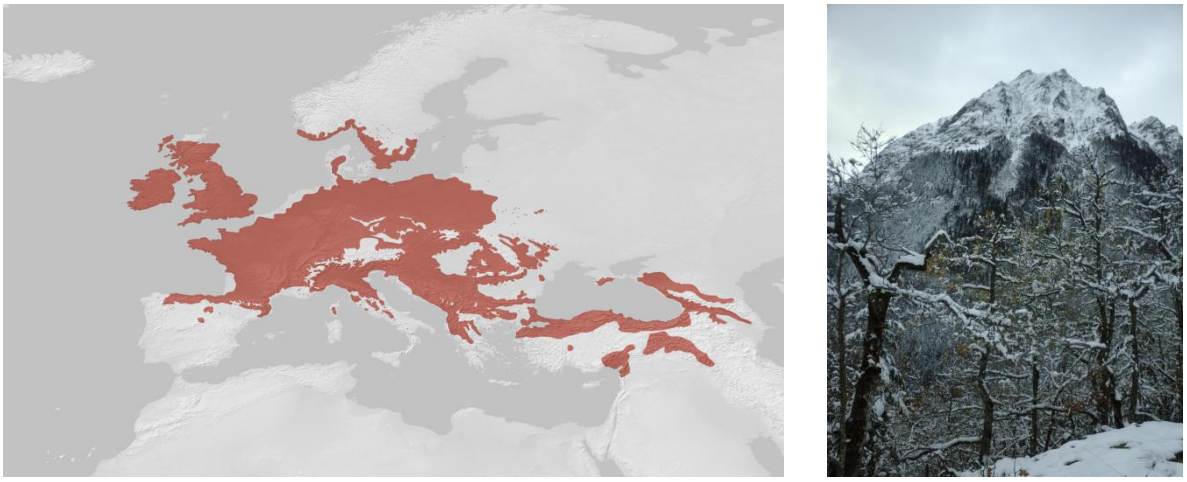


Fig. 13: *Sessile oak distribution and climate variation. Left: European distribution map of Q.petraea created with QGIS from data made available by the European Forest Genetic Resources Programme (EUFORGEN [63]). Right: Sessile oak trees in the snow. Photo taken by T. Leroy on November 22rd, 2015 at an elevation of 1200 meters in one of the French Pyrenees forests investigated in Leroy et al. [6] ('O12'* population*).*

3.3.3 From raw sequencing data to allele counts

The Illumina data can be downloaded from SRA or EMBL-EBI using the project ID PRJEB32209. We make available on github all the scripts used to download and perform the trimming, read mapping and to identify variants. The

pipeline is roughly similar to those used for the African rice data, at least for read trimming and mapping. A notable exception is the way in which variants are identified. As previously described (#3.1), variant calling methods have been developed to minimize the number of false-positive variants (e.g. sequencing errors). Indeed, each diploid individual possess either two copies of the reference allele (homozygous for the same allele than the reference genome), one copy (heterozygous, with both a reference and an alternative allele), or none (homozygous for the alternative allele). In other words, the frequency of the reference allele estimated for each individual is expected to be close to 1, 0.5 or 0. When the coverage is high enough (> 20), deviations from these situations can be informative of false positive SNPs. In contrast, such investigations are impossible to perform with pool-seq data because DNA from several individuals are mixed prior to sequencing. As a consequence, only few parameters can be used to exclude false positive SNPs, *i.e*. the minor allele frequency (MAF) and the depth of coverage per at each position. Illumina sequencing errors are expected to be about 1% or less, so it is generally recommended to use a MAF that exceeds this value (e.g. 2% or more). Similarly, coverage is expected to vary across the genome following a Poisson distribution [64]. Extreme values in the observed distribution of coverage depth are also informative from some read-mapping biases inducing an excess or deficit of coverage compared to the expectations assuming this distribution. For example, highly covered regions can be due to reads corresponding to two genomic loci with almost similar sequences (*e.g.* recent duplications) aligning to a unique location of the reference sequence. Such regions therefore present a high risk of identifying false positive SNPs. In practice, a matrix of allele counts (**Fig. 1** & Table 1) contains both allele frequencies and coverages that can be used to filter variants.

One thing must be kept in mind, however: errors in pool-seq data are necessarily more numerous than in individual-based sequencing. Even after using some MAF or coverage thresholds, the number of false positive SNPs can remain substantial. Population-level estimates of nucleotide diversity can be greatly inflated, especially for species with low to extremely low genetic diversity, for which the noise-to-signal ratio can be high. In this section, we choose not to cover diversity-related analyses (including comparisons of estimators, *e.g.* Tajima's D) based on pool-seq data to call for caution. It must however be noted that some methods already exists (e.g. Popoolation [65]) and some studies successfully reported similar range of estimates based both on individual and pool-seq datasets (e.g. oaks [62]).

Table 1: *a hypothetical example of a read count matrix with two SNPs in rows. The two pools are assumed to be sequenced at a mean pool coverage of 100X. Allele frequencies can be easily derived from this matrix (e.g. 30/(75+30)=0.29 for the pop1 of the SNP Chr1:47).*

Chromosome	Position	Ref allele	Major allele (all populations)	Minor allele (all populations)	Major allele counts (pop1)	Minor allele counts (pop1)	Major allele counts (pop2)	Minor allele counts (pop2)
Chr1	47	G	G	C	75	30	49	55
Chr1	112	T	A	T	68	20	79	14

3.3.4 Inferring the history of a set of populations

Allele frequencies are expected to be very informative about historical relatedness between populations. Indeed, two populations that have a recently shared history are expected to exhibit more similarities in allele frequencies because of a low influence of genetic drift, as compared to two genetically distant populations. As a consequence, inferring the history of a set of populations based on allele frequencies is expected to be possible. This is exactly what Treemix [66] aims to do. This genetic tool infers the relationships among populations as a bifurcating tree, which can therefore be considered as an analogous to phylogenetic trees. To do so, the software first infers the variance-covariance matrix of allele frequencies between population based on a large set of variants and then finds the maximum likelihood tree of populations than explains most of the observed variance in relatedness between populations.

In the case of sessile oak, TreeMix computes the 18x18 variance-covariance matrix using a huge set of SNPs (37 million SNPs). Because the allele frequencies at nearby SNPs are expected to be highly correlated due to linkage disequilibrium (*see* **Note 1**), we set the parameter k to 1,000 (blocks of 1,000 SNPs) to take into account this bias. TreeMix therefore first estimates the variance-covariance matrix based on 37,062 blocks of 1,000 SNPs.

Using the R scripts from the TreeMix suite, the total variance explained by a simple bifurcating tree can be estimated. Applied to the sessile oak dataset, drift alone accounted for more than 89% of the total variance in allele frequencies among populations. An example of phylogenetic visualization of the inferred best likelihood tree is shown in **Fig. 14**. As a first step, it provides a lot of information regarding the relatedness of populations. For example, sessile oak populations from the latitudinal gradient are genetically different from the populations from the altitudinal gradient,

especially the six populations at the highest elevation (**Fig. 14**). The population from Ireland however departs from this general pattern, since this population is more related to Pyrenean populations at high elevation.

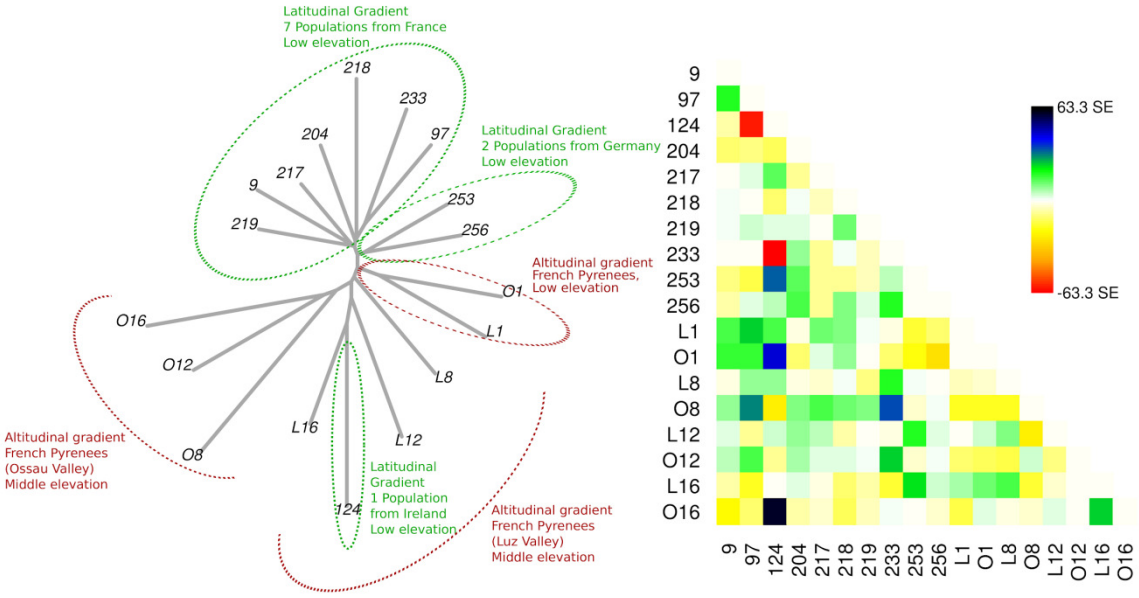


Fig. 14: *Population splits inferences under TreeMix assuming a simple bifurcating tree (no migration nodes). Left: Unrooted visualization of the best likelihood tree. Unlike in the study of Leroy et al. [6], we do not use additional species to root the tree, i.e. to find the most basal ancestor of the tree, but only perform the inference based on the 18 sessile oak populations. Right: Visualization of the matrix of residuals. For example this matrix shows that populations 124 and O16 have a remaining variation in relatedness (black square) that is not captured by the bifurcating tree.*

In the great majority of cases, a simple bifurcating tree cannot explain all the genetic variation observed in the variance-covariance matrix. TreeMix allows adding some additional edges connecting distant nodes or branches. These events can be interpreted as different migrations events, either ancient or contemporary, that have contributed to generate populations with a mixed ancestry (so-called admixed populations). We can therefore perform simulations for a range of migration events (m).

By adding different migration events, the likelihood of the model (or the total variance explained) is expected to increase (**Fig. 15**). For example, adding a single migration node substantially increase the proportion of explained variance (+3.1%, see **Fig. 16** for the corresponding tree topology).

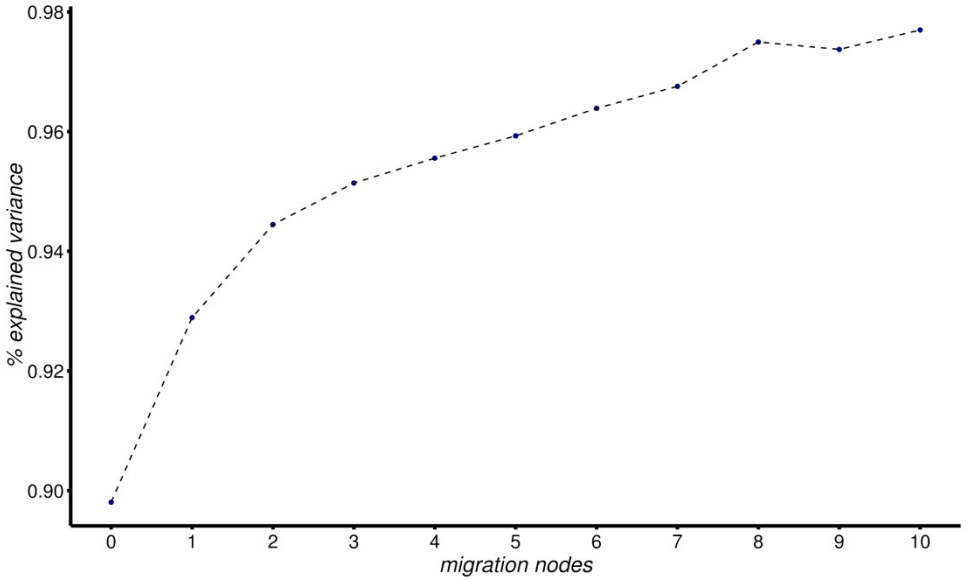


Fig. 15: *Proportion of the variance explained for a growing number of migration nodes. Only one simulation was performed per migration node.*

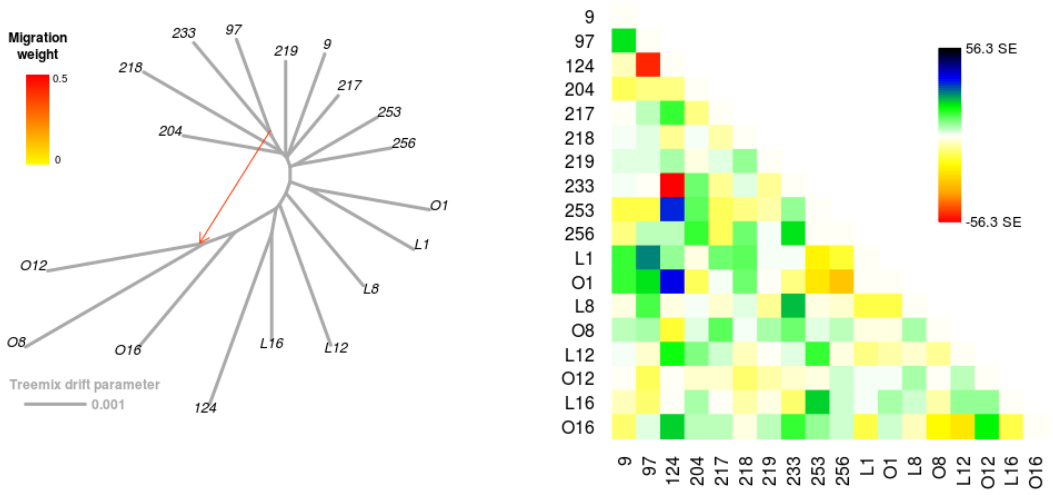


Fig. 16: *Population splits inferences under TreeMix assuming a simple bifurcating tree and a migration node. Left: Unrooted visualization of the best likelihood tree and the inferred migration node. Right: Visualization of the matrix of residuals for this best tree. Unlike in Fig. 14, no strong excess of remaining variation in relatedness between populations 124 and O16 is observed.*

Admixture between populations can be tested using three- and four-population tests. These *f3* and *f4* tests were developed by Reich et al. [67] and Keinan et al. [68], respectively, and are implemented in the TreeMix suite. The tree-population test *f3*(A;B;C) aims at testing if a given population A is admixed between two other

populations (B and C). Negative f3 values are indicative of admixture (see [67] for methodological details and [6] for empirical tests on oak data).

3.3.5 F_{ST} Fixation indices

Several bioinformatic solutions were developed to compute measures of differentiation between pools such as F_{ST} (see #3.2.7 for general information about F_{ST}). Popoolation2 [69] is probably the most widely used program for this purpose. In this section, we used the new estimator of F_{ST} recently developed by Hivert et al. [70], because of its higher robustness to different sources of bias associated with pool-seq ([70] for details). In addition, this new F_{ST} estimator is implemented in a R package ("poolfstat" [71]) which also generates input files for BayPass [54], the genome scan method used in the section 3.3.6.

Using the R package *poolfstat*, the *computePairwiseFSTmatrix* function can be used to calculate pairwise F_{ST} values over the whole dataset, which can be useful to have a rapid overview of the genetic structure among the different pools (**Fig. 17**).

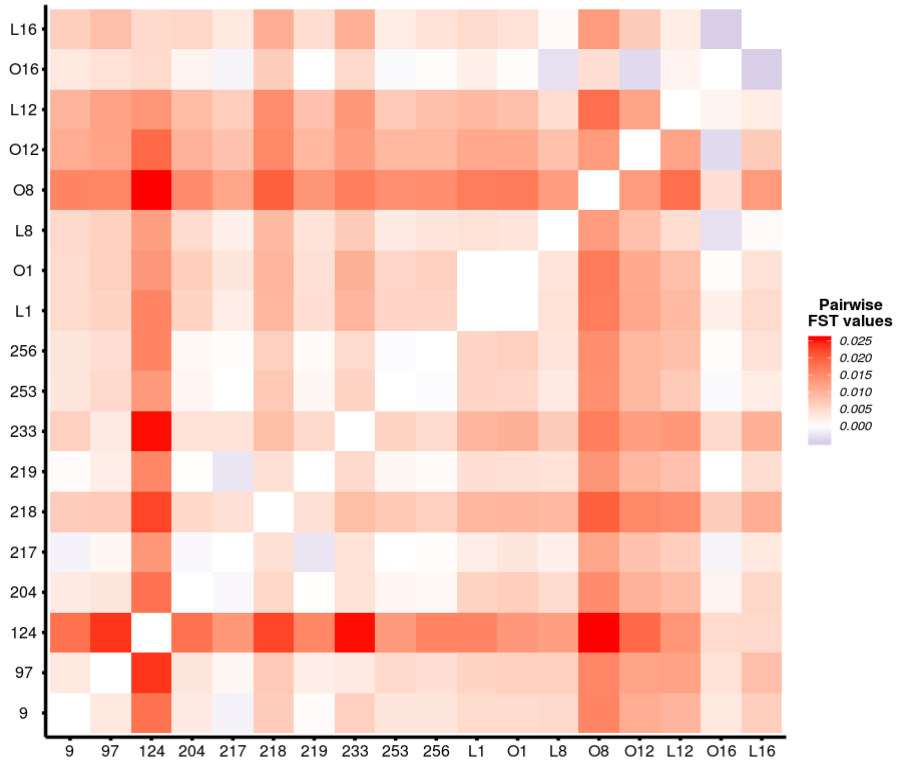


Fig. 17: *Pairwise F_{ST} values between the 18 sessile oak pools, as computed by the R package poolfstat. To speed up computations, computations were performed on a random selection of 100,000 SNPs among the whole SNP set.*

F_{ST} values can also be computed for each SNP using the *computeFST* function to detect SNPs that exhibits very high levels of differentiation among all pools (black line, **Fig. 18**). F_{ST} values can also be estimated for each SNP and each pair of pools

using the *computePairwiseFSTmatrix* function with the following argument "output.snp.values = TRUE" (grey lines, **Fig. 18**).

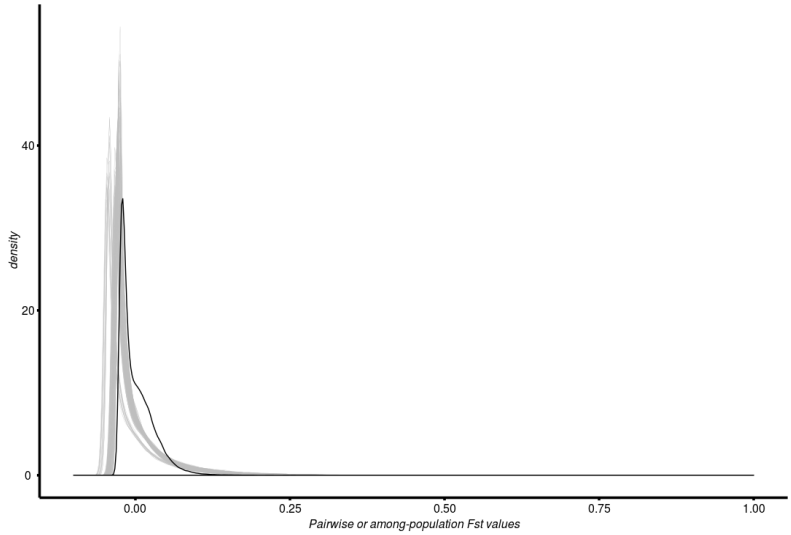


Fig. 18: *Distributions of Pairwise (grey) and among-population F_{ST} (black) values. Each grey line corresponds to the distribution of F_{ST} for one of the 153 (i.e. $\frac{18*(18-1)}{2}$) possible pairs.*

3.3.6 Genome-scans of selection

Unlike the genome scan for selection performed for the African rice (section #3.2.7), we use a Bayesian framework to detect footprints of natural selection. We have chosen the method implemented in BayPass [54], which is equally suited for pool-seq and individual sequencing data. Many other methods are available and of interest too, including Bayenv [72-73]. Core models of Bayenv and BayPass are indeed very similar. First, the population structure is captured by computing a covariance matrix of allele frequencies across all populations. This matrix is particularly convenient since it makes technically possible to perform extensive neutral simulations assuming this inferred covariance matrix in order to calibrate a measure of differentiation (Pseudo-Observed DataSets, PODS) and then identify threshold values based on these neutral simulations. Under Bayenv or BayPass, the differentiation metric used is the XtX, which can be considered as a SNP-specific F_{ST} explicitly accounting for the population structure. Outlier SNPs are the observed variants (red, in **Fig. 19**) deviating from neutral expectations, i.e. those exhibiting greater XtX values than expected based on the simulations (black, **Fig. 19**).

XtX outliers are not randomly distributed along the genome, but rather cluster in several genomic regions (black dots, **Fig. 20**). All these regions show an excess of differentiation among populations as compared to the expectations based on the variance-covariance matrix.

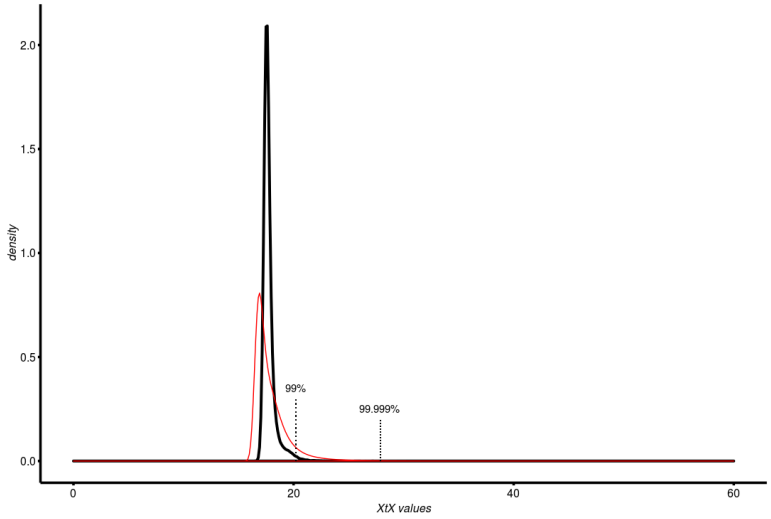


Fig. 19: *Distributions of the XtX values for the observed dataset (red) and for the simulations assuming the variance-covariance matrix (black). Thresholds corresponding to the top 1% and 0.001% of the XtX values based on simulations are shown by the dotted lines.*

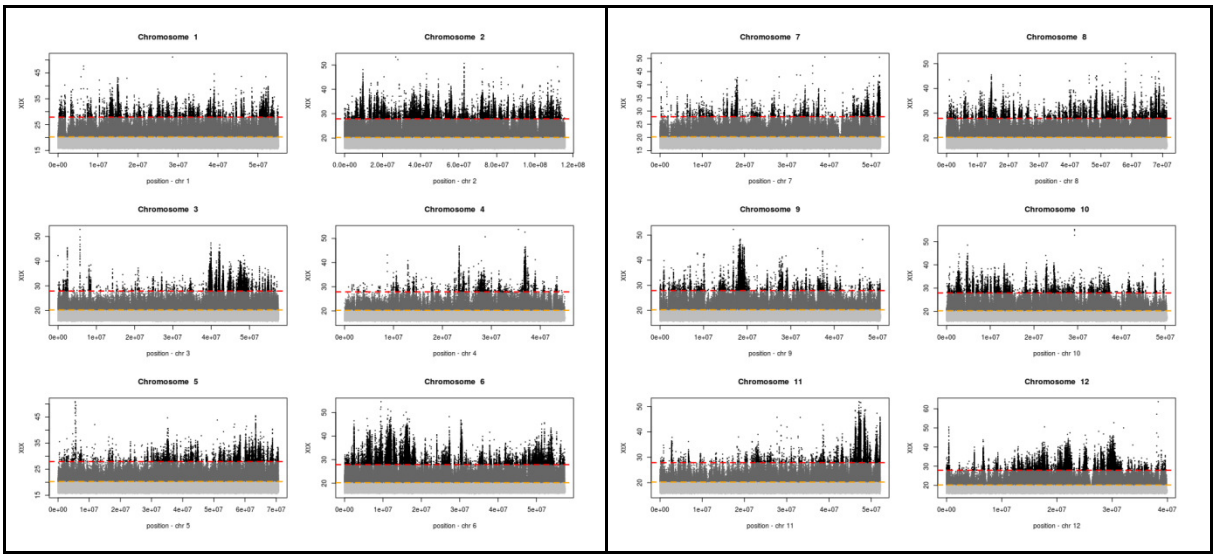


Fig. 20: *Manhattan plots showing the chromosome positions of all SNP and the corresponding XtX value as computed under BayPass. SNPs with empirical XtX values exceeding the 99% and 99.999% thresholds based on Pseudo-Observed DataSets (PODS, orange and red lines, respectively) are shown in dark grey and black, respectively.*

3.3.7 Genotype-environment association (GEA)

BayPass can also identify association between allele frequencies differences and population-specific covariables, such as environmental or phenotype data (e.g. temperature, height or yield). Assuming that climatic or phenotypic data is available

for the set of populations under investigation, it is possible to identify allele frequency variation along these climatic or phenotypic gradients (so-called genetic clines). Associations to environmental covariables are often referred as to Genotype-Environment Associations (GEA), while associations to phenotype are often referred as Genotype-phenotype associations (GPA) or population Genome-Wide Association Study (pGWAS). The strategy is to find correlations between allele frequencies at a given locus for a set of populations and mean values for a given trait for the same populations. In a nutshell, BayPass infer this "environmental effect" through a locus-specific regression coefficient parameter (β). In BayPass, the significance of this parameter can be tested using different decision rules (see [54] for details). Here, we use a simple comparison of models with and without association (i.e. a model assuming $\beta \neq 0$ vs. $\beta = 0$) and quantify this support using Bayes Factors (BF). The most positive BF values correspond to SNPs with the highest support for the model with a significant environmental or phenotypic effect. In general, SNPs of great interest are those simultaneously exhibiting both allele frequency differences among populations (highest XtX values) and associations (highest BF values, **Fig. 21**).

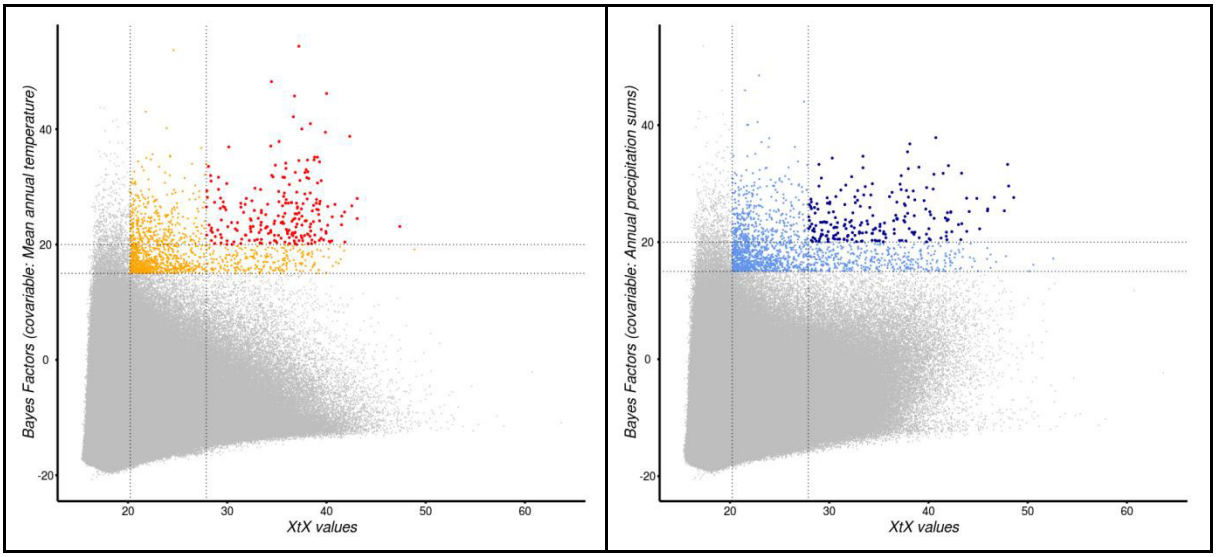


Fig. 21: *Whole-genome scan for genetic differentiation (XtX) and association (Bayes Factors, BF) with mean annual temperature (left) or precipitation sums (right) covariables and identification of SNPs of interests (orange or light blue, best candidates red and dark blue). A simple rule-of-thumb decision was used to identify the most strongly associated SNPs: BF=15 and BF=20. As an alternative, it is also possible to use the PODS to calibrate the BF metric, in the same way as for the XtX (see Leroy et al. [6]).*

Manhattan plots showing chromosome positions of the associated SNPs (**Fig. 22**) reveal clusters of associated SNPs in some genomic regions, particularly on

chromosomes 1, 9, 10 and 12. Such investigations can lead to the identification of important genes for local adaptation possible, for example here adaptations to cold/warm conditions or drought/waterlogging. It is however crucial to keep in mind the following statement when interpreting the results: correlation does not imply causation. GEA and GPA analyses can provide ecologically meaningful information but these analyses are also prone to over-interpretation and storytelling (e.g. [74]).

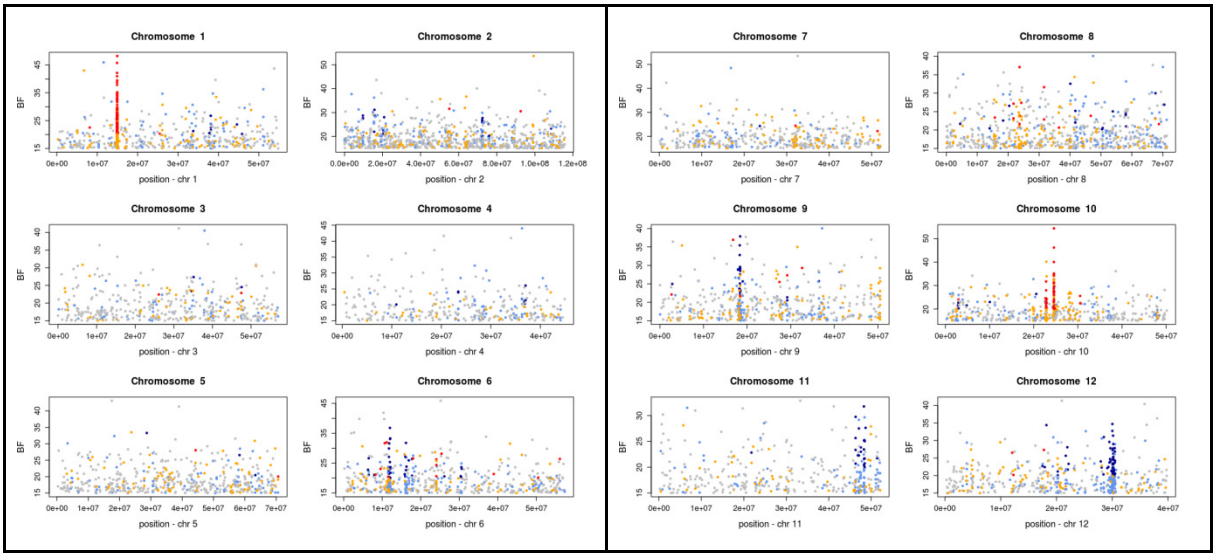


Fig. 22: *Manhattan plots showing the chromosome positions of the SNPs exhibiting elevated Bayes Factors (BF) as detected using BayPass. Significant SNPs in Fig. 20 are shown in colors. To facilitate readability, only SNPs with XtX > 15 and BF > 15 for either the mean annual temperature covariable or mean annual precipitation sums are shown.*

4. Notes

Note 1: Taking into account linkage disequilibrium

For several analyses (e.g. PCA, clustering methods) it is important to note that the linkage disequilibrium (LD), the non-random association of alleles within a genome between a given locus and its genomic neighborhood, is an important factor to control for. For species with a relatively limited extent of LD across the genome (in general native species with a high genetic diversity), this bias is expected to be limited, but can become substantial for some species, particularly domesticated ones. The use of SNP pruning methods (e.g. SNPrune [75]) is currently being increasingly used for that purpose. We recommend using these methods. Advices on how to use these methods are available on the github repository.

Note 2: Beyond TreeMix, demographic inferences

It is also important to note that TreeMix (#3.3.4) fits single admixture pulses assuming homogeneous gene flow along the genome. This assumption is likely to be

violated because migration is expected to be impeded at some genes maintaining genetic differences between hybridizing populations (e.g. [33] for empirical evidence). As a consequence, Treemix provides a good way to investigate potential migration events, but the exact direction of gene flow and the intensity of the migration edges should be interpreted with some caution. Some more advanced modeling approaches, albeit computationally intense, can decipher the evolutionary history of the investigated species with more confidence. These methods can explicitly account for heterogeneous migration rates (i.e. presence of barriers to gene flow). These methods provide considerably stronger statistical support for migration between populations, as well as temporal changes in effective population sizes, e.g. Approximate Bayesian Computation (ABC [33,76]) or dadi [32]. A growing number of empirical studies have used the former (*e.g.* [76-77]), the latter (*e.g.* [78-79]) or both methods (*e.g.* [80]).

Deciphering the evolutionary history of a given species is an important step, because demography can generate a substantial background noise weakening genome scan analyses [81]. To perform robust identification of variants under selection (or variants in close vicinity), one of the ongoing challenges is to better take into account the evolutionary history of the population. Extensive simulations under the inferred most-likely evolutionary scenario can provide an accurate distribution of the expected differences in allele frequencies (e.g. F_{ST}, XtX or similar), thereby allowing the identification of variants under selection among loci deviating from these demographic expectations. Some early attempts to explicitly taking into account the inferred demography to scan genome for selection have recently emerged (e.g. [61,78,82-83]). In future, we suspect the emergence of new methods inferring at once the most-likely demographic scenario and variants departing from neutrality assuming this scenario.

Acknowledgments

The analyses benefited from the Montpellier Bioinformatics Biodiversity (MBB) platform services, the genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul), the Bird Platform of the university of Nantes and Compute Canada (Graham servers). This work takes its source from a diverse range of research contributions and projects we achieved during the last 5 years. During this period, TL was supported by different postdoctoral fellowships from the French *Agence Nationale de la Recherche* (ANR, Genoak project, PI: Christophe Plomion, 11-BSV6-009-021 and BirdIslandGenomic, PI: Benoit Nabholz, ANR-14-CE02-0002), from the European Research council (ERC, Treepeace, PI: Antoine Kremr, Grant Agreement no. 339728) and from the University of Vienna, Austria (PI Christian Lexer). QR was supported by the government of Canada through Genome Canada, Genome British Columbia, and Genome Quebec and want to thanks Louis Bernatchez for the opportunity to develop various projects during his postdoctoral

research. We want to thank Jean-Marc Aury, Antoine Kremer & Christophe Plomion for providing access to the oak sequencing data. We also thank Philippe Vigouroux and Philippe Cubry for information concerning the African rice data, and Pierre-Alexandre Gagnaire and Nicolas Bierne for discussions on TreeMix.

References

1. Charlesworth B (2010) Molecular population genomics: a short history. Genetics Research 92:397–411. https://doi.org/10.1017/S0016672310000522
2. Wang W, Mauleon R, Hu Z, et al (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49. https://doi.org/10.1038/s41586-018-0063-9
3. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, 1001 Genomes Consortium (2016) 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell 166:481–491. https://doi.org/10.1016/j.cell.2016.05.063
4. Hartl DL, Clark AG (1998) Principles of population genetics
5. Cubry P, Tranchant-Dubreuil C, Thuillet A-C, et al (2018) The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes. Current Biology 28:2274-2282.e6. https://doi.org/10.1016/j.cub.2018.05.066
6. Leroy T, Louvet J-M, Lalanne C, et al (2019) Adaptive introgression as a driver of local adaptation to climate in European white oaks. bioRxiv 584847. https://doi.org/10.1101/584847
7. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997
8. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324
9. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
10. Makino T, Rubin C-J, Carneiro M, et al (2018) Elevated Proportions of Deleterious Genetic Variation in Domestic Animals and Plants. Genome Biology and Evolution 10:276–290. https://doi.org/10.1093/gbe/evy004
11. Meyer RS, Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. Nature Reviews Genetics 14:840
12. Falush D, Stephens M, Pritchard JK (2003) Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. Genetics 164:1567
13. Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Molecular Ecology Resources 9:1322–1332. https://doi.org/10.1111/j.1755-0998.2009.02591.x
14. Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. Genetics 155:945
15. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nature Genetics 40:646
16. Baird NA, Etter PD, Atwood TS, et al (2008) Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers. PLOS ONE 3:e3376. https://doi.org/10.1371/journal.pone.0003376
17. Durand E, Jay F, Gaggiotti OE, François O (2009) Spatial Inference of Admixture Proportions and Secondary Contact Zones. Molecular Biology and Evolution 26:1963–1973. https://doi.org/10.1093/molbev/msp106
18. Corander J, Marttinen P (2006) Bayesian identification of admixture events using multilocus molecular markers. Molecular Ecology 15:2833–2843. https://doi.org/10.1111/j.1365-294X.2006.02994.x
19. Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. Genetics 197:573. https://doi.org/10.1534/genetics.114.164350
20. Frichot E, François O (2015) LEA: An R package for landscape and ecological association studies. Methods in Ecology and Evolution 6:925–929. https://doi.org/10.1111/2041-210X.12382
21. Frichot E, Mathieu F, Trouillon T, et al (2014) Fast and Efficient Estimation of Individual Ancestry Coefficients. Genetics 196:973. https://doi.org/10.1534/genetics.113.160572
22. Caye K, Deist TM, Martins H, et al (2016) TESS3: fast inference of spatial population structure and genome scans for selection. Molecular Ecology Resources 16:540–548. https://doi.org/10.1111/1755-0998.12471
23. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289

24. Pont C, Leroy T, Seidel M, et al (2019) Tracing the ancestry of modern bread wheats. Nature Genetics 51:905–911. https://doi.org/10.1038/s41588-019-0393-z
25. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585
26. Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205. https://doi.org/10.1038/nrg2526
27. Sigwart J (2009) Coalescent Theory: An Introduction. Systematic Biology 58:162–165. https://doi.org/10.1093/schbul/syp004
28. Terhorst J, Kamm JA, Song YS (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet 49:303–309. https://doi.org/10.1038/ng.3748
29. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493
30. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nature Genetics 46:919
31. Excoffier L, Dupanloup I, Huerta-Sánchez E, et al (2013) Robust Demographic Inference from Genomic and SNP Data. PLOS Genetics 9:e1003905. https://doi.org/10.1371/journal.pgen.1003905
32. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. PLOS Genetics 5:e1000695. https://doi.org/10.1371/journal.pgen.1000695
33. Roux C, Fraïsse C, Romiguier J, et al (2016) Shedding Light on the Grey Zone of Speciation along a Continuum of Genomic Divergence. PLOS Biology 14:e2000234. https://doi.org/10.1371/journal.pbio.2000234
34. Akashi H, Osada N, Ohta T (2012) Weak Selection and Protein Evolution. Genetics 192:15. https://doi.org/10.1534/genetics.112.140178
35. Lu J, Tang T, Tang H, et al (2006) The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends in Genetics 22:126–131. https://doi.org/10.1016/j.tig.2006.01.004
36. Yang J, Mezmouk S, Baumgarten A, et al (2017) Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. PLOS Genetics 13:e1007019. https://doi.org/10.1371/journal.pgen.1007019
37. Liu Q, Zhou Y, Morrell PL, Gaut BS (2017) Deleterious Variants in Asian Rice and the Potential Cost of Domestication. Molecular Biology and Evolution 34:908–924. https://doi.org/10.1093/molbev/msw296
38. Ramu P, Esuma W, Kawuki R, et al (2017) Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nature Genetics 49:959
39. Zhou Y, Massonnet M, Sanjak JS, et al (2017) Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. Proc Natl Acad Sci USA 114:11715. https://doi.org/10.1073/pnas.1709257114
40. Stein JC, Yu Y, Copetti D, et al (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nature Genetics 50:285–296. https://doi.org/10.1038/s41588-018-0040-0
41. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP, et al (2016) Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci USA 113:152. https://doi.org/10.1073/pnas.1512501113
42. Ng PC, Henikoff S (2001) Predicting Deleterious Amino Acid Substitutions. Genome Research 11:863–874
43. Choi Y, Sims GE, Murphy S, et al (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS One 7:e46688–e46688. https://doi.org/10.1371/journal.pone.0046688
44. Peischl S, Excoffier L (2015) Expansion load: recessive mutations and the role of standing genetic variation. Molecular Ecology 24:2084–2094. https://doi.org/10.1111/mec.13154
45. Henn BM, Botigué LR, Bustamante CD, et al (2015) Estimating the mutation load in human genomes. Nature Reviews Genetics 16:333
46. Henn BM, Botigué LR, Peischl S, et al (2016) Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci USA 113:E440. https://doi.org/10.1073/pnas.1510805112
47. Simons YB, Turchin MC, Pritchard JK, Sella G (2014) The deleterious mutation load is insensitive to recent population history. Nat Genet 46:220–224. https://doi.org/10.1038/ng.2896
48. Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the selective neutrality of polymorphisms. Genetics 74:175
49. Bierne N, Roze D, Welch JJ (2013) Pervasive selection or is it…? why are FST outliers sometimes so frequent? Molecular Ecology 22:2061–2064. https://doi.org/10.1111/mec.12241
50. Bierne N, Welch J, Loire E, et al (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. Molecular Ecology

20:2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x
51. Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Molecular Ecology 24:1031–1046. https://doi.org/10.1111/mec.13100
52. Nei M, Maruyama T (1975) Lewontin-Krakauer test for neutral genes. Genetics 80:395
53. Robertson A (1975) Remarks on the Lewontin-Krakauer test. Genetics 80:396
54. Gautier M (2015) Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. Genetics 201:1555. https://doi.org/10.1534/genetics.115.181453
55. Whitlock MC, Lotterhos KE (2015) Reliable Detection of Loci Responsible for Local Adaptation: Inference of a Null Model through Trimming the Distribution of FST. The American Naturalist 186:S24–S36. https://doi.org/10.1086/682949
56. Luu K, Bazin E, Blum MGB (2017) pcadapt: an R package to perform genome scans for selection based on principal component analysis. Molecular Ecology Resources 17:67–77. https://doi.org/10.1111/1755-0998.12592
57. Abdellaoui A, Hottenga J-J, Knijff P de, et al (2013) Population structure, migration, and diversifying selection in the Netherlands. European Journal Of Human Genetics 21:1277
58. Jackson DA (1993) Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. Ecology 74:2204–2214. https://doi.org/10.2307/1939574
59. Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals — mining genome-wide polymorphism data without big funding. Nature Reviews Genetics 15:749
60. Gautier M, Foucaud J, Gharbi K, et al (2013) Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Molecular Ecology 22:3766–3779. https://doi.org/10.1111/mec.12360
61. Leroy T, Rougemont Q, Dupouey J-L, et al (2018) Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. bioRxiv. https://doi.org/10.1101/246637
62. Plomion C, Aury J-M, Amselem J, et al (2018) Oak genome reveals facets of long lifespan. Nature Plants 4:440–452. https://doi.org/10.1038/s41477-018-0172-3
63. de Vries S, Murat A, Bozzano M, et al (2015) Pan-European strategy for genetic conservation of forest trees and establishment of a core network of dynamic conservation units
64. Lindner MS, Kollock M, Zickmann F, Renard BY (2013) Analyzing genome coverage profiles with applications to quality control in metagenomics. Bioinformatics 29:1260–1267. https://doi.org/10.1093/bioinformatics/btt147
65. Kofler R, Orozco-terWengel P, De Maio N, et al (2011) PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. PLOS ONE 6:e15925. https://doi.org/10.1371/journal.pone.0015925
66. Pickrell JK, Pritchard JK (2012) Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. PLOS Genetics 8:e1002967. https://doi.org/10.1371/journal.pgen.1002967
67. Reich D, Thangaraj K, Patterson N, et al (2009) Reconstructing Indian population history. Nature 461:489
68. Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nature Genetics 39:1251
69. Kofler R, Pandey RV, Schlötterer C (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27:3435–3436. https://doi.org/10.1093/bioinformatics/btr589
70. Hivert V, Leblois R, Petit EJ, et al (2018) Measuring Genetic Differentiation from Pool-seq Data. Genetics 210:315. https://doi.org/10.1534/genetics.118.300900
71. Gautier M, Hivert V, Vitalis R poolfstat: Computing F-Statistics from Pool-Seq Data
72. Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. Genetics 185:1411–1423. https://doi.org/10.1534/genetics.110.114819
73. Günther T, Coop G (2013) Robust Identification of Local Adaptation from Allele Frequencies. Genetics 195:205. https://doi.org/10.1534/genetics.113.152462
74. Pavlidis P, Jensen JD, Stephan W, Stamatakis A (2012) A Critical Assessment of Storytelling: Gene Ontology Categories and the Importance of Validating Genomic Scans. Molecular Biology and Evolution 29:3237–3248. https://doi.org/10.1093/molbev/mss136
75. Calus MPL, Vandenplas J (2018) SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. Genetics Selection Evolution 50:34. https://doi.org/10.1186/s12711-018-0404-z
76. Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly

divergent Ciona intestinalis species. Mol Biol Evol 30:1574–1587
77. Fraïsse C, Roux C, Gagnaire P-A, et al (2018) The divergence history of European blue mussel species reconstructed from Approximate Bayesian Computation: the effects of sequencing techniques and sampling strategies. PeerJ 6:e5198. https://doi.org/10.7717/peerj.5198
78. Rougemont Q, Gagnaire P-A, Perrier C, et al (2017) Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. Molecular Ecology 26:142–162. https://doi.org/10.1111/mec.13664
79. Tine M, Kuhl H, Gagnaire P-A, et al (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat Commun 5:5770
80. Leroy T, Rougemont Q, Dupouey J-L, et al (2018) Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. bioRxiv. https://doi.org/10.1101/246637
81. Hermisson J (2009) Who believes in whole-genome scans for selection? Heredity 103:283–284
82. Fraïsse C, Roux C, Welch JJ, Bierne N (2014) Gene-Flow in a Mosaic Hybrid Zone: Is Local Introgression Adaptive? Genetics 197:939. https://doi.org/10.1534/genetics.114.161380
83. Le Moan A, Gagnaire P-A, Bonhomme F (2016) Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. Mol Ecol 25:3187–3202. https://doi.org/10.1111/mec.13627