Statistical Methods to Identify Local Adaptation in Continuous and Admixed Populations

Helena Martins

Communauté UNIVERSITÉ Grenoble Alpes

THÈSE

Pour obtenir le grade de

DOCTEUR DE LA COMMUNAUTÉ UNIVERSITÉ GRENOBLE ALPES

Spécialité : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement

Arrêté ministériel : 25 mai 2016

Présentée par

Helena MARTINS

Thèse dirigée par **Olivier FRANCOIS (EDISCE)**, professeur, G-INP
et codirigée par **Michael BLUM (EDISCE)**, CNRS
préparée au sein du **Laboratoire Techniques de L'Ingénierie Médicale et de la Complexité - Informatique, Mathématiques et Applications.**
dans **l'École Doctorale Ingénierie pour la santé la Cognition et l'Environnement**

Méthodes statistiques pour identifier l'adaptation locale dans les populations continues et mélangées

Statistical Methods to Identify Local Adaptation in Continuous and Admixed Populations

Thèse soutenue publiquement le **26 septembre 2018**,
devant le jury composé de :

Monsieur OLIVIER FRANÇOIS
PROFESSEUR, UNIVERSITE GRENOBLE ALPES, Directeur de thèse
Monsieur MATHIEU GAUTIER
CHARGE DE RECHERCHE, INRA MONTPELLIER, Rapporteur
Madame STEPHANIE MANEL
PROFESSEUR, CNRS DELEGATION LANGUEDOC-ROUSSILLON, Rapporteur
Madame LAURENCE DESPRES
PROFESSEUR, UNIVERSITE GRENOBLE ALPES, Président
Monsieur MICHAEL BLUM
DIRECTEUR DE RECHERCHE, CNRS DELEGATION ALPES, Co-directeur de thèse
Monsieur YVES VIGOUROUX
DIRECTEUR DE RECHERCHE, IRD FRANCE-SUD, Examinateur

Table of Contents

Acknowledgements

I would like to express my sincere gratitude to my supervisors Olivier François and Michael Blum for the continuous support of my PhD study and related research, for his patience, motivation, and immense knowledge. Their guidance helped me in all the time of study and writing of this thesis. I thank all the BCM team for the excellent atmosphere during these three and a half years.

I thank my reviewers Stephanie Mane and Mathieu Gautier, as well as members of the jury Laurence Desprès, Michael Blüm and Yves Vigouroux, who accepted to evaluate my work.

Special thanks to my parents João Luiz and Silvia who showed me the importance of a good education. They always encouraged me to dream big and to be patient because the dreams never get old. I am also grateful to my two brothers, Ednardo and João Vitor, and all other family members that supported me along the way.

Most importantly, I would like to thank my husband Luis, without whose love, encouragement and editing assistance, I would not have finished this thesis.

Résumé court

Titre: Méthodes Statistiques pour Identifier l'Adaptation Locale dans les Populations Continues et Métissées

La recherche des signatures génétiques de l'adaptation locale est d'un grand intérêt pour de nombreuses études en génétique des populations. Les approches pour détecter les loci sélectifs à partir de leur contexte génomique, se concentrent souvent sur les valeurs extrêmes de l'indice de fixation, F_{ST}, pour les loci. Cependant, le calcul de l'indice de fixation devient difficile lorsque la population est génétiquement continue, lorsque la définition des sous-populations est difficile, ou en présence d'individus métissés dans l'échantillon. Dans cette thèse, nous présentons une nouvelle méthode pour identifier des loci sous sélection basée sur une extension de la statistique F_{ST} à des échantillons comportant des individus métissés. Des scans génomiques pour la sélection ont été appliqués en utilisant notre nouvelle statistique F_{ST} pour la plante *A. thaliana*, et dans les données de génomique humaine de l'échantillon de référence `POPRES`. Les résultats ont montré l'utilité de notre méthode pour détecter les signaux de sélection naturelle. Considérant des méthodes statistiques pour identifier l'adaptation locale dans une population métissée, nous avons inclus des données spatiales pour calculer les coefficients d'ascendance et les fréquences d'allèles ancestrales. Des tests de sélection utilisant notre nouvelle mesure F_{ST} et le logiciel `tess3` ont fourni des preuves de signaux de sélection naturelle dans le génome de *A. thaliana* en Europe. Pour enrichir notre travail, nous avons recherché les effets du déséquilibre de liaison (DL) et des méthodes d'élagage de DL dans les analyses pour détecter la sélection. Nous avons réaffirmé l'importance de l'ajustement pour le DL dans les données génomiques. La présence de DL peut modifier les résultats de

l'analyse sur les données avec des individus métissés et il peut augmenter le taux de faux positifs. En conclusion, ce travail de thèse permet des scans génomiques en présence d'individus métissés, ouvrant des espaces pour des analyses biologiques variées.

Mot-clés: génétique des populations, scan génomique, populations mélangées, tests de différenciation des populations, déséquilibre de liaison, biostatistiques.

Abstract

Title: Statistical Methods to Identify Local Adaptation in Continuous and Admixed Populations

Finding genetic signatures of local adaptation is of great interest for many population genetic studies. Conventional approaches to sorting selective loci from their genomic background focus on the extreme values of the fixation index, F_{ST}, across loci. However, the computation of the fixation index becomes challenging when the population is genetically continuous, when predefining subpopulations is a difficult task, and in the presence of admixed individuals in the sample. In this thesis, we present a new method to identify loci under selection based on an extension of the F_{ST} statistic to samples with admixed individuals. Genome scans for selection applied using our new F_{ST} statistic on genetic data of the plant *A. thaliana* and, in human genomic data from the population reference sample, `POPRES`, showed the usefulness of our method to detect targets of natural selection. Considering statistical methods to identify local adaptation in admixed population, we included spatial data to compute ancestry coefficients and allele frequencies. Tests for selection using our new F_{ST} statistic and the `tess3` software provided evidence of signals of natural selection in shaping the genome-wide variation of the plant species *A. thaliana* in Europe. To extend our work, we investigated the effects of linkage disequilibrium and LD-pruning methods in genome scans for selection. We reiterated the importance of adjusting for LD in genomic data since it can change the results of analysis on data with admixed individuals and can increase the false discovery rate. In conclusion, this PhD work, makes possible the application of genome scans for selection in presence of admixed individuals and, open new spaces for numerous and varied biological analyses.

Key-words: population genetics, genome scans for selection, admixed populations, population differentiation tests, linkage disequilibrium, biostatistics.

Chapter 1

Introduction

1.1 Population Genetics

Population genetics examines the amount of genetic variation within populations and processes such as adaptation, speciation, and population structure (Okasha, 2006). At a specific gene or locus in the genome, individuals may carry different genetic variants that are called alleles. The different alleles carried by various individuals of the same species capture genetic variation within a species. Population genetics describes this genetic change, which could be the consequence of four different evolutionary processes: mutation, migration and divergence, genetic drift and natural selection (Hartl et al., 1997).

The first process is the mutational process. Mutations are permanent alterations of the nucleotide sequence in the genome of an organism and appear at a particular rate called the mutation rate. Considering that migration rate for most organisms is pretty low, the impact of brand new mutations on allele frequencies from one generation to the next is usually not substantial (Griffiths, 2002).

Another evolutionary process that affects genetic variation is concerned to migration and divergence between populations. Migration occurs when individuals occasionally cross these barriers becoming migrants. Divergence between populations occurs for instance in the presence of natural barriers, such as sea or mountain (Hey and Pinho, 2012). Because of divergence, sexual reproduction is more likely to occur between individuals that are on the same side of a natural barrier. While divergence between populations increases genetic differences between individuals, migration tends to make

individuals that are part of different populations more similar (Lenormand, 2002).

The third process is genetic drift, and it is a more abstract concept. It describes random fluctuations of allele frequency as a function of time. Genetic drift takes place when allele frequency increases and decreases by chance over time. Genetic drift is more pronounced in small populations, where alleles at low frequency face a higher chance of being lost. Genetic drift is enhanced after population bottlenecks, which are events that drastically decrease population size. Genetic drift can result in the loss of rare alleles and decrease genetic variation. Genetic drift can cause a new population to be genetically distinct from its original population, which has led to the hypothesis that genetic drift plays a role in the evolution of new species (Chen, Boeger, et al., 1994).

The last evolutionary process that affects genetic variation is natural selection. There are several forms of natural selection; positive natural selection promotes mutations and new alleles, while a negative or purifying selection happens when individuals carrying new variants have a lower fitness. Another type of natural selection is diversifying selection when extreme values of a quantitative trait are favoured. Selective pressures may vary according to the environment resulting in one allele selected in one population only or in different alleles selected in different populations (Savolainen et al., 2013). Variation of selective pressure can result in local adaptation where individuals are locally adapted to their environment (Kawecki and Ebert, 2004).

Genetic variation between individuals in a population is the result of those processes mentioned above. One of the goals of population genetics is to find the signals left by these four evolutionary processes on genetic variation. This thesis work is a part of this research effort. We propose a statistical methodology to find genomic variation involved in local adaptation, and we evaluate it using numerical simulations of different scenarios of local adaptation.

1.2 Local Adaptation

Local adaptation happens when individuals adapt to particularities of their local environment, and it is relevant to climate change, crop and animal production, and conservation of genetic resource (Savolainen et al., 2013). When local adaptation occurs, the survival and reproductive success of an individual (fitness) in a site must be better for a native individual than for an individual coming from another habitat, as shown in Figure 1.1 (Kawecki and Ebert, 2004).

In general, an individual will be adapted to its local habitat when one or more alleles increase its fitness, which determines an adaptive trait. An allele can be selected in a population due to different situations (Kliman et al., 2008). One situation is when the allele is already in the population and becomes adaptive after a change in the environment. Another option is when the allele is the result of a mutation or when it comes, by gene flow, from another population. Sexual reproduction plays an essential role in natural selection in both situations cited above. Through breeding, individuals from new generation must inherit adapted alleles. The identification of alleles responsible for adaptation is part crucial to the study of evolutionary processes in populations, and it can have numerous applications. In the following paragraphs, we will describe some real examples of local adaptation.

Peppered moth colouration in response to the industrial revolution Before the Industrial Revolution in England, the pale peppered moths were the majority, and black moths were rare between the species. The early decades of the Industrial Revolution pollution blackened the trees, turning the black moths less vulnerable to predators (Figure 1.2). This event gave the black variety more chance to survive and reproduce. Over time, the black peppered moths became far more numerous in urban areas than the pale variety. Peppered moth colouration remains a classic example of Darwin's natural selection (Majerus, 2009).

Antibiotic resistance in diseases Bacteria reproduce very fast and can evolve in a short period. The bacterium E. *coli*, for example, can have its DNA damaged

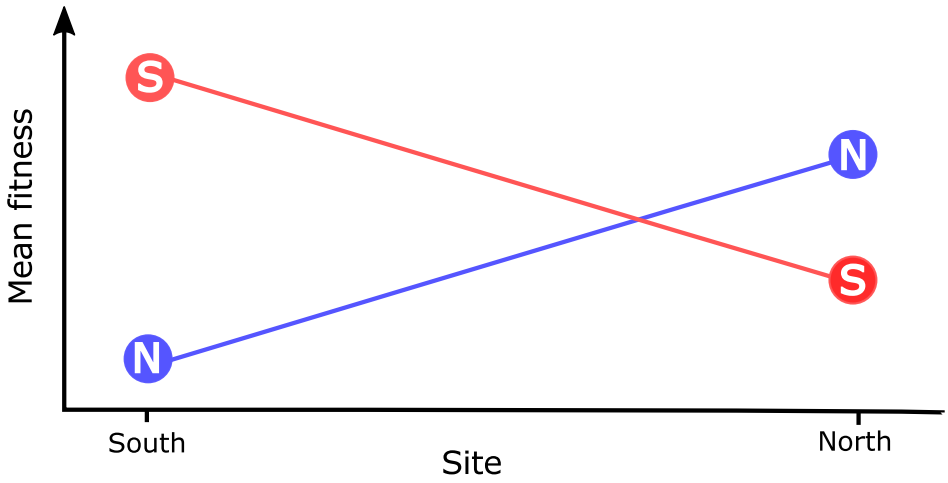


Figure 1.1 – **Fitness comparisons between populations from North (N) and South (S).** Each locally adapted population in its native site has higher fitness than any other population in the same site (Savolainen et al., 2013).

or changed during replication, which most of the time causes the death of the cell. Occasionally, these mutations can make the bacteria able to survive and even multiply in the presence of an antibiotic. The resistance from antibiotics is a severe public health problem; while it can be prevented by minimising unnecessary prescribing and overprescribing of antibiotics (Levy, 1998).

High-altitude adaptation in humans Humans are adapted to an environment where oxygen is abundant. High-altitude hypoxia, or decreased oxygen levels caused by low barometric pressure, challenges the ability of humans to live and reproduce. Despite that, natives of Tibet have been living at high altitudes for generations and exhibit unique circulatory, respiratory, and haematological systems. Researchers identified that Tibetans have special genes that allow the use of smaller quantity of oxygen efficiently, which enables their members to receive enough oxygen while exercising at high altitude (Wu and Kayser, 2006). For example, changes in the gene EPAS1, which controls haemoglobin production, have been positively selected in Tibetans. This mutation occurs in higher frequency in Tibetans than their Han neighbours and is correlated with decreased haemoglobin concentrations amongst Tibetans (Fig. 1.3). The low haemoglobin concentration at high altitude is considerate a significant genetic adaptation to high-altitude (Beall, Brittenham, et al., 1998).

Figure 1.2 – **Peppered moth local adaptation.** The picture shows a pale moth and a dark moth rest side-by-side on a soot-covered trunk near Birmingham (Kettlewell, 1956).

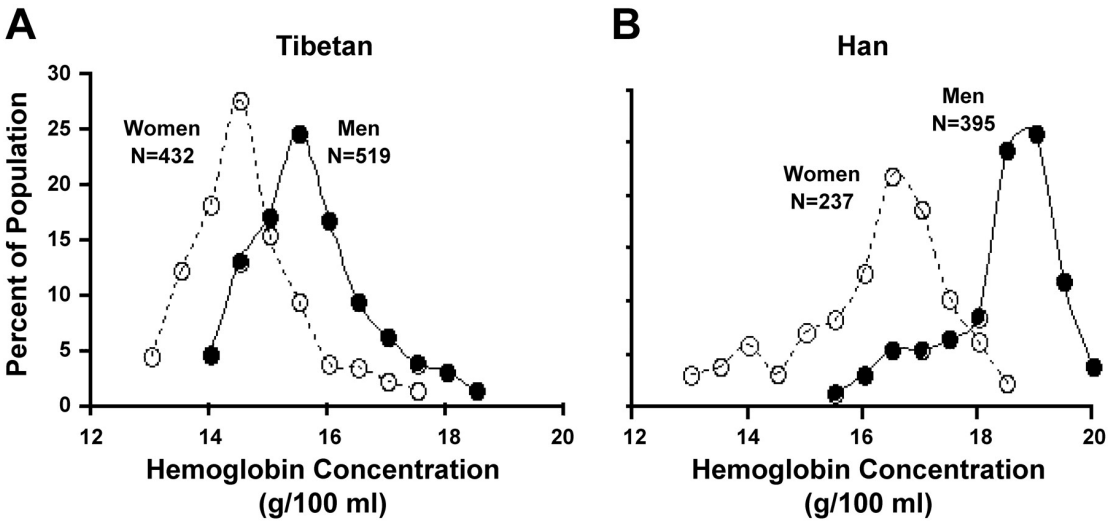


Figure 1.3 – **High-altitude adaptation in Tibain and Han.** Frequency distribution of [Hb] among adult (ages 16–60 yr) Tibetan (A) and Han (B) subjects at 4,525m. • And solid lines, males; ∘ and broken lines, females. Note that the [Hb] are skewed to the right in Tibetans and to the left in Han, showing that Han men and women had higher haemoglobin levels than Tibetan men and women (Wu and Kayser, 2006).

1.3 Population genetic data

1.3.1 Genetic markers

New technologies for DNA sequencing have contributed to increase the amount of dataset available to study genetic diversity within species. These datasets are known as genetic markers, which are DNA sequences with known physical locations on chromosomes and consist of polymorphisms whose allelic type can vary between individuals. Genetic markers can be used to identify regions of the genome responsible for a disease, for phenotypic variation or adaptive variation. Examples of genetic markers include single nucleotide polymorphisms (SNPs) and microsatellites. During this thesis, we consider SNP data.

Single nucleotide polymorphism (SNP)

Single nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide (A, T, G, or C) in the genome differs between individuals of a species (Figure 1.4). A single nucleotide has a variation classified as a SNP when more than 1% of a population does not hold the same nucleotide at a specific position in the DNA sequence. For example, at a specific base position in the human genome, the base A could be present in most individuals, however, in a small group of individuals, the position is occupied by base G (Figure 1.5). We say that there is a SNP at this specific base position, and the two possible nucleotide variations – A or G – are alleles for this base position.

Genetic Data Representation

The genetic data as SNPs can be represented using a matrix, called genotype matrix, which is composed only of 0, 1 and 2. To compute the genotype matrix, in this work we consider diploid individuals and reference allele as the less frequent allele. Considering a given locus and individual, we represent the genotype data as the number of times that the reference allele is observed. For example, assuming that the genotype of individuals can be AA, AG, GG; and considering that G is the reference allele, AA is coded as 0, AG as 1 and GG as 2. The choice of reference allele can

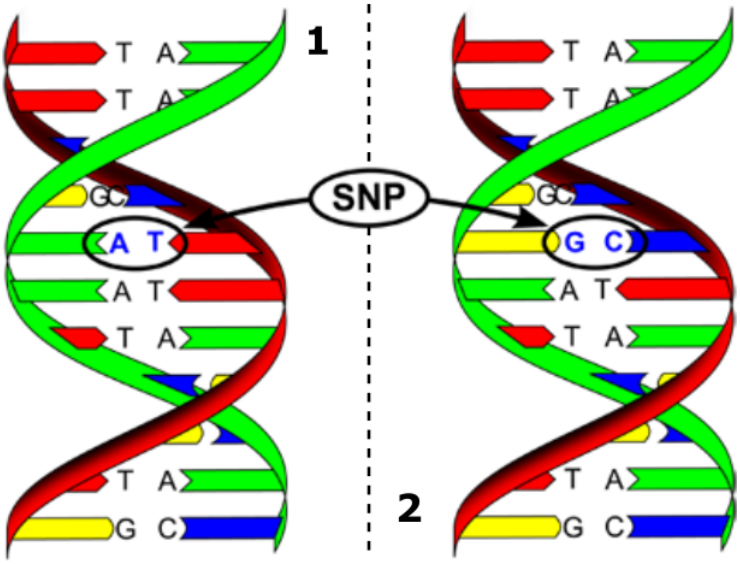


Figure 1.4 – **Example of a Single Nucleotide Polymorphism (SNP)**. DNA strand 1 differs from DNA strand 2 at a single base-pair location (a T/C polymorphism).

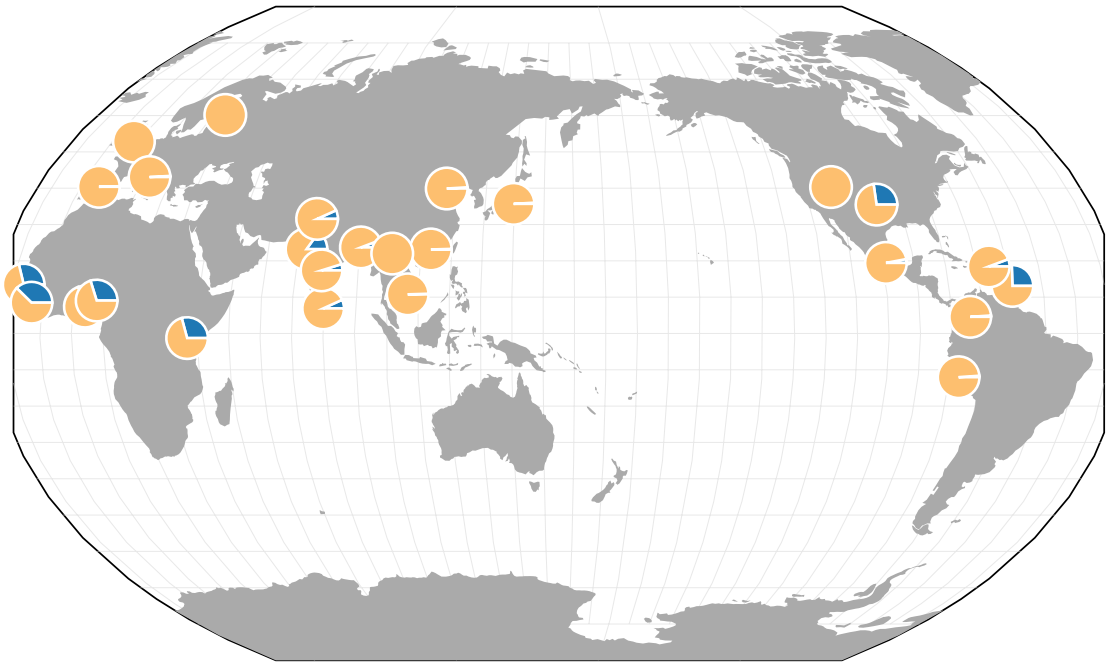


Figure 1.5 – **Allele frequency map for the SNP rs6602666.** Frequency proportions for the effect (G) and non-effect (A) alleles are represented respectably in dark blue and orange, respectively. Obtained from the Geography of Genetic Variants Browser (Majerus, 2009).

$$\mathbf{Y} = \begin{bmatrix} 0 & 1 & 2 & 2 & \cdots & \cdots & \cdots \\ 1 & 1 & 0 & 1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & \cdots & \cdots \\ 0 & 0 & 2 & 0 & \cdots & \cdots & \cdots \end{bmatrix}$$

Figure 1.6 – **Example of Genotype Matrix** In the matrix Y each element represents the number of times that the mutated allele is observed for a given individual at a given locus. Each line corresponds to the genotype of a single diploid individual.

be made arbitrarily without affecting the statistical methods based on the genotype matrix. In this study, we denote genotype matrices as the matrix Y of size $n \times p$, where n is the number of individuals and p is the number of loci (Figure 1.6).

1.3.2 Reference data sets

In the next paragraphs, we will talk briefly about two reference data sets in population genetics: European Lines of the plant species *Arabidopsis thaliana* and human genomic data from the population reference sample, POPRES. Both of the data were used for applications during this thesis.

Arabidopsis thaliana

Arabidopsis thaliana is the first plant to have a complete genome sequenced and published. Arabidopsis is a member of the mustard (Brassicaceae) family, which includes cultivated species such as cabbage and radish. *A. thaliana* has key features for basic research in genetics and molecular biology, mostly because of its small genome, 135 megabase pairs (Mbp) and a haploid chromosome number of 5, and because it can be cultivated in a controlled environment. The complete genome sequence of *Arabidopsis thaliana* was first published by the Arabidopsis Genome Initiative in 2000. During our work, we investigated genomic data from 120 European lines of *A. thaliana* genotyped for 216k SNPs, with a density of one SNP per 500 bp

(Atwell et al., 2010).

POPRES: Population Reference Sample

Advances in technology and science resulted from the Human Genome and HapMap projects, have made large-scale genomic data available. This data can be used to identify genetic factors that contribute to variation in disease risk. To facilitate exploratory genetic research, Nelson et al. assembled a DNA resource from a large number of subjects participating in multiple studies throughout the world. The POPRES project includes nearly 6,000 subjects of African American, East Asian, South Asian, Mexican, and European origin. During this thesis we consider 1385 European individuals from the POPRES dataset (447k SNPs in 22 chromosomes) (Nelson et al., 2008).

1.4 Genome Scans

Typically, a selected allele predominance increase or decrease within one or several but not all populations. Therefore, the observation of a high allele frequency in one population relative to others suggests that this allele has been positively selected.

Genome scans methods are used to screen genome-wide patterns of DNA polymorphism and to detect signatures of positive selection. These methods can make use of a one-dimensional statistic to test if populations have allelic frequencies that are significantly different from each other. The fixation index, F_{ST}, is the most common statistic used to scan the genome and can be computed for each marker. This statistic is associated with the variance in the allele frequency between populations and to the similarity among individuals inside populations (Holsinger and Weir, 2009).

The standard definition F_{ST} is based on Wright's studies, and is related to the variance in allele frequency that is explained by population structure (Wright, 1949). Therefore, F_{ST} value can be calculated using analysis of variance (ANOVA) (Holsinger and Weir, 2009). Lewontin and Krakauer (1973) approximated F_{ST} distribution with a chi-squared distribution to perform genome-wide scans. They pioneered the theory of statistical tests to detect selection.

Unfortunately, scan for selection based on F_{ST} can generate large numbers of false-positive test (Duforet-Frebourg et al., 2015). To reduce the rate of false positives, Bonhomme et *al.* 2010 proposed another chi-square test statistic based on allele frequency. The program `Bayescan` introduced another statistic method based on a Bayesian estimation of F_{ST} (Foll and Gaggiotti, 2008). Another software, `pcadapt`, implements a test for selection, which is based on principal component analysis (Luu et al., 2017a).

In this section, we will describe the island model, demographic simulations model commonly used in population genetics, and explain in more details the statistical methods mentioned above.

1.4.1 Demographic Simulation Model

Island Model

The island model has been proposed by Wright and it assumes that n populations are at demographic equilibrium and exchange a certain proportion of migrants at each generation (Figure 1.7). In this case, an allele that appears in one population through mutation can potentially be dispersed to any other population in a single generation, the probability of which is determined by the migration rate. As all populations are sharing migrants, they will converge on a unified allele frequency, defined by the global average allele frequency, $\bar{p}$. In this point, we say that the populations are in equilibrium (Slatkin and Voelm, 1991). The key parameter in the island model is the migration rate that measures the intensity of migration between populations. High migration rates tend to homogenise genetic variation between populations while low migration rates cause a more significant differentiation (Landguth et al., 2010).

In this PhD thesis, we considered Wright's 2-Island Model to simulated population genetics data. We used the computer program `ms` to perform coalescent simulations of neutral and selected SNP loci (Hudson, 2002). The justification for the use of Wright's models to simulate selection is that there is an overall migration rate for neutral markers, and a smaller migration rate that reflects action of selection, which increases population differentiation. (Bazin et al., 2010).

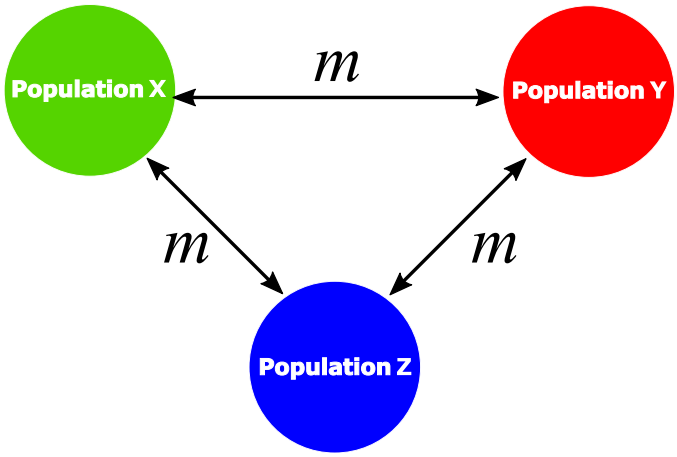


Figure 1.7 – **An n-island model with three populations** In the n-island model each population has its own allele frequencies, with a constant and symmetric migration rate (m).

1.4.2 Fixation Index

In the 1950's, Sewall Wright introduced F-statistics as a tool to describe genetic diversity in and with populations. Wright demonstrated that the amount of genetic differentiation among populations is related to the rates of migration, mutation and genetic drift. For example, large populations with high migration rate tend to show little differentiation, on another hand, small populations with low migration rate tend to be highly differentiated. The Fixation Index (F_{ST}), one of Wright's F-statistics, is a convenient measure of differentiation. Estimates of F_{ST} can identify regions of the genome that have been the target of selection, and comparisons of F_{ST} from different parts of the genome can provide insights into the demographic history of populations. For these reasons, F_{ST} and related statistics, have a central role in population and evolutionary genetics (Holsinger and Weir, 2009).

To obtain the F_{ST} value for a specific two-alleles locus, we consider N as the number of populations, p_i as the allelic frequency of the two alleles (reference allele) in the population i and $\bar{p}$ as the frequency of the reference allele across all populations, F_{ST} can be defined as follow (Wright, 1949)

$$F_{ST} = \frac{\frac{1}{N-1}\sum_{i=1}^{N}(p_i - \bar{p})^2}{\bar{p}(1-\bar{p})} \tag{1.1}$$

where $\bar{p} = \frac{1}{N-1} \sum_{i=1}^{N} p_i$.

Considering equation 1.1, the F_{ST} can also be related to the genetic variance due to the population structure. A classical definition for F_{ST}, that corresponds to the proportion of the genetic variation in sampled allele frequency and is defined as

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2} \tag{1.2}$$

where σ_T^2 is the variance of the allelic state in the total population, and σ_S^2 is the variance in the frequency of the allele between different subpopulations (Weir, 1996). Therefore, F_{ST} value can be calculated using analysis of variance (ANOVA) of allele frequencies.

Genome scans using F_{ST}

Genome scans that look for loci with an atypical F_{ST} value are a standard method to identify genome regions that are under local adaptation. Consider the example of local adaptation in Tibetan individuals mentioned in the Section 1.2. Tibetans are adapted to the high altitude environment. To detect markers that are involved in this process, it is possible to consider a second subpopulation that is not adapted to these conditions, and scan the genome looking for strong F_{ST} values. These strong values of F_{ST} will indicate, between Tibetans and the non adapted population, a considerable genetic differentiation when compared to the rest of the genome. In 2010, Beall et *al.*, presented a scan comparing the Tibetans and population of the Hans, China. The Manhattan plot of Beall, Cavalleri, et al. (2010), reveals signals of selection on the chromosome 2, in a region containing the gene EPAS1, known for control of haemoglobin production (Figure 1.8).

1.4.3 Lewontin-Krakauer's Test

Using the F_{ST} statistic of equation 1.1, Lewontin and Krakauer proposed a test to identify if a genetic marker is an outlier and if it is possibly involved in biological adaptation. The principle is to scan the genome computing F_{ST} values at each genetic

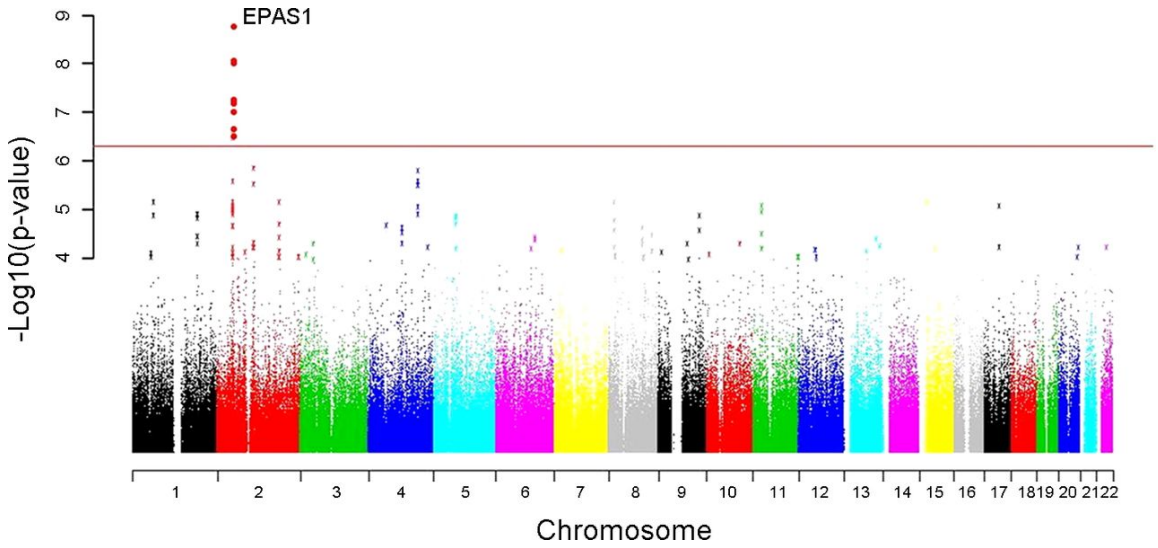


Figure 1.8 – **A genome-wide scan of allelic differentiation between population samples of Tibetans and Han Chinese.** The vertical axis shows the negative log of site-specific p-values for allele frequency differences between the Tibetan and Han Chinese population samples (low p-values denote allele frequency differences that are too large to explain by genetic drift). The horizontal axis of the graph shows the genomic positions of each assayed nucleotide site, arranged by chromosome number. The red line indicates the threshold for genome-wide statistical significance ($P = 5 \times 10^{-7}$).

marker available. Using this method, it is possible to detect the markers that have "strong" F_{ST}. The high value of F_{ST} for a given marker can be interpreted as a signal resulting from a selection pressure in one of the considered subpopulations. The idea of Lewontin and Krakauer is that in the absence of selection, the evolutionary forces, drift, migrations or mutations have a uniformly distributed effect on the genome. The variations caused by these forces are called the neutral structure (Lewontin and Krakauer, 1973). Therefore, all neutral markers should have F_{ST} values following a neutral distribution. Markers under selection can have F_{ST} values that depart from the neutral distribution, by having outlier values of F_{ST}.

Considering N the number of the populations, Lewontin and Krakauer (1973) proposed for the F_{ST} value a χ^2 test, denoted T_{LK} and defined as follow:

$$T_{LK} = \frac{N-1}{\bar{F}_{ST}} F_{ST} \tag{1.3}$$

Under the assumption that allelic frequencies are distributed according to a normal or binomial distribution, T_{LK} follows a χ^2 law with $N-1$ degrees of freedom. In fact, considering F_{ST} defined as the equation 1.1, we have (Luu 2017, PHd Manuscript)

$$(N-1)F_{ST} = \frac{1}{\bar{p}(1-\bar{p})} \sum_{i=1}^{N} (p_i - \bar{p})^2 = \left(\frac{p - \bar{p}}{\sqrt{\bar{p}(1-\bar{p})}} \right) \left(\frac{p - \bar{p}}{\sqrt{\bar{p}(1-\bar{p})}} \right)^T . \tag{1.4}$$

The degree of freedom is often estimated by the number of population samples minus 1. However, other methods to choose the number of freedom have been proposed. Lotterhos and Whitlock (2015) proposed to estimate the number of degree of freedom based on a maximum-likelihood principle, leading to values smaller than the actual number of populations. Duforet-Frebourg et al. (2015) estimate the degree of freedom of their tests by the number of principal components. As Caye et al. (2015), during this thesis, we estimate the degree of freedom of our test by the number of genetic clusters inferred from the data.

1.4.4 F_{LK} Model

An issue with genome scans based on F-statistics as Lewontin and Krakauer (1973), is that they can generate a high rate of false positives for both biological and statistical reasons (Bierne et al., 2013b). To minimise false positive rate and to increase the power of genome scans for selection, F_{LK} test compares patterns of differences in allele frequencies among populations to the values expected under a scenario of neutral evolution (Bonhomme et al., 2010). To test selective neutrality, F_{LK} reconstructs a topology modelling population divergence with a hierarchical structure.

To reflect the hierarchical structure, Bonhomme et al. 2010 use a matrix

$$\mathcal{F} = (f_{ij})_{1 \leq i,j \leq N} \in \mathcal{M}N(\mathbb{R})$$

,

where fij is the probability that an individual from the population i and an individual from the population j have inherited this allele from the same common ancestor.

The F_{LK} statistic is similar to the one in the equation 1.4 and can be defined as

$$F_{LK} = \left(\frac{p - \hat{p}_0}{\sqrt{\hat{p}_0(1 - \hat{p}_0)}} \right) \mathcal{F}^{-1} \left(\frac{p - \hat{p}_0}{\sqrt{\hat{p}_0(1 - \hat{p}_0)}} \right)^T \tag{1.5}$$

where $\hat{p}_0$ is the estimator of p_0, the frequency of the allele in ancestral population. The method F_{LK} is implemented by the software `FLK` (Bonhomme et al., 2010) and more recently by the software `hapFLK` (Fariello et al., 2013).

1.4.5 Bayesian Model

In 2004, Beaumont and Balding developed another statistic method based on F_{ST} in the context of the Island Model with an infinite number of islands (Figure 1.7).

Beaumont and Balding defined a parameter F_{ij} for the marker i, from a subpopulation j, estimated using the likelihood of a multinomial-Dirichlet model (Beaumont and Balding, 2004). To detect if a marker is involved in local adaptation, they propose the following model

$$\log\left(\frac{F_{ST}}{1-F_{ST}}\right) = \alpha_i + \beta_j, \tag{1.6}$$

where α_i is a parameter specific to marker i (e.g. mutation rate) and β_j measures the amount of drift in population j. Neutral markers are assumed to have α_i values equal to zero. In the case of atypical values for α_i, Beaumont and Balding proposed two interpretations. A strong positive value would indicate a selection pressure related to local adaptation. In the case of a strongly negative value, it would indicate homogeneous allelic frequency within populations possibly because of balancing selection. The Bayesian model is implemented in the software `Bayescan` (Foll and Gaggiotti, 2008).

1.4.6 Principal Components Analysis

Principal component analysis (PCA) is an approach used to highlight of multivariate data. It is often used to make data accessible to explore and visualise. PCA is mathematically characterised as an orthogonal linear transformation that converts the data to a different coordinate system. In this way, the highest variance by some projection of the data appears to lie on the first coordinate (denominated the first principal component), the second largest variance on the second coordinate, and so on (Jolliffe, 1986) (Figure 1.9).

Population genetics often applies PCA as a visualisation tool to assess population structure. Recently, it has also been used for genome scans (Duforet-Frebourg et al., 2015; Galinsky et al., 2016).

The idea to use PCA in a genomic scan for selection is based on the fact that the fixation index, F_{ST}, can be seen as the proportion of variance explained by the principal components (Duforet-Frebourg et al., 2015). The correlation between genetic variants and principal components enables the detection of markers involved in local adaptation without the need to define populations a priori.

To detect outliers Luu et al. (2017a) consider SNPs that are excessively related with population structure as measured by principal components. Consider Y $(n \times p)$ the genotype matrix where n is the number of individuals, and p is the number of

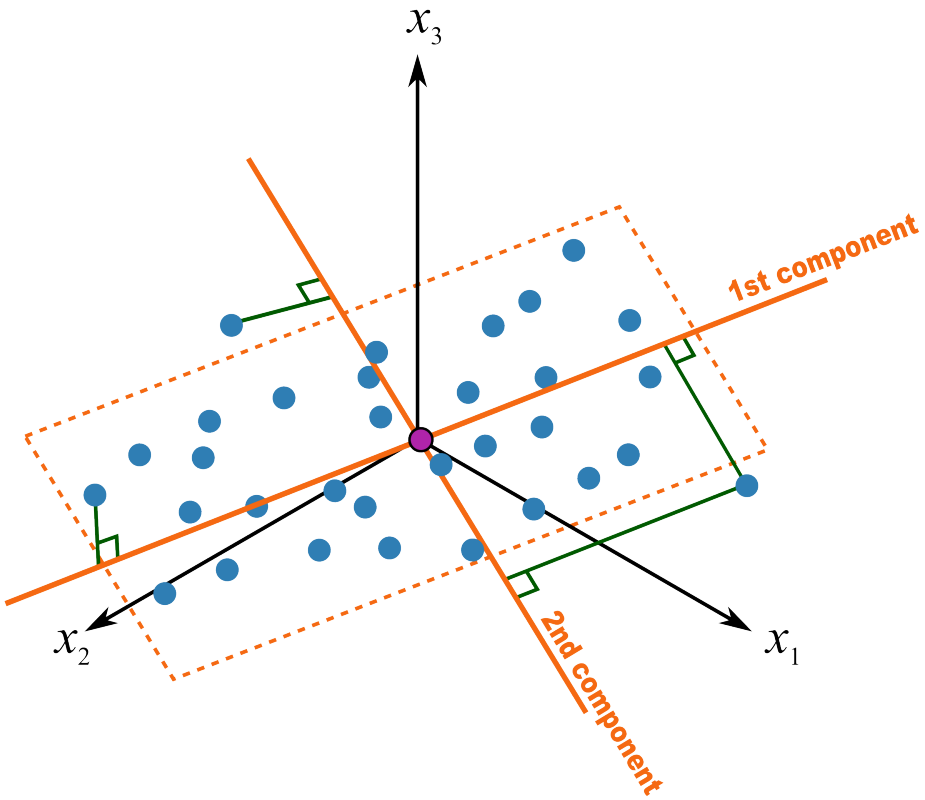


Figure 1.9 – **Principal Component Analysis of Distributed Data.** By definition, the 1st component line corresponds to the projection axis maximising the variance. The 2nd component line is deduced from the first one thanks to the orthogonality constraint and the fact that there are just two variables.

loci. The truncated singular value decomposition (SVD) that approximates the data matrix Y by a matrix of smaller rank is

$$Y \approx U\Sigma V^T, \tag{1.7}$$

where U is a $(n \times K)$ orthonormal matrix containing the principal components, V is a $(p \times K)$ orthonormal matrix, Σ is a diagonal $(K \times K)$ matrix and K corresponds to the rank of the approximation. Regressing each of the p SNPs by the K principal components $U_1, \ldots, U_K$

$$G_j = \sum_{k=1}^{K} \beta_{jk} U_k + \epsilon_j, ; j = 1, \ldots, p, \tag{1.8}$$

where β_{jk} is the regression coefficient corresponding to the j-th SNP regressed by the k-th principal component, and ϵ_j is the residuals vector. To summarise the result of the regression analysis for the j-th SNP, Luu et al. return a vector of z-scores $z_j = (z_{j1}, \ldots, z_{jK})$ where z_{jk} corresponds to the z-score obtained when regressing the j-th SNP by the k-th principal component.

The next step is to look for outliers based on the vector of z-scores. Luu et al. consider a classical approach in multivariate analysis for outlier detection. The test statistic is a robust Mahalanobis distance D defined as

$$D_j^2 = (z_j - \bar{z})^T \Sigma^{-1} (z_j - \bar{z}) \tag{1.9}$$

where Σ is the $(K \times K)$ covariance matrix of the z-scores and $\bar{z}$ is the vector of the K z-score means (Maronna and Zamar, 2002). When $K > 1$, the covariance matrix Σ is estimated with the orthogonalized Gnanadesikan–Kettenring method that is a robust estimate of the covariance able to handle large-scale data (Maronna and Zamar, 2002). When $K = 1$, the variance is estimated with another robust estimate.

1.5 Controlling false discoveries in genome scans for selection

Yoav Benjamini and Hochberg (1995) formally described the false discovery rate (FDR) theory (Benjamini-Hochberg approach). Considering a list of SNPs where the hypothesis of selective neutrality is rejected, the FDR is defined as the proportion of false discoveries among the positive tests. Control of the FDR ensures that the expected value of the FDR is lower than a pre-specified level. Therefore, if we define α as the expected FDR, candidate lists of loci are expected to contain less than a proportion α of false positives. In genome scans, FDR control methods are employed, especially in organisms where there are fewer genetic markers and no much information about the species evolutionary history and population structure. The challenge of FDR control algorithms is to minimise the number of false positives without being overly conservative and miss essential associations. In this section, we will describe the FDR control assumptions and talk about a unified testing framework for genome scans for selection.

1.5.1 FDR control algorithms

The underlying principle of FDR control algorithms relies on the fact that significance values (P-values) corresponding to truly null hypotheses, i.e., selectively neutral loci, are uniformly distributed over the interval (0,1). To see why the uniform distribution assumption is critical here, let us recall that the FDR is the expected value of the ratio F/S where F is the number of false positive tests, and S is the number of significant (positive) tests (Storey and Tibshirani, 2003). Let L_0 be the total number of truly null hypotheses. To provide control of the FDR at level, Benjamini and Hochberg (1995) proposed to sort the set of P-values and considered the largest value k such that $P_{(k)} \leq k\alpha/L$ (L is the total number of tests). The list of discoveries included all tested items with P-values lower than $P_{(k)}$. To compute the expected value of the ratio F/S, let us assume that the random value S is equal to $S = k$. According to the uniform distribution, the expected number of times a truly null hypothesis is rejected is equal to $E[F|S = k] = L_0 k\alpha/L$ in the Benjamini-Hochberg

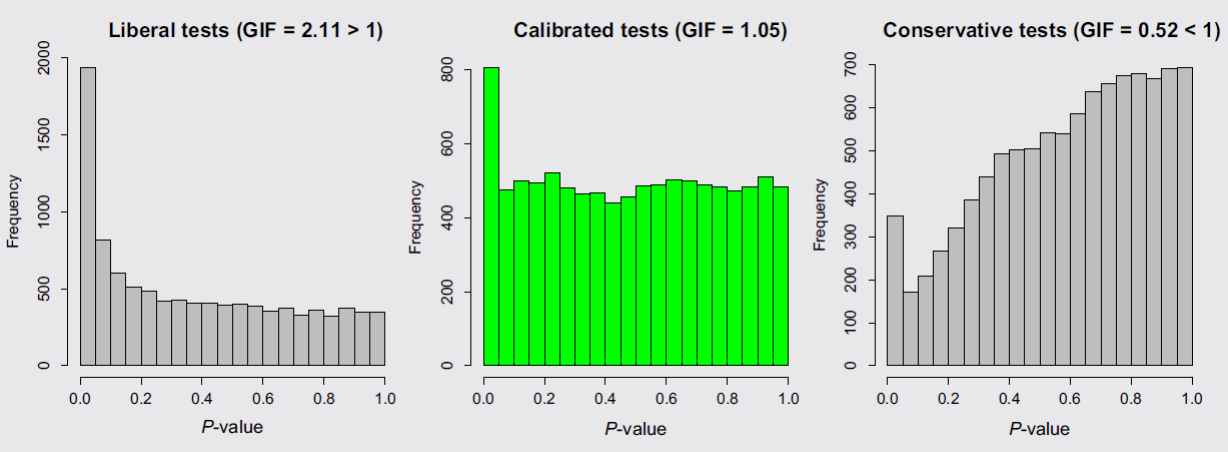


Figure 1.10 – **Histograms of test significance values (P-values) before the application of FDR control algorithms (artificial data).** GIF is the genomic inflation factor for each data set.

algorithm. Assuming $k \geq 1$, we have

$$E[F/S|S = k] = E[F/k|S = k] = L_0/L \times \alpha \leq \alpha.$$

Under these expectations, we obtain an FDR that is under control. To check that the uniform distribution assumption is correct, standard graphical approaches can be used. These methods display histograms of test P-values as in Figure 1.10.

1.5.2 A unified testing framework for genome scans for selection

Chi-squared distributions and genomic control

A statistical framework for genome scans approaches is based on the use of the chi-squared distribution for rejecting the null hypothesis of selective neutrality at a given locus. The ubiquity of the chi-squared distribution enables a unified treatment of test calibration and FDR control in genome scans for selection, which can be achieved by applying techniques developed for the analysis of GWAS and genome-wide gene expression analysis. The procedure to control FDR consists of modifying the null hypothesis to match the levels of neutral genetic background variation observed in the dataset. This procedure is sometimes called genomic control in GWAS, and empirical null hypothesis testing in studies of differential gene expression (Devlin and Roeder,

1999; Efron, 2007).

Genomic control relies on the introduction of inflation factors. Inflation factors are constant values, λ, that are used to rescale the test statistic to limit inflation due to population structure and confounding factors. The goal of the rescaling method is to define a modified test statistic leading to a flat histogram for the significance values. The test statistics will be designated as squared z-scores in this manuscript. For chi-squared tests, the rescaled statistic is z_l^2/λ where z_l is the score computed at locus l, and the degree of freedom of the test is left unaltered. This procedure has been called an empirical null-hypothesis testing approach by statisticians because it modifies the base-line null hypothesis, $H_0 : z_l^2 = 1$, and replaces it by a new null hypothesis, $H_0 : z_l^2 = \lambda$, in which λ is estimated from the data. Following GWAS approaches, an estimate of λ is commonly obtained after computing the genomic inflation factor, defined by the median of the squared z-scores divided by the median of a chi-squared distribution with $d-1$ degrees of freedom, d number of subpopulations. (Devlin and Roeder, 1999).

1.6 Motivations

Genome scans for selection identify loci that demonstrate significantly higher or lower among-population genetic differentiation than expected under neutrality (outliers). Identification of loci under selection is a crucial step in understanding the evolutionary process because those loci are responsible for the genetic variations that affect fitness in different environments.

During the last years many approaches focused on finding potentially adaptive character have been developed. Typically, these approaches focus on examining the variation in allele frequencies between populations. Those methods consider a significant number of single nucleotide polymorphisms (SNPs) and loci that have a target of selection can be identified as outliers in the upper tail of the empirical distribution of F_{ST} (Lewontin and Krakauer, 1973; Beaumont and Nichols, 1996; Akey et al., 2002; Weir, Cardon, et al., 2005). Related to statistical analysis of variance, F_{ST}

is the proportion of the total genetic variance contained in a subpopulation relative to the total genetic variance.

An important characteristic of the methods based on F_{ST} is that they require individuals to be assigned to a predefined populations. When the background levels of F_{ST} are weak and when populations are genetically homogeneous (Waples and Gaggiotti, 2006) or the samples contains admixed individuals, defining subpopulations during the outlier tests may be a challenge (Pritchard et al., 2000). Considering this situation, new methods of genome scans for selection that can handle admixed and in continuous populations must be developed. In parallel of those genome scans, spatial data can be used to provide more clues for selective forces in the real landscape. Also, it can complement and support quality of the final set of loci identified as potentially under selection (Feng et al., 2015).

Another situation that can interfere in the identification of loci under selection is the presence of linkage disequilibrium in genetic data sets. Although the LD is widely used to provide insight into the evolutionary history and for mapping genes in humans and other species, it remains a confounding factor in genome-wide association studies. Considering this problem, analyse possible effects of linkage disequilibrium is important before genome scan for selection (Laurie et al., 2010).

In this PhD manuscript we proposed a new statistic for genome scans for selection. To address limitation of F_{ST} methods, the statistic does not require predefined populations. Using simulations we investigate statistical properties especially in presence of LD.

1.7 Main Results

In this thesis, we present a definition of F_{ST}, which can be applied when there is continuous population structure and admixed individuals. This statistic requires computation of ancestry coefficients. More specifically, we used factor models to estimate F_{ST}, and we compared our neutrality tests with those derived from a principal component analysis approach. The performances of the tests were illustrated using simulated data and by re-analysing genomic data from European lines of the plant

species *Arabidopsis thaliana* and human genomic data from the population reference sample, `POPRES`. In addition, we compared our results for Europeans from the POPRES data sets with the genome-wide patterns of selection in 230 ancient Eurasians. This work is presented in Chapter 2 of this manuscript and corresponds to the paper Martins et al. (2016).

Considering our goal of exploring statistical methods to identify local adaptation in admixed population, we included spatial data to compute ancestry coefficients and allele frequencies. To perform computations, we used the software `tess3`, a spatial ancestry estimation program. Genome scans for selection were conducted using our statistic and `tess3` in simulated data and among European lines of *A. thaliana*. This work corresponds to the paper Caye et al. (2015) and is presented in Chapter 3 of this manuscript.

The last part of this thesis, Chapter 4, presents our investigation of the effects of linkage disequilibrium and LD-pruning methods in genome scans for selection. We conducted simulations using the statistic presented in Chapter 2. Intensity of LD varies in the simulated data as well as the number of admixed individuals. Accounting for the impact of linkage disequilibrium in our data, we applied an LD-pruning method using the toolset `PLINK`. We compare our new statistic for selection with other methods for genome scans.

Chapter 2

A new method to identify loci under selection based on an extension of the F_{ST} statistic to samples with admixed individuals

2.1 Abstract

Finding genetic signatures of local adaptation is of great interest for many population genetic studies. Common approaches to sorting selective loci from their genomic background focus on the extreme values of the fixation index, F_{ST}, across loci. However, the computation of the fixation index becomes challenging when the population is genetically continuous, when predefining subpopulations is a difficult task, and in the presence of admixed individuals in the sample. In this study, we present a new method to identify loci under selection based on an extension of the F_{ST} statistic to samples with admixed individuals. In our approach, F_{ST} values are computed from the ancestry coefficients obtained with ancestry estimation programs. More specifically, we used factor models to estimate F_{ST}, and we compared our neutrality tests with those derived from a principal component analysis approach. The performances of the tests were illustrated using simulated data, and by re-analyzing genomic data from European lines of the plant species *Arabidopsis thaliana* and human genomic data from the population reference sample, POPRES.

2.2 Introduction

Natural selection, the process by which organisms that are best adapted to their environment have an increased contribution of genetic variants to future generations, is the driving force of evolution (Darwin, 1859). Identifying genomic regions that have been the targets of natural selection is one of the most important challenge in modern population genetics (Vitti et al., 2013). To this aim, examining the variation in allele frequencies between populations is a frequently applied strategy (Cavalli-Sforza, 1966). More specifically, by sampling a large number of single nucleotide polymorphisms (SNPs) throughout the genome, loci that have been affected by diversifying selection can be identified as outliers in the upper tail of the empirical distribution of F_{ST} (Lewontin and Krakauer, 1973; Beaumont and Nichols, 1996; Akey et al., 2002; Weir, Cardon, et al., 2005). For selectively neutral SNPs, F_{ST} is determined by migration and genetic drift, which affect all SNPs across the genome in a similar way. In contrast, natural selection has locus-specific effects that can cause deviations in F_{ST} values at selected SNPs and at linked loci.

Outlier tests based on the empirical distribution of F_{ST} across the genome requires that the sample is subdivided into K subsamples, each of them corresponding to a distinct genetic group. For outlier tests, defining subpopulations may be a difficult task, especially when the background levels of F_{ST} are weak and when populations are genetically homogeneous (Waples and Gaggiotti, 2006). For example, Europe is genetically homogeneous for human genomes, and it is characterized by gradual variation in allele frequencies from the south to the north of the continent (Lao et al., 2008), in which genetic proximity mimics geographic proximity (Novembre et al., 2008). Studying evolution in the field, most ecological studies use individual-based sampling along geographic transects without using prior knowledge of populations (Manel, Schwartz, et al., 2003; Schoville et al., 2012). For example, the 1001 genomes project for the plant species *Arabidopsis thaliana* used a strategy in which individual ecotypes were sampled with a large geographic coverage of the native and naturalized ranges (Horton et al., 2012; Weigel and Mott, 2009). One last difficulty with F_{ST} tests arises from the presence of individuals with multiple ancestries (admixture), for which the genome

exhibits a mosaic of fragments originating from different ancestral populations (Long, 1991). The admixture phenomenon is ubiquitous over sexually reproducing organisms (Pritchard et al., 2000). Admixture is pervasive in humans because migratory movements have brought together peoples from different origins (Cavalli-Sforza et al., 1994). Striking examples include the genetic history of African American and Mestizo populations, for which the contributions of European, Native American, and African populations had been studied extensively (Bryc et al., 2010; Tang, Choudhry, et al., 2007).

Most of the concerns raised by definitions of subpopulations are commonly answered by the application of clustering or ancestry estimation approaches such as `structure` or principal component analysis (PCA) (Pritchard et al., 2000; Patterson et al., 2006). These approaches rely on the framework of factor models, where a factor matrix, the Q-matrix for `structure` and the score matrix for PCA, is used to define individual ancestry coefficients, or to assign individuals to their most probable ancestral genetic group (Engelhardt and Stephens, 2010). To account for geographic patterns of genetic variation produced by complex demographic histories, spatially explicit versions of the `structure` algorithm can include models for which individuals at nearby locations tend to be more closely related than individuals from distant locations (François and Durand, 2010).

In this study, we propose new tests to identify outlier loci in admixed and in continuous populations by extending the definition of F_{ST} to this framework (Long, 1991). Our tests are based on the computation of ancestry coefficient and ancestral allele frequency, Q and F, matrices obtained from ancestry estimation programs. We develop a theory for the derivation of this new F_{ST} statistic, defining it as the proportion of genetic diversity due to allele frequency differences among populations in a model with admixed individuals. Then we compute our new statistic using the outputs of two ancestry estimation programs: `snmf` which is used as fast and accurate version of the `structure` algorithm, and `tess3` a fast ancestry estimation program using genetic and geographic data (Frichot, Mathieu, et al., 2014; Caye et al., 2015). Using simulated data sets and SNPs from human and plants, we compared the results

of genome scans obtained with our new F_{ST} statistic with the results of PCA-based methods (Hao et al., 2015; Duforet-Frebourg et al., 2015; Chen, Lee, et al., 2016; Galinsky et al., 2016; Luu et al., 2017b).

2.3 F-statistics for populations with admixed individuals

In this section, we extend the definition of F_{ST} to populations containing admixed individuals, and for which no subpopulations can be defined a priori. We consider SNP data for n individuals genotyped at L loci. The data for each individual, i, and for each locus, ℓ, are recorded into a genotypic matrix Y. The matrix entries, $y_{i\ell}$, correspond to the number of derived or reference alleles at each locus. For diploid organisms, $y_{i\ell}$ is an integer value 0, 1 or 2.

2.3.1 A new definition of F_{ST}

Suppose that a population contains admixed individuals, and the source populations are unknown. Assume that individual ancestry coefficients, Q, and ancestral population frequencies, F, are estimated from the genotypic matrix Y by using an ancestry estimation algorithm such as `structure` (Pritchard et al., 2000). Consider a particular locus, ℓ, and let f_k be the reference allele frequency in ancestral population k at that locus. We set

$$f = \sum_{k=1}^{K} q_k f_k \,,$$

where q_k is the average value of the population k ancestry coefficient over all individuals in the sample, and the ancestral allele frequencies are obtained from the F matrix. Our formula for F_{ST} is

$$F_{\mathrm{ST}} = 1 - \frac{\sum_{k=1}^{K} q_k f_k (1 - f_k)}{f(1-f)} \,. \tag{2.1}$$

The above definition of F_{ST} for admixed populations is obviously related to the original definition of Wright's fixation index. Assuming K predefined subpopulations,

Wright's definition of F_{ST} writes as follows (Wright, 1949)

$$F_{\mathrm{ST}} = 1 - \frac{H_{\mathrm{S}}}{H_{\mathrm{T}}},$$

where $H_{\mathrm{S}} = \sum_{k=1}^{K} n_k f_k (1 - f_k)/n$, $H_{\mathrm{T}} = f(1-f)$, n_k is the sample size, f_k is the allele frequency in subpopulation k, and f is the allele frequency in the total population. For admixed samples, the estimates of the sample sizes, n_k, are obtained by setting $n_k = nq_k$, and the sampled allele frequencies are replaced by their ancestral allele frequencies. The interpretation of the new F_{ST} statistic is thus similar to the interpretation of Wright's fixation index. The main distinction is its application to 'idealized' ancestral populations inferred by `structure` or a similar algorithm. For recently admixed populations, our new statistic represents a measure of population differentiation due to population structure prior to the admixture event. Mathematically rigorous arguments for this analogy will be given in a subsequent paragraph.

2.3.2 Admixture estimates

While many algorithms can compute the Q and F matrices, our application of the above definition will focus on ancestry estimates obtained by nonnegative matrix factorization algorithms (Frichot, Mathieu, et al., 2014). Frichot, Mathieu, et al. (2014)'s algorithm runs faster than the Monte-Carlo algorithm implemented in `structure` and than the optimization methods implemented in `faststructure` or `admixture` (Alexander, Novembre, et al., 2009; Raj et al., 2014). Estimates of Q and F matrices obtained by the `snmf` algorithm can replace those obtained by the program `structure` advantageously for large SNP data sets (Wollstein and Lao, 2015).

The `snmf` algorithm estimates the F matrix as follows. Assume that the sampled genotype frequencies can be modelled by a mixture of ancestral genotype frequencies

$$\delta_{(y_{i\ell}=j)} = \sum_{k=1}^{K} Q_{ik} G_{k\ell}(j), \quad j = 0, 1, \ldots, p,$$

where $y_{i\ell}$ is the genotype of individual i at locus ℓ, the Q_{ik} are the ancestry coefficients for individual i in population k, the $G_{k\ell}(j)$ are the ancestral genotype frequencies in population k, and p is the ploidy of the studied organism (δ is the Kronecker delta

symbol indicating the absence/presence of genotype j). For diploids ($p = 2$), the relationship between ancestral allele and genotype frequencies can be written as follows

$$F_{k\ell} = G_{k\ell}(1)/2 + G_{k\ell}(2).$$

The above equation implies that the sampled allele frequencies, $x_{i\ell}$, satisfy the following equation

$$x_{i\ell} = y_{i\ell}/2 = \sum_{k=1}^{K} Q_{ik} F_{k\ell}\,,$$

which makes the estimates consistent with the definition of F_{ST}.

2.3.3 Population differentiation tests

The regression framework explained in the next paragraph leads to a direct approximation of the distribution of F_{ST} under the null-hypothesis of a random mating population (Sokal and Rohlf, 1981). In this framework, we define the squared z-scores as follows

$$z^2 = (n - K)\frac{F_{\mathrm{ST}}}{1 - F_{\mathrm{ST}}}\,.$$

Assuming random mating at the population level, we have

$$z^2/(K-1) \sim F(K-1, n-K)\,,$$

where $F(K-1, n-K)$ is the Fisher distribution with $K-1$ and $n-K$ degrees of freedom. In addition, we assume that the sample size is large enough to approximate the distribution of squared z-scores as a chi-squared distribution with $K-1$ degrees of freedom.

A naive application of this theory would lead to an increased number of false positive tests due to population structure. In genome scans, we adopt an empirical null-hypothesis testing approach which recalibrates the null-hypothesis. The principle of test calibration is to evaluate the levels of population differentiation that are expected at selectively neutral SNPs, and modify the null-hypothesis accordingly

(François, Martins, et al., 2016). Following GWAS approaches, this can be achieved after computing the genomic inflation factor, defined by the median of the squared z-scores divided by the median of a chi-squared distribution with $K-1$ degrees of freedom (genomic control, Devlin and Roeder (1999)).

2.3.4 Software

The methods described in this section were implemented in the R package LEA (Frichot and François, 2015). A short tutorial on how to compute the F_{ST} statistic and implement the tests is available at `http://goo.gl/OsRhLQ`.

2.3.5 Mathematical theory

A classical definition for the fixation index, F_{ST}, corresponds to the proportion of the genetic variation (or variance) in sampled allele frequency that can be explained by population structure

$$F_{\mathrm{ST}} = \frac{\sigma_T^2 - \sigma_S^2}{\sigma_T^2} \tag{2.2}$$

where, in the analysis of variance terminology, σ_T^2 is the total variance and σ_S^2 is the error variance (Weir, 1996). This definition of F_{ST}, which uses a linear regression framework, can be extended to models with admixed individuals in a straightforward manner. Suppose that a population contains admixed individuals, and assume we have computed estimates of the Q and F matrices. For diploid organisms, a genotype is the sum of two parental gametes, taking the values 0 or 1. In an admixture model, the two gametes can be sampled either from the same or from distinct ancestral populations. The admixture model assumes that individuals mate randomly at the moment of the admixture event. Omitting the locus subscript ℓ, a statistical model for an admixed genotype at a given locus can be written as follows

$$y = x_1 + x_2$$

where x_1 and x_2 are independent Bernoulli random variables modelling the parental gametes. The conditional distribution of x_1 (resp. x_2) is such that $\mathrm{prob}(x_1 = 1|\mathrm{Anc}_1 =$

$k) = f_k$ where f_k is the allele frequency in ancestral population k, Anc is an integer value between 1 and K representing the hidden ancestry of each gamete. The sampled allele frequency is defined as $x = y/2$ (x taking its values in 0, 1/2, 1). Thus the expected value of the random variable x is given by the following formula

$$f = \mathrm{E}[x] = \sum_{k=1}^{K} q_k f_k \,,$$

where $q_k = \mathrm{prob}(\mathrm{Anc} = k)$. The total variance of x satisfies

$$2\sigma_T^2 = 2\mathrm{Var}[x] = f(1-f).$$

Using the Q and F matrices, q_k can be estimated as the average value of the ancestry coefficients over all individuals in the sample, and the ancestral allele frequencies can be estimated as $f_k = F_k$.

To compute the error variance, σ_S^2, we consider that the two gametes originate from the same ancestral population. Assuming Hardy-Weinberg equilibrium in the ancestral populations, the error variance can be computed as follows

$$2\sigma_S^2 = \sum_{k=1}^{K} q_k f_k (1-f_k) \,,$$

and the use of equation 2.2 for F_{ST} concludes the proof of equation 2.1.

2.4 Simulation experiments and data sets

2.4.1 Simple simulation models

In a first series of simulations, we created replicate data sets close to the underlying assumptions of population differentiation tests (Lewontin and Krakauer, 1973; Beaumont and Nichols, 1996). While relying on simplified assumptions, those easily reproducible simulations have the advantage of providing a clear 'proof-of-concept' framework which connects our new statistic to the classical theory. Admixed genotypes from a unique continuous population were obtained from two ancestral gene pools. In this continuous population, individual ancestry varied gradually along a longitudinal axis. The samples contained 200 individuals genotyped at 10,000 unlinked

SNPs. Ancestral polymorphisms were simulated based on Wright's two-island models. Two values for the proportion of loci under selection were considered (5% and 10%). To generate genetic variation at outlier loci, we assumed that adaptive SNPs had migration rates smaller than the migration rate at selectively neutral SNPs. In this model, adaptive loci experienced reduced levels of ancestral gene flow compared to the genomic background (Bazin et al., 2010). The effective migration rate at a neutral SNP was equal to one of the four values $4Nm = 20, 15, 10, 5$. The effective migration rate at an adaptive SNP was equal to one of the four values $4Nm_s = 0.1, 0.25, 0.5, 1$. A total number of 32 different data sets were generated by using the computer program ms (Hudson, 2002).

The model for admixture was based on a gradual variation of ancestry proportions across geographic space (Durand et al., 2009). Geographic coordinates (x_i, y_i) were created for each individual from Gaussian distributions centered around two centroids put at distance 2 on a longitudinal axis (standard deviation [SD] = 1). As it happens in a secondary contact zone, we assumed that the ancestry proportions had a sigmoidal shape across space (Barton and Hewitt, 1985),

$$p(x_i) = \frac{1}{(1 + e^{-x_i})}\,.$$

For each individual, we assumed that each allele originated in the first ancestral population with probability $p(x_i)$ and in the second ancestral population with probability $1 - p(x_i)$ (Durand et al., 2009).

2.4.2 Complex simulation models

To evaluate the power of tests in realistic landscape simulations, we used six publicly available data sets previously described by Lotterhos and Whitlock (2015). In those scenarios, the demographic history of a fictive species corresponded to nonequilibrium isolation by distance due to expansion from two refugia. The simulations mimicked a natural population whose ranges have expanded since the last glacial maximum, potentially resulting in secondary contact (Hewitt, 2000). The study area was modelled as a square with 360 × 360 demes. Migration was determined by a dispersal kernel

with standard deviation $\sigma = 1.3$ demes, and the carrying capacity per deme was 124. The data sets consisted of 9900 neutral loci and 100 selected loci. Twenty unrelated individuals were sampled from thirty randomly chosen demes. For each replicate data set, a selective landscape was randomly generated based on spherical models described as 'weak clines' (details in Lotterhos and Whitlock (2015)). All selected loci adapted to this landscape.

2.4.3 Computer programs

We performed genome scans for selection using three factor methods: `snmf` (Frichot, Mathieu, et al., 2014), `tess3` (Caye et al., 2015), `pcadapt` (Luu et al., 2017b; Duforet-Frebourg et al., 2015). A fourth method used the standard F_{ST} statistic where subpopulations were obtained from the assignment of individuals to their most likely genetic cluster. Like for `snmf`, the `tess3` estimates of the Q and G matrices are based on matrix factorization techniques. The main difference between the two programs is that `tess3` computes ancestry estimates by incorporating information on individual geographic coordinates in its algorithm whereas the `snmf` algorithm is closer to `structure` (Caye et al., 2015). The default values of the two programs were implemented for all their internal parameters. Each run of the two programs was replicated five times, and the run with the lowest cross-entropy value was selected for computing F_{ST} statistics according to formula (1). We compared the results of `snmf` and `tess3` with the results of the program `pcadapt` (Luu et al., 2017b). The test statistic of the latest version of `pcadapt` is the Manhanalobis distance relative to the z-scores obtained after regressing the SNP frequencies on the $K - 1$ principal components. As for `snmf` and for `tess3`, test calibration in `pcadapt` was based on the computation of the genomic inflation factor. For genome scans based on the F_{ST} statistic where subpopulations are obtained from the assignment of individuals to their most likely genetic cluster, we used a chi-squared distribution with $K - 1$ of freedom after recalibration of the null-hypothesis using genomic control. Before applying the methods to the simulated data sets, the SNPs were filtered out and only the loci with minor allele frequency greater than 5% were retained for analysis.

2.4.4 Real data sets

To provide an application of our method to natural populations, we reanalyzed data from the model plant organism *Arabidopsis thaliana*. This annual plant is native to Europe and central Asia, and within its native range, it goes through numerous climatic conditions and selective pressures (Mitchell-Olds and Schmitt, 2006). We analyzed genomic data from 120 European lines of *A. thaliana* genotyped for 216k SNPs, with a density of one SNP per 500 bp (Atwell et al., 2010). To reduce the sensitivity of methods to an unbalanced sampling design,fourteen ecotypes from Northern Scandinavia were not included in our analysis. Those fourteen ecotypes represented a small divergent genetic cluster in the original data set. In addition to the plant data, we analyzed human genetic data for 1,385 European individuals genotyped at 447k SNPs (Nelson et al., 2008).

2.4.5 Candidate lists

After recalibration of the null-hypothesis using genomic inflation factors, histograms of test significance values were checked for displaying their correct shape. Then, False Discovered Rate (FDR) control algorithms were applied to significance values using the Storey and Tibishirani algorithm (Storey and Tibshirani, 2003). For simulated data, lists of outlier loci were obtained for an expected FDR value of 10%. The same nominal level was applied for the analysis of the human data set. For *A. thaliana*, an expected FDR value of 1% was applied, and a consensus list of loci was obtained by including all peak values present in Manhattan plots for `snmf` and `tess3`.

2.5 Results

2.5.1 Simple simulation models

We evaluated the performances of genome scans using tests based on `snmf`, `tess3`, `pcadapt`, and F_{ST}, in the presence of admixed individuals. For `snmf` and for `tess3`, we used $K = 2$ ancestral populations. This value of K corresponded to the minimum of the cross-entropy criterion when K was varied in the range 1 to 6, and it also

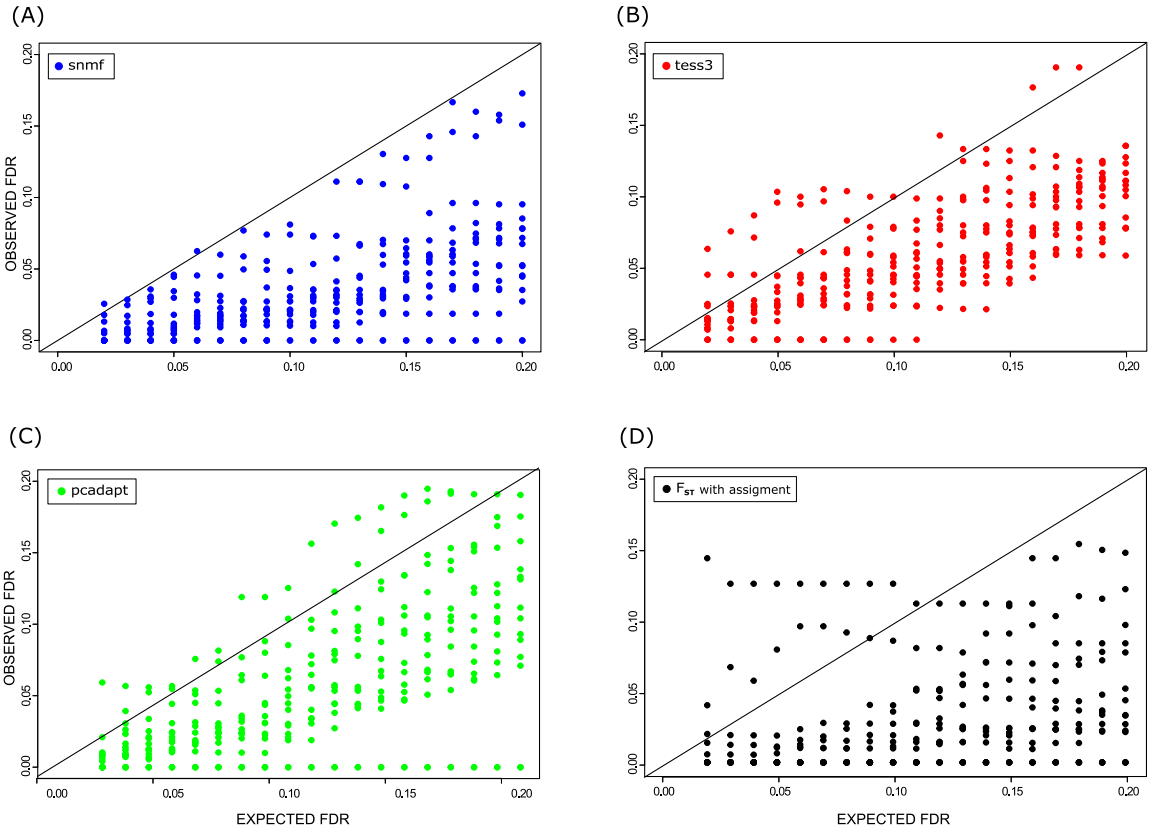


Figure 2.1 – **FDR for simulations of admixed populations.** Simulation of ancestral populations based on 2-island models with various levels of population differentiation and selection. Sixteen data sets contained 5% of truly selected loci. Observed false discovery rates for an expected level of FDR equal to 0.1. (A) F_{ST} tests based on `snmf` Q and F matrices, (B) F_{ST} tests based on `tess3` Q and F matrices, (C) Luu et al. (2017a) `pcadapt` statistic, (D) Standard F_{ST} test based on assignment of individuals to their most likely genetic cluster.

corresponded to the true number of ancestral populations in the simulations. We used `pcadapt` with its first principal component. Considering expected FDR values between 0.01 and 0.2, we computed observed FDR values for the lists of outlier loci produced by each test. The observed FDR values remained generally below their expected values (Figure 2.1 for data sets with 5% of loci under selection, Figure S1 for data sets with 10% of loci under selection). These observations confirmed that the use of genomic inflation factors leads to overly conservative tests (François, Martins, et al., 2016). Since similar levels of observed FDR values were observed across the 4 tests, we did not implement other calibration methods than genomic control.

Next, we evaluated the sensitivity (power) of the four tests in each simulation

scenario. Our experiments confirmed that the use of approaches that estimate ancestry coefficients is appropriate when no subpopulation can be predefined (Figure 2.2A for ancestry coefficient estimates). As we expected from the simulation process, the tests had higher power when the relative levels of selection intensity were higher. For $4Nm = 5$ and $4Nm_s = 0.1, 0.25, 0.5$, and 1, the power of tests for `snmf`, `tess3`, `pcadapt` was close to 27% for data sets with 5% of outliers (Figure 2.2B, expected FDR equal to 10%). The F_{ST} test based on assignment of individuals to their most likely cluster failed to detect outlier loci (power value equal to 0%). For $4Nm = 10$, the power of the tests ranged between 40% and 45% for `snmf`, `tess3`, `pcadapt`, and it was equal to 26% for the F_{ST} test (Figure 2.2B). For $4mN \geq 15$, corresponding to the highest selection rates, the power was approximately equal to 50% for all methods considered. The relatively low power values confirmed that the tests were conservative, and truly-adaptive loci were difficult to detect. To provide an upper bound on the power of outlier tests in the context of admixed populations, we applied an F_{ST} test to the samples obtained prior to admixture, estimating allele frequencies from their true ancestral populations. For $4Nm = 5$ and 10, the power of the tests for `snmf`, `tess3`, `pcadapt` was similar to the power obtained when we applied outlier tests to the data before admixture (Figure 2.2B). The results for data sets with 10% of selected loci were similar to those obtained with 5% of selected loci (Figure S2).

2.5.2 Complex simulation models

We compared the power of factor methods to the power of tests based on assignment of individuals to their most likely cluster in realistic landscape simulations (Lotterhos and Whitlock, 2015). As a consequence of isolation by distance, the cross-entropy curve for `snmf` decreased with the value of the number of clusters, but the curve did not exhibit a minumum. A plateau reached at $K = 6$ indicated that this value of K could be the best choice for modelling the mixed levels of ancestry in the data (Figure 2.3A). In agreement with this result, `pcadapt` consistently found 5 axes of variation in the data. For values of $K = 4 - 7$ and for an expected level of FDR of 10%, the power of tests based on factor methods ranged between 0.82 and 0.87 (Figure 2.3B).

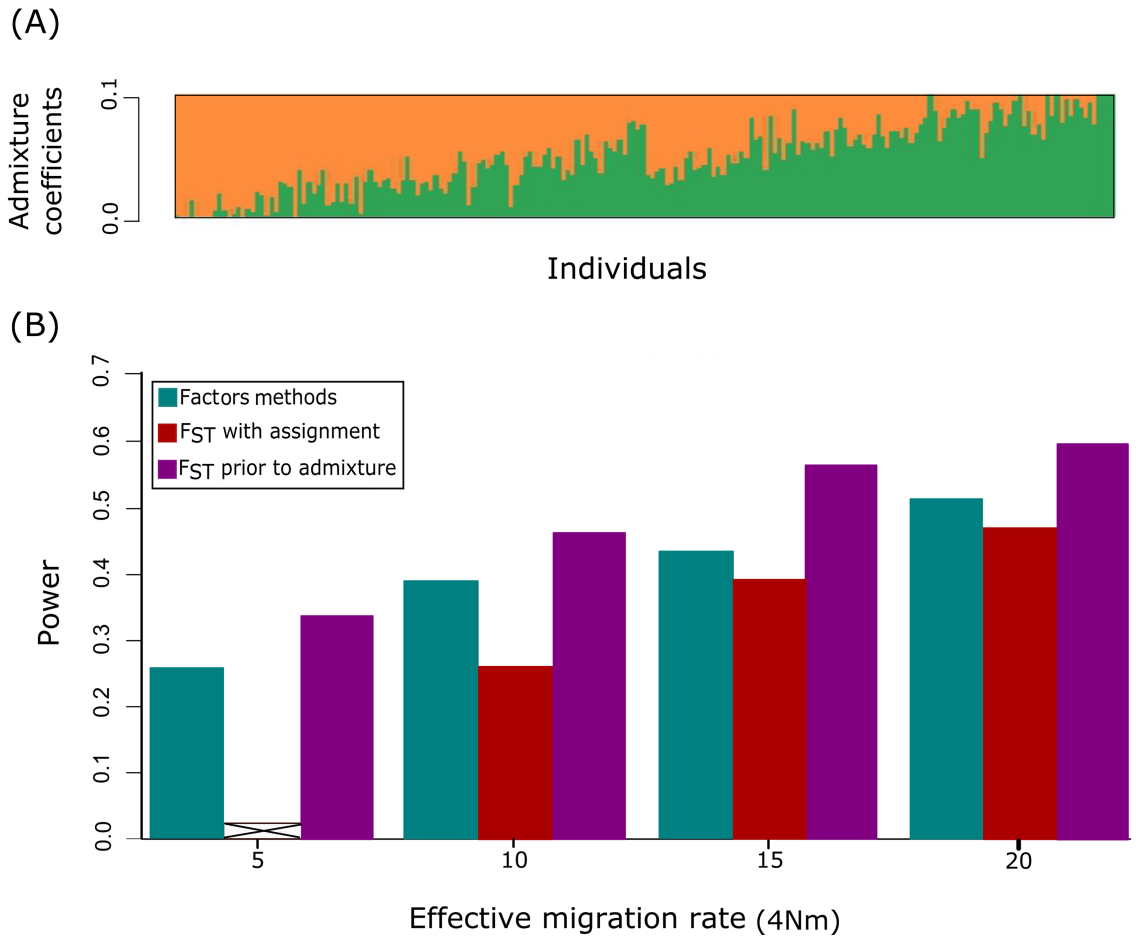


Figure 2.2 – **Power in simulations of admixed populations**. Simulations of ancestral populations based on 2-island models with various levels of selection and background of levels of population differentiation ($4Nm$). Sixteen data sets contained 5% of truly selected loci. (A) Individual ancestry coefficients estimated from neutral loci using `snmf` with $K = 2$. (B) Power estimates for tests based on factor methods (grouping `snmf`, `tess3` and `pcadapt`), for F_{ST} tests in which individuals were assigned to their most likely cluster, and for F_{ST} tests prior to admixture. Power values were computed by considering an expected FDR value equal to 0.1. For $4Nm = 5$ (relatively weak selection intensity), the F_{ST} test based on assignment failed to detect outlier loci.

Although SNP rankings were not different for `pcadapt`, the `pcadapt` tests were less conservative than the tests based on the default values of `snmf` (values not reported). Classical tests that assigned individuals to their most likely cluster had power ranging between 0.44 and 0.48. The power values for classical F_{ST} tests were substantially lower than those obtained with the new tests.

2.5.3 Arabidopsis data

We applied `snmf`, `tess3` and `pcadapt` to perform genome scans for selection in 120 European lines of *Arabidopsis thaliana* (216k SNPs). Each ecotype was collected from a unique geographic location, and there were no predefined populations. To study adaptation at the continental scale, a small number ecotypes from Northern Scandinavia, which were grouped by clustering programs, were removed from the original data set of Atwell et al. (2010). For `snmf` and `tess3`, the cross-entropy criterion indicated that there are two main clusters in Europe, and that finer substructure could be detected as a result of historical isolation-by-distance processes. For $K = 2$, the western cluster grouped all lines from the British Isles, France and Iberia and the eastern cluster grouped all lines from Germany, and from Central and Eastern Europe (Figure 2.4). For implementing genome scans for selection, we used two clusters in `snmf` and `tess3`, and one principal component in `pcadapt`. The genomic inflation factor was equal to $\lambda = 11.5$ for the test based on `snmf`, and it was equal to $\lambda = 13.1$ for the test based on `tess3`. The interpretation of these two values is that the background level of population differentiation that was tested in `snmf` and `tess3` is around 0.09 (François, Martins, et al., 2016). For the three methods, the Manhattan plots exhibited peaks at the same chromosome positions (Figure 2.5). For an expected FDR level equal to 1%, the Storey and Tibshirani (2003) algorithm resulted in a list of 572 chromosome positions for the `snmf` tests and 882 for the `tess3` tests. Figure S3 displays a Manhattan plot for the plant genome showing the main outlier loci detected by our genome scans for selection for $K = 2$. Unlike for simulated data, the tests based on PCA were more conservative than the tests based on genetic clusters. Generally, the differences between test significance values among methods could be

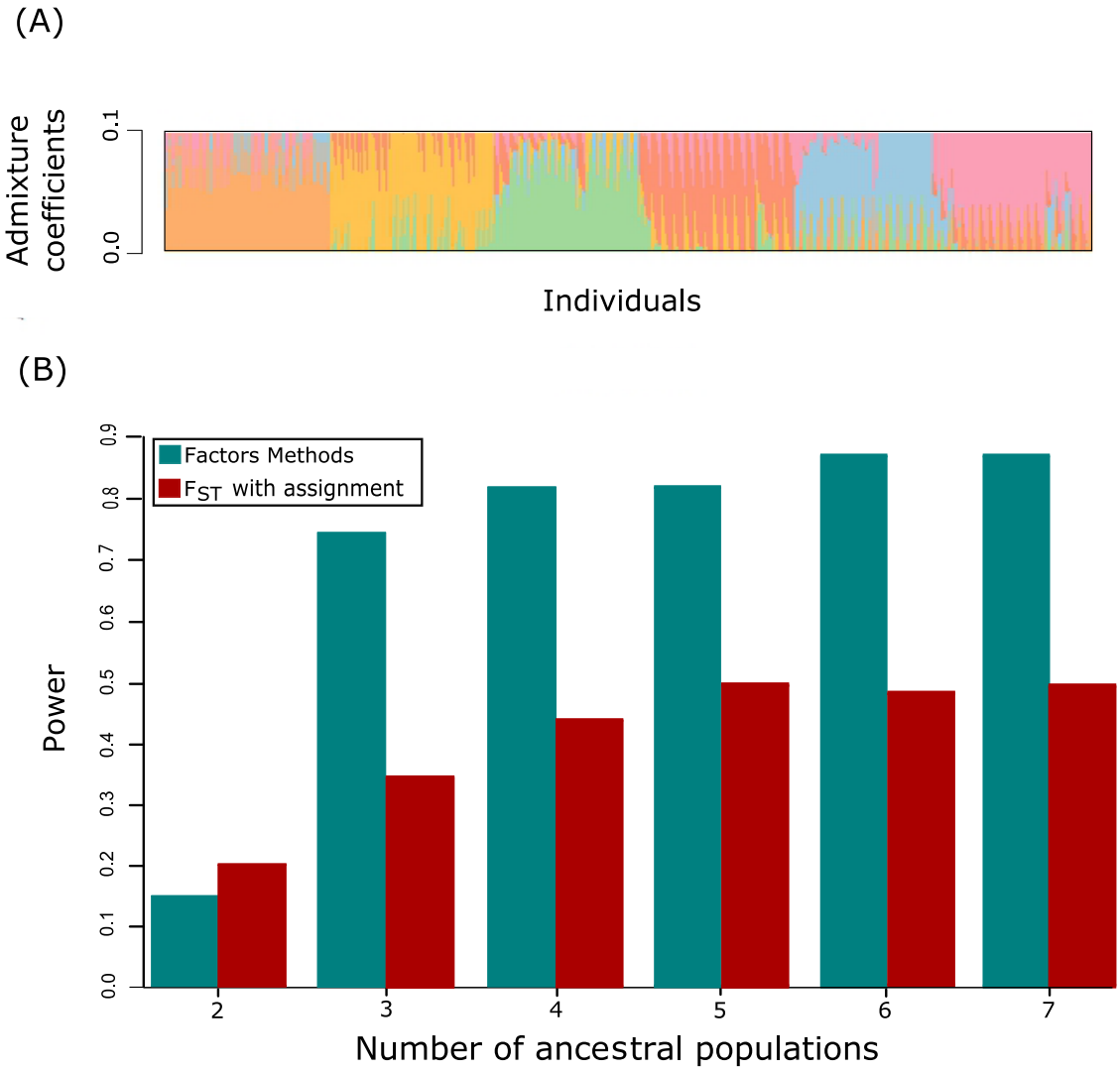


Figure 2.3 – **Power in simulations of range expansions**. (A) Individual ancestry coefficients estimated using `snmf` with $K = 6$ ancestral populations. (B) Power estimates for tests based on factor methods and for F_{ST} tests in which individuals were assigned to their most likely cluster. Power values were computed by considering an expected FDR value equal to 0.1. Factor methods included `snmf` and `pcadapt`.

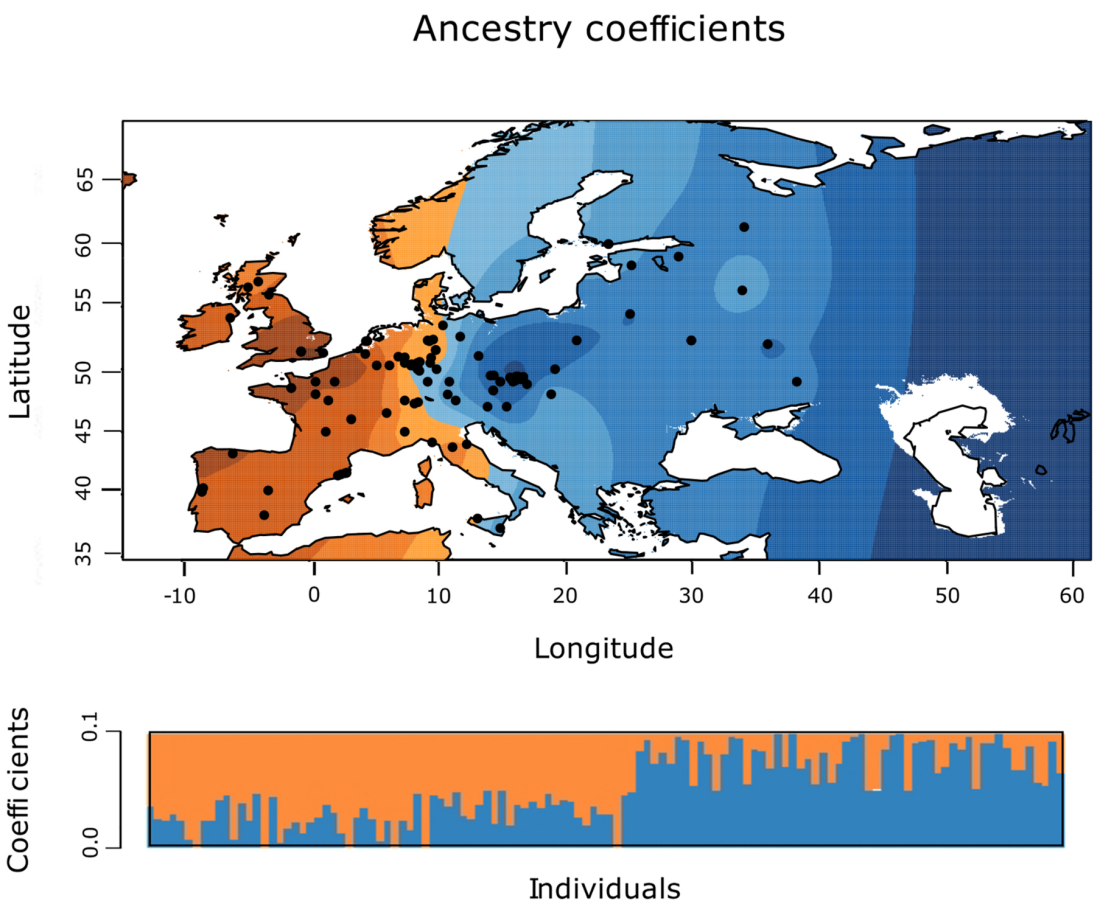


Figure 2.4 – **Ancestry coefficients for** *Arabidopsis thaliana***.** Coefficients estimated using `snmf` with $K = 2$ ancestral populations interpolated on a geographic map of Europe.

attributed to the estimation of the genomic inflation factor and test calibration issues rather than to strong differences in SNP ranking. The results of genome scans for selection were also investigated for values of K greater than 2. The higher values of K revealed additional candidate genomic regions that were consistently discovered by the three factor methods (Figures S4-S6).

Table 2.1 reports a list of 33 candidate SNPs for European *A. thaliana* lines in the 10% top hits, based on the peaks detected by the factor methods. For chromosome 1, the list contains SNPs in the gene AT1G80680 involved in resistance against

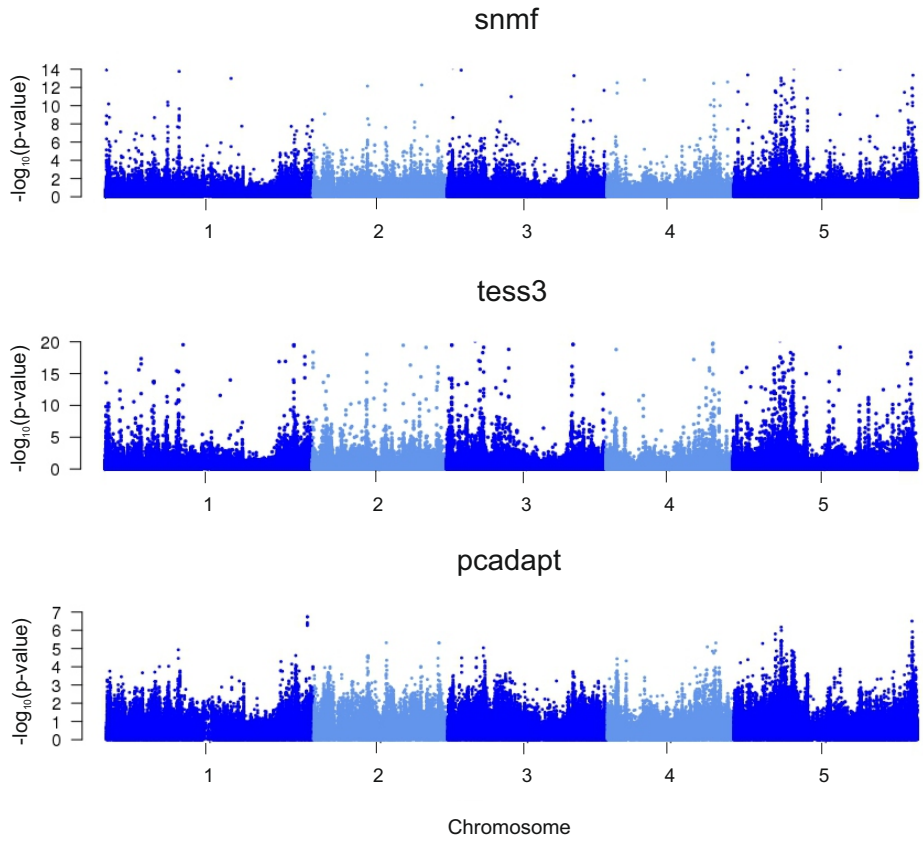


Figure 2.5 – **Manhattan plots of minus log10(p-values) for** *A. thaliana*. Tests using (A) `snmf`, (B) `tess3` and (C) `pcadapt`. The tests based on `pcadapt` were more conservative than the tests based on the other methods.

bacterial pathogens. For chromosome 2, the list contains SNPs in the gene AT2G18440 (AtGUT15), which can be used by plants as a sensor to interrelated temperatures, and which has a role for controlling growth and development in response to a shifting environment (Lu et al., 2005). For chromosome 3, the list contains SNPs in the gene AT3G11920 involved in cell redox homeostasis. Fine control of cellular redox homeostasis is important for integrated regulation of plant defense and acclimatory responses (Muhlenbock et al., 2007). For chromosome 4, we found SNPs in the gene AT4G31180 (IBI1) involved in defense response to fungi. The most important list of candidate SNPs was found in the fifth chromosome. For example, the list of outlier SNPs contained SNPs in the gene AT5G02820, involved in endoreduplication, that might contribute to the adaptation to adverse environmental factors, allowing the maintenance of growth under stress conditions (Chevalier et al., 2011), in the genes AT5G18620, AT5G18630 and AT5G20620 (UBIQUITIN 4) involved in response to temperature stress (Kim, Kim, et al., 2005), and in the gene AT5G20610 which is involved in response to blue light (DeBlasio et al., 2005). Several additional candidates were found with values of K greater than two for the `snmf` tests. For $K = 3$ and $K = 4$ those additional outlier regions included one SNP in the flowering locus FRIGIDA and four SNPs in COP1-interacting protein 4.1 on chromosome 4 (Horton et al., 2012), Figure S6). For the tests with $K = 4$, outlier regions included two SNPs in the FLOWERING LOCUS C (FLC) and five SNPs in the DELAY OF GEMINATION 1 (DOG1) locus (Horton et al., 2012), Figure S6).

2.5.4 Human data

We applied the `snmf` and `pcadapt` tests to 1,385 European individuals from the POPRES data set (447k SNPs in 22 chromosomes). We used $K = 2$ ancestral populations in `snmf` and one principal component for PCA. For `snmf`, the genomic inflation factor was equal to $\lambda = 9.0$, indicating a background level of population differentiation around 0.006 between northern and southern European populations (Figure 2.6). For an expected FDR equal to 10%, we found 205 outlier loci using `snmf` tests, and 165 outlier loci with `pcadapt`. For chromosome 2, the most important

Chromosome	Position (kb)	Gene	Unknown	References
1	132330	AT1G01340	Salt tolerance	Guo et al. (2008)
	490925	AT1G02410	Plant growth and pollen germination	Radin et al. (2015)
	2191723	AT1G07140(SIRANBP)	Encodes a putative ran-binding protein	Wang et al. (2008)
	10779171	AT1G30470	Unknown	
	26503961	AT1G70340	Unknown	
	29516989	AT1G78450	Unknown	
	30324008	AT1G80680	Defense response	Roth and Wiermer (2012)
2	7995729	AT2G18440 (AtGUT15)	Encodes a noncoding RNA	
3	2048905	AT3G06580 (GAL1)	Galactose metabolic process	Wang et al. (2008)
	3772311	AT3G11920	Cell redox homeostasis	
	5476074	AT3G16170 (AAE13)	Fatty acid biosynthetic process	Chen, Kim, et al. (2011)
	18595731	AT3G50150	Unknown	
	18362443	AT3G49530	Response to cold	Chawade et al. (2007)
4	15155879	AT4G31180 (IBI1)	Defense response	Rajjou et al. (2006)
5	642558	AT5G02820	Endoreduplication	
	644279	AT5G02830	Unknown	
	6092682	AT5G18400 (ATDRE2)	Apoptotic process	Wang et al. (2008)
	6195917	AT5G18620	Response to cold	Kim, Kim, et al. (2005)
	6202633	AT5G18630	Lipid metabolic process	Wang et al. (2008)
	6947843	AT5G20540	Unknown	
	6952417	AT5G20550	Oxidation-reduction process	
	6956660	AT5G20570 (ATRBX1)	Protein ubiquitination	Ascencio-Ibanez et al. (2008)
	6958628	AT5G20580	Unknown	
	6963438	AT5G20590	Cell wall organization or biogenesis	Xin et al. (2007)
	6968690	AT5G20610	Response to blue light	DeBlasio et al. (2005)
	6973071	AT5G20620 (UBIQUITIN 4)	Cellular protein modification process	Sun and Callis (1997)
	8500476	AT5G24770	Defense response	Catinot et al. (2015)
	8773789	AT5G25280	Unknown	
	8823283	AT5G25400	Carbohydrate transport	Wang et al. (2008)
	10856791	AT5G28830	Unknown	
	26161831	AT5G65460 (KAC2)	Photosynthesis	He et al. (2005)
	26176021	AT5G65480	Unknown	Wang et al. (2008)
	26225832	AT5G65630 (GTE7)	Defense response	Wang et al. (2008)

Table 2.1 – **List of 33 candidate SNPs for European ecotypes of** *A. thaliana***.** The list was based on the list of p-values obtained by using an expected FDR of 1% for `snmf` and `tess3` tests.

signal of selection was found at the lactase persistence gene (*LCT*) (Bersaglieri et al., 2004). For chromosome 4, 5 SNPs were found at the *ADH1C* locus that is involved in alcohol metabolism (Han et al., 2007), close to the *ADH1B* locus reported by Galinsky et al. (2016). For chromosome 6, a signal of selection corresponding to the human leukocyte antigen (*HLA*) region was identified. For chromosome 15, there was an outlier SNP in the *HERC2* gene, which modulates human pigmentation (Visser et al., 2012), Figure 2.6).

2.6 Discussion

When no subpopulation can be defined a priori, analysis of population structure commonly relies on the computation of the Q (and F) ancestry matrix obtained through the application of the program `structure` or one of its improved versions (Pritchard et al., 2000; Tang, Peng, et al., 2005; Chen, Durand, et al., 2007; Alexander, Novembre, et al., 2009; Raj et al., 2014; Frichot, Mathieu, et al., 2014; Caye et al., 2015). In this context, we proposed a definition of F_{ST} based on the Q and F matrices, and we used this new statistic to screen genomes for signatures of diversifying selection. By modelling admixed genotypes, our definition of F_{ST} was inspired by an analysis of variance approach for the genotypic data (Weir and Cockerham, 1984; Holsinger and Weir, 2009).

The estimator for F_{ST} presented here is related to the estimator proposed by Long (1991) for population data. Long's estimator was obtained from the variance of allele frequencies with respect to their expectations based on an admixture model, that enables estimating the effect of genetic drift and the effective size of the hybrid population. In order to obtain Long's estimate, multiple locus samples are required from the hybrid population and from all contributing parental populations. For the method proposed in our manuscript, information on ancestral genetic diversity is evaluated with less prior assumptions by the application of ancestry estimation programs.

Ancestry coefficients computed by `structure` or similar programs are conceptual abstractions that do not always reflect demographic history correctly (Kalinowski,

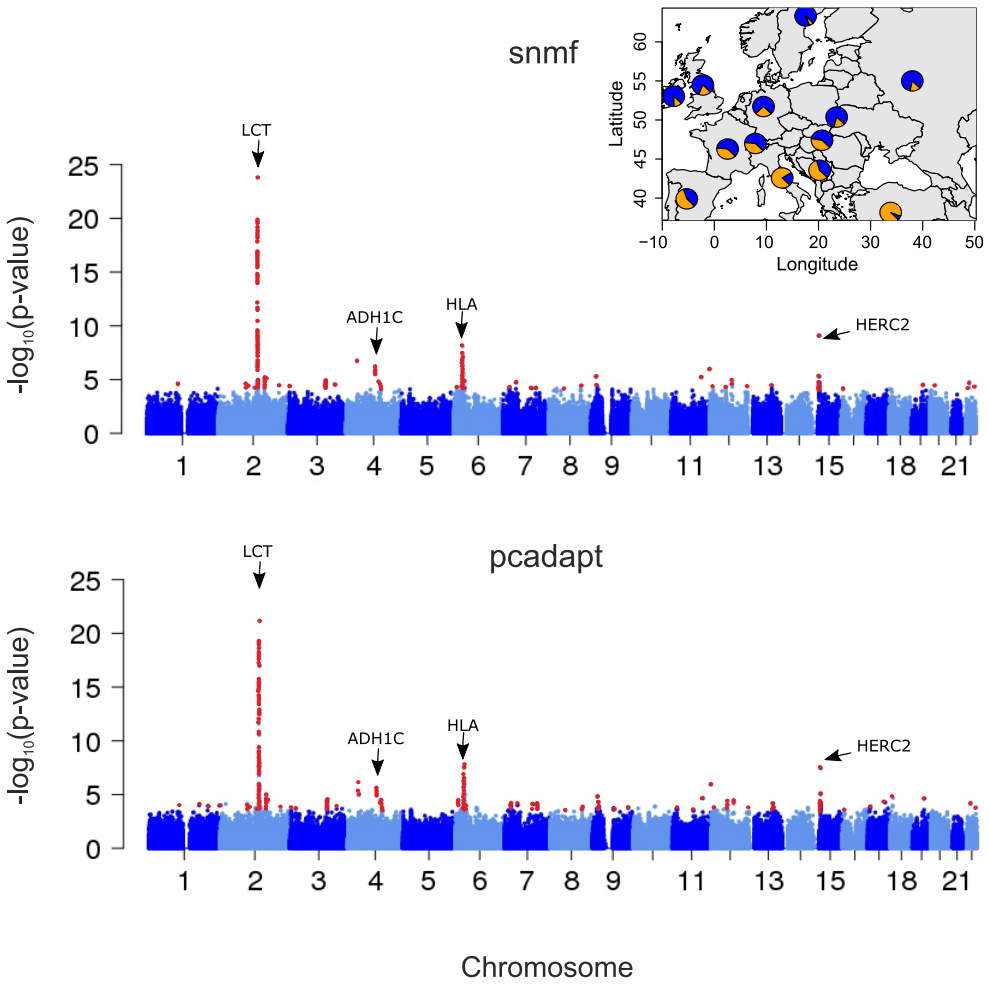


Figure 2.6 – **Manhattan plots of minus log10(p-values) for Europeans (POPRES data set).** Tests using (A) `snmf` and (B) `pcadapt`. Candidate loci detected by genome scans for selection are colored in red for an expected FDR level of 10%. The inserted figure displays population structure estimated with `snmf` with $K = 2$ populations.

2010; Puechmaille, 2016; Falush et al., 2016). Assuming that a large number of SNPs are genotyped across multiple populations, the calibration of statistical tests of neutrality do not require assumptions about population demographic history. Our simulations of admixed populations provided evidence that the tests based on this new statistic had an increased power compared to tests in which we assigned individuals to their most probable cluster. Interestingly, the power of those tests was only slightly lower than standard F_{ST} tests based on the truly ancestral allele frequencies. Going beyond simplified simulation scenarios, we evaluated the power of our tests in range expansion scenarios with complex patterns of isolation by distance. In those scenarios, genetic correlation among samples inflates the variance of population differentiation statistics (Bierne et al., 2013a). We observed that inflation factor corrections reduced this problem when using numbers of clusters (K) greater than 2. Although a 'true' value for K did not exist, we found that the power of our tests was optimal for K estimated from a PCA or by cross-validation using our factor model. In this case, the ancestry coefficients disagreed with the known demographic history (simulated organisms expanded from two refugia), but the gain in performance in favor of the new tests was even higher than in the simple proof-of-concept simulations tailored to the new method.

Our reanalysis of European *A. thaliana* genetic polymorphisms provided a clear example of the usefulness of our F_{ST} statistic to detect targets of natural selection in plants. European ecotypes of *Arabidopsis thaliana* are continuously distributed across the continent, with population structure influenced by historical isolation-by-distance processes (Atwell et al., 2010; Hancock et al., 2011; François, Blum, et al., 2008). The application of our F_{ST} statistic to the SNP data suggested several new candidate loci involved in resistance against pathogens, in growth and development in response to a shifting environment, in the regulation of plant defense and acclimatory responses, in the adaptation to adverse environmental factors, in allowing the maintenance of growth under stress conditions, in response to temperature stress or response to light.

An alternative approach to investigating population structure without predefined populations is by using principal component analysis (Patterson et al., 2006). Statis-

tics extending the definition of F_{ST} were also proposed for PCA (Hao et al., 2015; Duforet-Frebourg et al., 2015; Galinsky et al., 2016; Chen, Lee, et al., 2016). The performances of PCA statistics and our new F_{ST} statistic were highly similar. The small differences observed for the two tests could be ascribed to the chi-squared distribution approximation and to the estimation of inflation factors to calibrate the null-hypothesis. The idea of detecting signatures of selection in an admixed population has a considerable history and has been explored since the early seventies (Blumberg and Hesser, 1971; Adams and Ward, 1973; Tang, Choudhry, et al., 2007). The connection between our definition of F_{ST} and previous works shows that the methods studied in this study, including PCA or ancestry programs, are extensions of classical methods of detection of selection using admixed populations (Long, 1991). Our results allow us to hypothesize that the age of selection detected by PCA and by our new method is similar. Thus it is likely that the selective sweeps detected by PCA and F_{ST} methods correspond to ancient selective sweeps already differentiating in ancestral populations. A comparison of our results for Europeans from the POPRES data sets and the genome-wide patterns of selection in 230 ancient Eurasians provides additional evidence that the signals detected by our F_{ST} were already present in the populations that were ancestral to modern Europeans (Mathieson et al., 2015).

While only minor differences between the ranking of p-values with 4 methods were observed, the results might be still sensitive to the algorithm used to estimating the ancestry matrices. Wollstein and Lao (2015) performed an extensive comparison of 3 recently proposed ancestry estimation methods, `admixture`, `faststructure`, `snmf` (Alexander and Lange, 2011; Raj et al., 2014; Frichot, Mathieu, et al., 2014), and they concluded that the accuracy of the methods could differ in some simulation scenarios. In practice, it would be wise to apply several methods and to combine their results by using a meta-analysis approach as demonstrated in François, Martins, et al. (2016).

2.7 Data Accessibility

Simulated data are available from Lotterhos KE, Whitlock MC (2015) Data from: The relative power of genome scans to detect local adaptation depends on sampling

design and statistical method. Dryad Digital Repository:

`http://dx.doi.org/10.5061/dryad.mh67v`.

The Atwell et al. (2010) data are publicly available from

`https://github.com/Gregor-Mendel-Institute/atpolydb`.

The POPRES data were obtained from dbGaP (accession number phs000145.v1.p1).

2.8 Acknowledgements

We are grateful to three anonymous reviewers for their time and efforts in evaluating our manuscript. Helena Martins acknowledges support from the "Ciências sem Fronteiras" scholarship programme from the Brazilian government. This work has been partially supported by the LabEx PERSYVALLab (ANR-11-LABX-0025-01) funded by the French programme Investissement d'Avenir and by the ANR AGRHUM project (ANR-14-CE02-0003-01). Of acknowledges support from Grenoble INP and from the "Agence Nationale de la Recherche" (project AFRICROP ANR-13-BSV7-0017).

2.9 Supplementary material

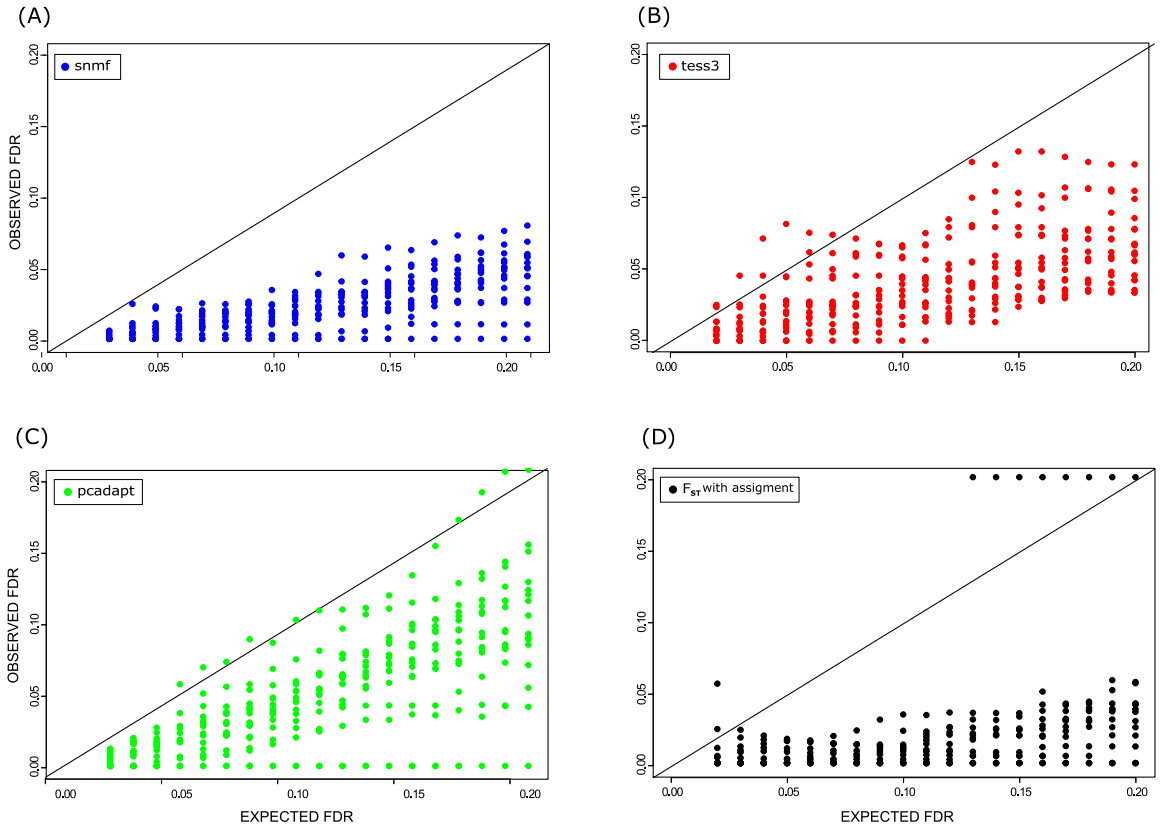


Figure S1. FDR for simulations of admixed populations (10% of outliers). Simulation of ancestral populations based on 2-island models with various levels of population differentiation and selection. Sixteen data sets contained 10% of truly selected loci. Observed false discovery rates for an expected level of FDR equal to 0.1. (A) F_{ST} tests based on `snmf` Q and F matrices, (B) F_{ST} tests based on `tess3` Q and F matrices, (C) Luu et al.'s (2016) `pcadapt` statistic, (D) Standard F_{ST} test based on assignment of individuals to their most likely genetic cluster.

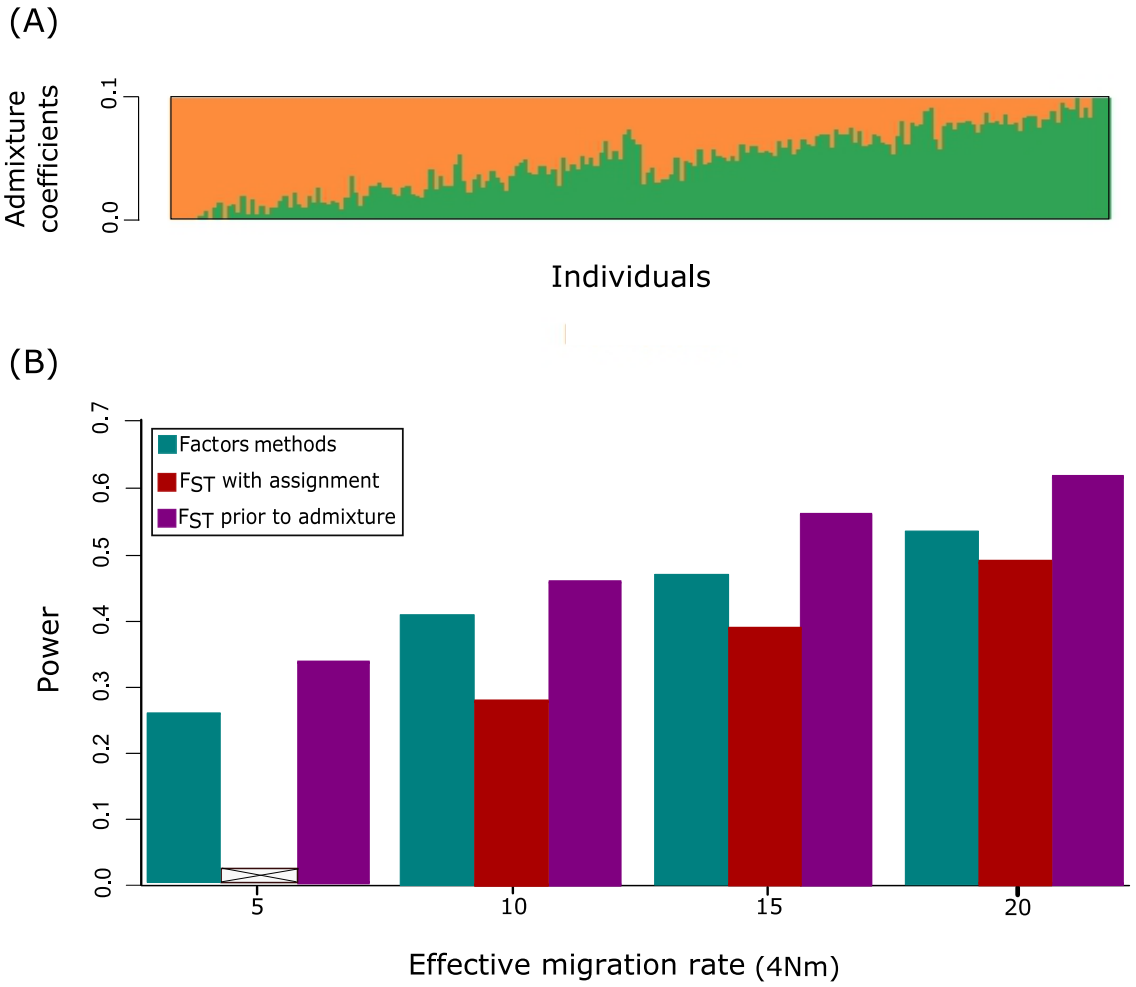


Figure S2. Power in simulations of admixed populations (10% of outliers). Simulations of ancestral populations based on 2-island models with various levels of selection and background of levels of population differentiation ($4Nm$). Sixteen data sets contained 10% of truly selected loci. (A) Individual ancestry coefficients estimated from neutral loci using `snmf` with $K = 2$. (B) Power estimates for tests based on factor methods (grouping `snmf`, `tess3` and `pcadapt`), for F_{ST} tests in which individuals were assigned to their most likely cluster, and for F_{ST} tests prior to admixture. Power values were computed by considering an expected FDR value equal to 0.1. For $4Nm = 5$ (weak selection intensity), the F_{ST} test based on assignment failed to detect outlier loci.

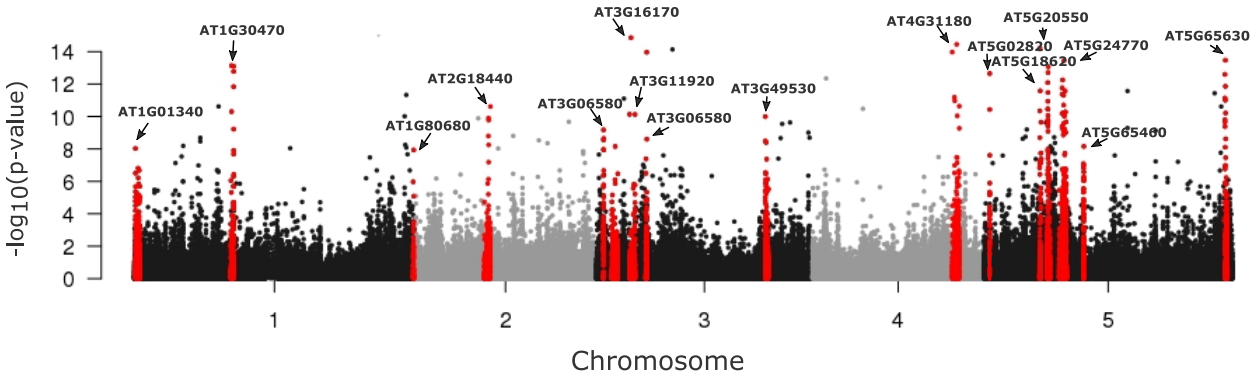


Figure S3. Manhattan plot of minus log10(p-values) for *A. thaliana*. The candidate regions are colored in red. Those regions correspond to an expected FDR level of 1% for `snmf` and `tess3` having more than 5 SNPs in each region.

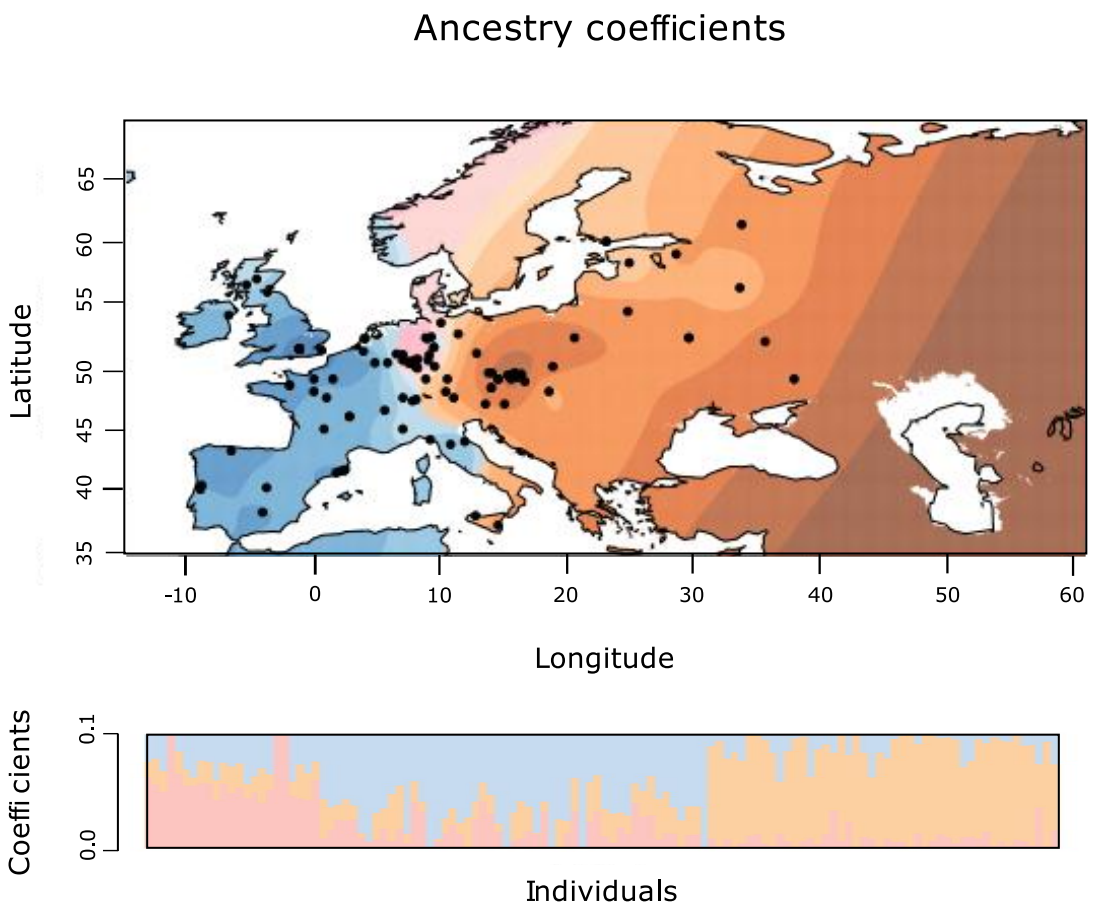


Figure S4. Geographic map of ancestry coefficients for *Arabidopsis thaliana* **using** `snmf` **with** $K = 3$ **ancestral populations.**

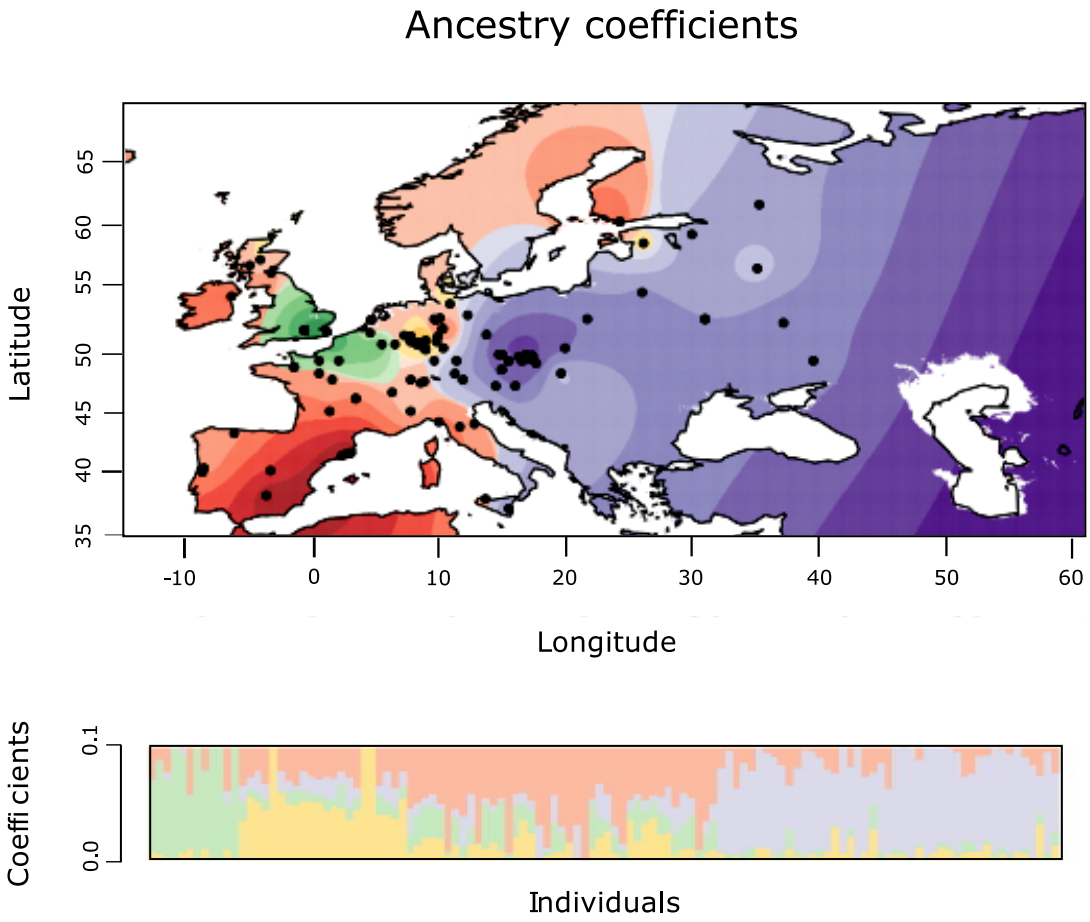


Figure S5. Geographic map of ancestry coefficients for *Arabidopsis thaliana* **using `snmf` with $K = 4$ ancestral populations.**

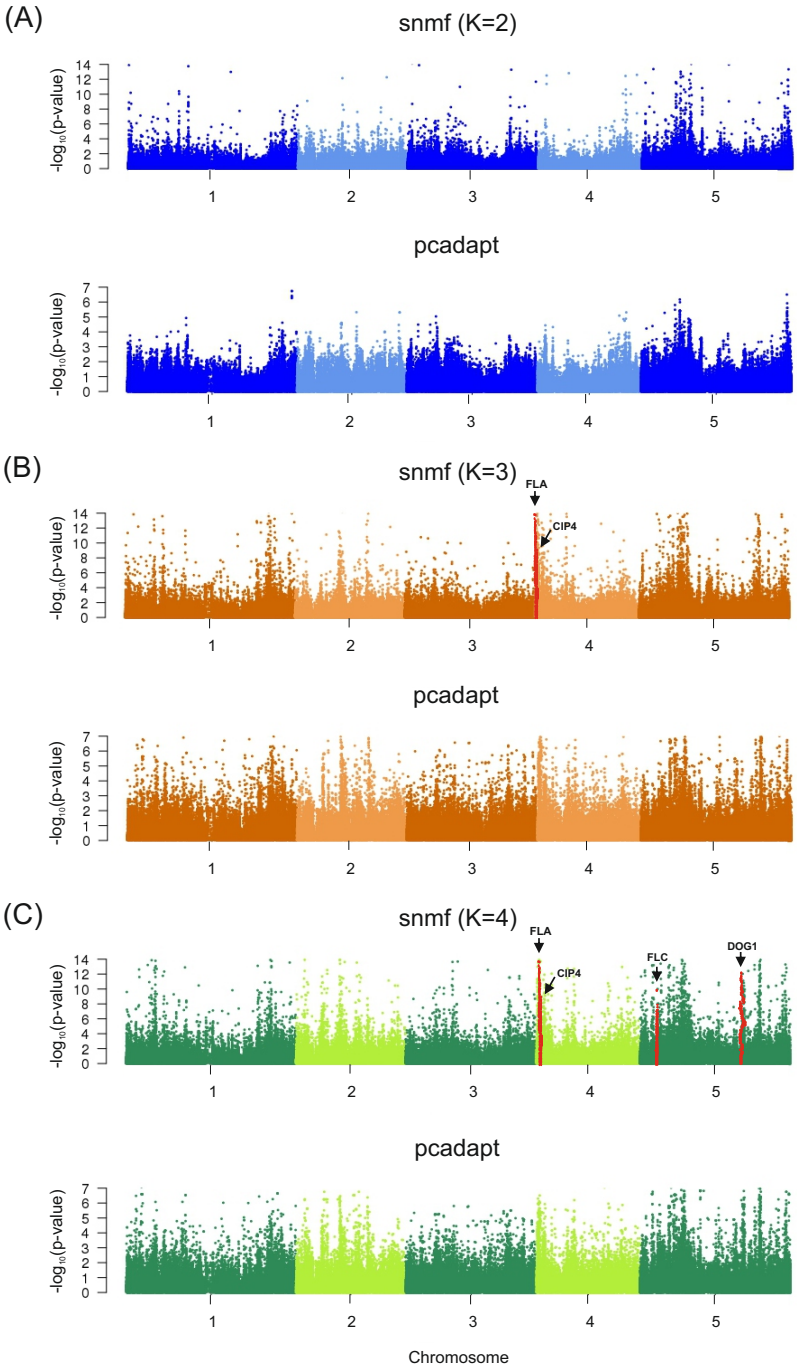


Figure S6. Manhattan plots of minus log10(p-values) for *A. thaliana*. Tests using (A) `snmf` with $K = 2$ ancestral populations and `pcadapt` with 1 principal component, (B) `snmf` with $K = 3$ ancestral populations and `pcadapt` with 2 principal components, (C) `snmf` with $K = 4$ ancestral populations and `pcadapt` with 3 principal components.

Chapter 3

Fast inference of spatial population structure and genome scans for selection

3.1 Abstract

Geography and landscape are important determinants of genetic variation in natural populations, and several ancestry estimation methods have been proposed to investigate population structure using genetic and geographic data simultaneously. Those approaches are often based on computer-intensive stochastic simulations, and do not scale with the dimensions of the data sets generated by high-throughput sequencing technologies. There is a growing demand for faster algorithms able to analyze genome-wide patterns of population genetic variation in their geographic context.

In this study, we present `tess3`, a major update of the spatial ancestry estimation program `tess`. By combining matrix factorization and spatial statistical methods, `tess3` provides estimates of ancestry coefficients with accuracy comparable to `tess` and with run-times much faster than the Bayesian version. In addition, the `tess3` program can be used to perform genome scans for selection, and separate adaptive from non-adaptive genetic variation using ancestral allele frequency differentiation tests. The main features of `tess3` are illustrated using simulated data and analyzing genomic data from European lines of the plant species *Arabidopsis thaliana*.

3.2 Introduction

In population genetics, geography is recognized as an important determinant of genetic variation in natural populations (Wright, 1943; Malécot et al., 1948; Kimura and Weiss, 1964; Cavalli-Sforza, 1966; Epperson, 2003). Normally, spatial patterns of genetic variation can be influenced by geographical distances and by the processes of divergence, by admixture resulting from the colonization of new areas and by landscape barriers.

Statistical approaches to analyze spatial patterns of genetic variation often rely on the inference of population genetic structure from multi-locus genotype data, which is commonly performed using the Bayesian approach implemented in the computer program `STRUCTURE` (Pritchard et al., 2000). Assuming K unobserved ancestral gene pools, `STRUCTURE` computes allele frequencies in each pool, and estimates individual *ancestry coefficients* representing the proportion of an individual genome that originates from each pool. Using `STRUCTURE`, ancestry coefficients are estimated without prior knowledge on geographic proximity among individuals.

The approach implemented in `STRUCTURE` has been substantially improved by a number of approaches that include spatial proximity information based on individual geographic coordinates (reviewed by François and Durand (2010)). Among those spatially explicit approaches, the computer program `tess` is one of the most frequently used algorithms (Chen, Kim, et al., 2011; François, Ancelet, et al., 2006). In the `tess` model, ancestry proportions are continuously distributed over geographic space, and the parameters that specify the shape of the clines are estimated from the genetic and the geographic data. Using geographic information, `tess` provides better estimates of ancestry coefficients than `STRUCTURE` when the levels of ancestral population divergence are low (Durand et al., 2009).

The Bayesian approaches implemented in `STRUCTURE` and `tess` rely on Markov Chain Monte Carlo algorithms. Monte Carlo algorithms are based on computer intensive stochastic simulations, and have the advantage of sampling the posterior distribution of the model parameters. However the application of stochastic algorithms can be difficult when the data include more than a few hundreds of individuals or a few

thousands of allelic markers. With the availability of next generation sequencing data, there is a need to analyze genotypic matrices that represent thousands of individuals and hundreds of thousands of markers. While fast versions of `STRUCTURE` have already been proposed (Raj et al., 2014; Frichot, Mathieu, et al., 2014; Alexander and Lange, 2011; Wollstein and Lao, 2015), developing fast and accurate estimation algorithms for ancestry coefficients in a geographic framework remains an important computational challenge.

In this chapter, we present the study published for us in Caye et al. 2016; a spatially explicit algorithm that provides fast estimation of ancestry coefficients with accuracy comparable to `tess` 2.3 (Durand et al., 2009). The new algorithms are based on least-squares optimization and on geographically constrained non-negative matrix factorization (Cai et al., 2011; Frichot, Mathieu, et al., 2014). These improvements of `tess` are implemented in the computer program `tess3`. We show that `tess3` is substantially faster than `tess` 2.3, with an increase in computational speed of one or two orders of magnitude. In addition, we show that ancestral allele frequencies are correctly estimated, and we illustrate the use of the `tess3` program to perform genome scans for selection based on ancestral allele frequency differentiation. To illustrate our approach, `tess3` was applied to genomic data from European lines of the model species *Arabidopsis thaliana* for which an individual-based sampling design was available (Atwell et al., 2010).

3.3 Material and Methods

The computer program `tess3` computes ancestry estimates for large genotypic matrices using the geographic coordinates of sampled individuals. The program also returns locus-specific estimates of ancestral genotypic frequencies, and computes locus-specific estimates of a population-based differentiation statistic that can be used in genome scans for adaptive alleles. The `tess3` program is particularly suited to the analysis of large genomic data sets, for which the number of loci (L) ranges between thousands to hundreds of thousands genetic polymorphisms and the number of individuals (n) ranges between hundreds to thousands individuals.

3.3.1 Input data

`tess3` requires that the data consists of n multi-locus genotypes and two geographic coordinates for each genotype. A genotypic matrix, X, records allelic data for each individual (i) and each locus (ℓ). With data representing single nucleotide polymorphisms (SNPs), the genotypic matrix records the number of derived or mutant alleles at each locus. Considering autosomes in a diploid organism, the genotype at locus ℓ corresponds to the number of derived alleles at this locus, which is encoded as an integer number 0, 1 or 2. For SNPs, the `geno` format is accepted by the program, which can also process other types of allelic data, such as short tandem repeats or amplified fragment length polymorphisms. Geographic coordinates can be expressed using several coordinate systems, for example longitude and latitude, and they are provided to the software in a separate input file.

3.3.2 Geographically constrained least-squares estimates of ancestry coefficients

Similarly to `tess` 2.3 or `STRUCTURE`, `tess3` supposes that the genetic data originate from the admixture of K ancestral populations, where K is unknown. `tess3` estimates a Q-matrix, $Q = (Q_{ik})$, which represents the individual ancestry coefficients ($n \times K$ dimensions), and a G-matrix, $G = (G_{k\ell}(j))$, which represents the ancestral genotypic frequencies. The dimension of G is equal to $K \times (p+1)L$ where p is the ploidy of the studied organism genome. The ancestry coefficient Q_{ik} is the fraction of individual i's genome that originates from the ancestral population k, and the coefficient $G_{k\ell}(j)$ represents the frequency of genotype j at locus ℓ in population k.

The principle underlying the `tess3` algorithm differs from the likelihood methods implemented in `STRUCTURE` or in `TESS` 2.3, and it can be considered to be model-free. The main idea is that the probability that an individual i carries the genotype j at locus ℓ is determined by the *law of total probability*

$$P(X_{i\ell} = j) = \sum_{k=1}^{K} Q_{ik} G_{k\ell}(j)\,.$$

The above formula establishes that each individual genotype is sampled from K

pools of ancestral genotypes, and that the sampling probabilities correspond to their admixture coefficients. The formula is equivalent to the factorization of the genotypic probability matrix, P, using the matrices Q and G as factors (Frichot, Mathieu, et al., 2014). In the `tess3` algorithm, probabilities are replaced by zero/one values depending on the absence or the presence of each genotype at each locus, and the resulting matrix is denoted by $\tilde{X}$. Estimates of Q and G are obtained by factorizing $\tilde{X}$ as follows $\tilde{X} = \hat{Q}\hat{G}$. Matrix factorization is performed according to a least-squares minimization algorithm (see Appendix). During the miminization process, spatial constraints are introduced to ensure that individuals that are geographically close to each other are more likely to share the same ancestral genotypes than individuals that are far apart. A regularization parameter, α, controls the regularity of ancestry estimates over the geographic space. Large values of α imply that ancestry coefficients have similar values for nearby individuals, whereas small values produce results close to `STRUCTURE`. The least-squares method leads to algorithms that are substantially faster than the Bayesian algorithms implemented in other programs. In addition, the approach makes no assumptions about linkage or Hardy-Weinberg equilibrium (HWE). The above framework is thus appropriate to deal with departures from HWE created by inbreeding or geographically restricted mating.

3.3.3 Number of populations

In `tess3`, the number of ancestral populations, K, is chosen after the evaluation of a cross-entropy criterion for each K (Frichot, Mathieu, et al., 2014). The choice of K is then based on a cross-validation method that partitions the genotypic matrix entries into a training set and a test set in which 5% of all entries are masked to the algorithm. The cross-entropy criterion compares the genotypic frequencies predicted from the training set to those computed from the test set at each locus. Smaller values of the criterion often indicate better estimates for `tess3`. In practice, the best choice for K corresponds to a plateau in the cross-entropy plot (Frichot and François, 2015)..

3.3.4 Outlier locus tests

In addition to the inference of spatial population structure, `tess3` can perform genome scans for selection when the program is applied to large genomic data sets. More specifically, `tess3` uses the ancestral genotype frequency matrix, G, to derive the allele frequencies in the K ancestral populations. Then the algorithm evaluates a locus-specific F_{ST}-statistic based on the estimated ancestral allele frequencies. Using standard population genetic theory, F_{ST}-statistics can be transformed into squared z-scores, and p-values can be computed using a chi-square distribution with $K-1$ degrees of freedom (Weir, 1996). To correct for the test inflation statistic due to neutral population structure, the z-scores were recalibrated using estimates of the inflation factor. Here, inflation factors were determined using an "empirical-null hypothesis" approach. The values of the inflation factor were determined graphically on the basis on quantile-quantile plots of p-values. This approach is less conservative than the method based on the median of the chi-square distribution with $K-1$ degrees of freedom (Devlin and Roeder, 1999; Frichot and François, 2015). Multiple testing issues were addressed by applying the Benjamini-Hochberg algorithm to the recalibrated p-values (Benjamini and Hochberg, 1995).

3.3.5 Simulated data sets and program runs

We created simulated data sets containing 200 admixed genotypes with levels of ancestry that varied continuously across geographic space. To generate the data, we used the computer program `MS` to perform coalescent simulations of neutral and outlier SNPs under island models with two populations (Hudson, 2002). One hundred genotypes were sampled from each source population, and admixed genotypes were created according to a longitudinal gradient of ancestry (Durand et al., 2009; François and Durand, 2010). Individuals at each extreme of the longitunal range were representative of ancestral populations, while individuals at the center of the range shared intermediate levels of ancestry in the two source populations. The number of loci was varied in the range $L = 1\text{k-}50\text{k}$ SNPs.

Our first series of simulations considered selectively neutral SNPs and used mi-

gration parameters, $M = 4mN_e$, between $M = 0.01$ and $M = 10$. The population differentiation statistic, F_{ST}, ranged from 0.007 to 0.42. Our second series of simulations included a proportion of outlier SNPs equal to 5%. Outlier loci were generated using two values of the effective migration rate $4m_sN = 0.1$ and $4m_sN = 1$. In simulations with outlier loci, the neutral migration rate was set to the value $4mN = 20$. The justification for using neutral migration-drift equilibrium models for simulating selection is that loci with selection have an effectively reduced migration rate, as compared to the neutral migration m in migration-selection-drift equilibrium models (Bazin et al., 2010).

The simulated data were used to compare `tess3` estimates to those of `TESS` 2.3 (Durand et al., 2009). The number of ancestral populations ranged from $K = 1$ to $K = 6$. Each run was replicated five times for each computer program. The number of cycles in the Markov chain Monte Carlo algorithm of `TESS` 2.3 was set to 1,000, and the optimal number of ancestral population was determined using the deviance information criterion. All other parameters were set to their default values. Statistical errors were measured as root mean squared errors (RMSE) between the estimated Q-matrix and the matrix of coefficients (Q^0) that were used to generate the data

$$\text{RMSE} = \left(\frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} (Q_{ik} - Q^0_{ik})^2 \right)^{1/2}.$$

A similar RMSE criterion was defined for comparing the estimates of G matrices obtained from `tess3` or `TESS` 2.3 to the estimates of the ancestral genotypic frequency matrix resulting from the coalescent simulations.

3.3.6 *Arabidopsis thaliana* data.

We applied `tess3` to genomic data from 170 European lines of the model plant *Arabidopsis thaliana* genotyped for 216k SNPs (Atwell et al., 2010). For these data, we determined the number of ancestral populations using the cross-entropy criterion, and we computed ancestry estimates for the sample. The results were projected onto a map of the European continent using a raster file and `R` graphic functions (Jay, Manel, et al., 2012). We also used `tess3` to perform a genome scan for selection on

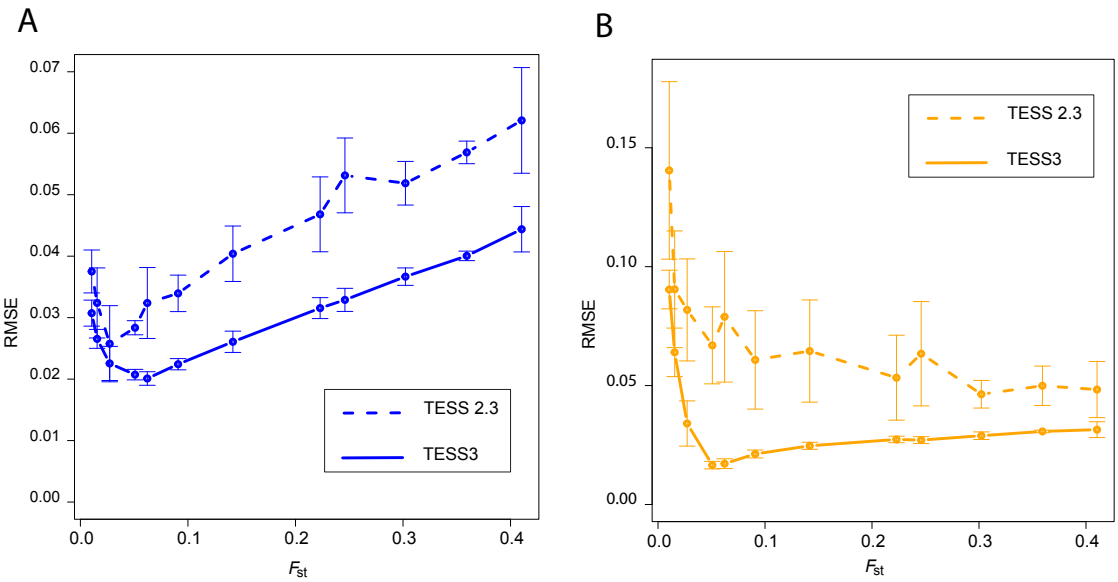


Figure 3.1 – **Statistical errors of `tess3` and `TESS 2.3` estimates.** Computer simulations of admixed populations using known individual ancestry proportions from two ancestral gene pools. (A) RMSEs of G estimates as a function of the level of ancestral population differentiation (F_{ST}). (B) RMSEs of Q estimates as a function of the level of ancestral population differentiation (F_{ST}).

chromosome 5 using $K = 3$ ancestral populations (54k SNPs).

3.4 Results

3.4.1 Comparison of ancestry estimates.

We used computer simulations of admixed populations to evaluate the ability of `tess3` to reproduce the ancestry estimates of `TESS` 2.3 using known individual ancestry proportions from two ancestral gene pools. Simulating 2k unlinked SNPs, we varied the level of ancestral population differentiation, measured by F_{ST}, to create difficult as well as easier data sets. For all data sets, the information criterion of each version of `TESS` led to $K = 2$ clusters. Statistical errors, measured by RMSEs for estimated Q and G matrices, ranged between 0.02 and 0.15 (Figure 3.1). Statistical errors increased as the levels of differentiation between the two source populations decreased, but they remained in an acceptable range for values of F_{ST} greater than 0.016. Overall, the statistical performances were of the same order for both versions of `TESS`.

3.4.2 Run-time analysis

Next we compared the run-times of `tess3` and `TESS 2.3` for increasing values of the number of ancestral populations and increasing numbers of loci. For `TESS 2.3` the total number of cycles in the MCMC algorithm was set to 1,000, a value for which the Monte-Carlo sampler reached its equilibrium state. Run-times were averaged over distinct random seed values for each K and number of loci. For both algorithms, the run-times increased with the number of loci and with the number of ancestral populations (Figure 3.2). For $L = 10$k loci, `tess3` and `TESS 2.3` runs took less than 6 minutes on an Intel Xeon 2.40 GHz CPU. With $L = 50$k loci and $K = 5$ ancestral populations, `TESS 2.3` took on average 30 minutes to complete a single run, whereas the `tess3` average run-time was about 4 minutes.

3.4.3 Outlier locus tests

We evaluated the capacity of `tess3` to detect outlier loci on simulated data containing 5% of outlier loci. For each locus, we performed a population differentiation test based on the estimated ancestral allele frequencies. Although the ratios m_s/m took large values, the probability distributions of F_{ST} statistics computed from neutral and selected ancestral allele frequencies overlapped substantially. Thus the power of neutrality tests were expected to be low. For a data set with $m_s/m = 0.005$, the estimate of the genomic inflation factor was equal to $\lambda = 4.4$. For a data set with $m_s/m = 0.05$ this value was equal to $\lambda = 10.0$. After correction of the test statistic, the observed levels of the false discovery rate were close to their expected values. The power to reject the null hypothesis was lower when the intensity of selection was low (Table 3.1). For an expected FDR of $q = 0.1$, the power of the test was approximately equal to 60% for the higher selection rate and it was equal to 30% for the lower selection rate. The power values were close to those obtained when we applied outlier tests to the data before admixture. This experiment showed that the power to reject neutrality in continuous populations was similar to the power of traditional population differentiation tests applied to the discrete (ancestral) population data.

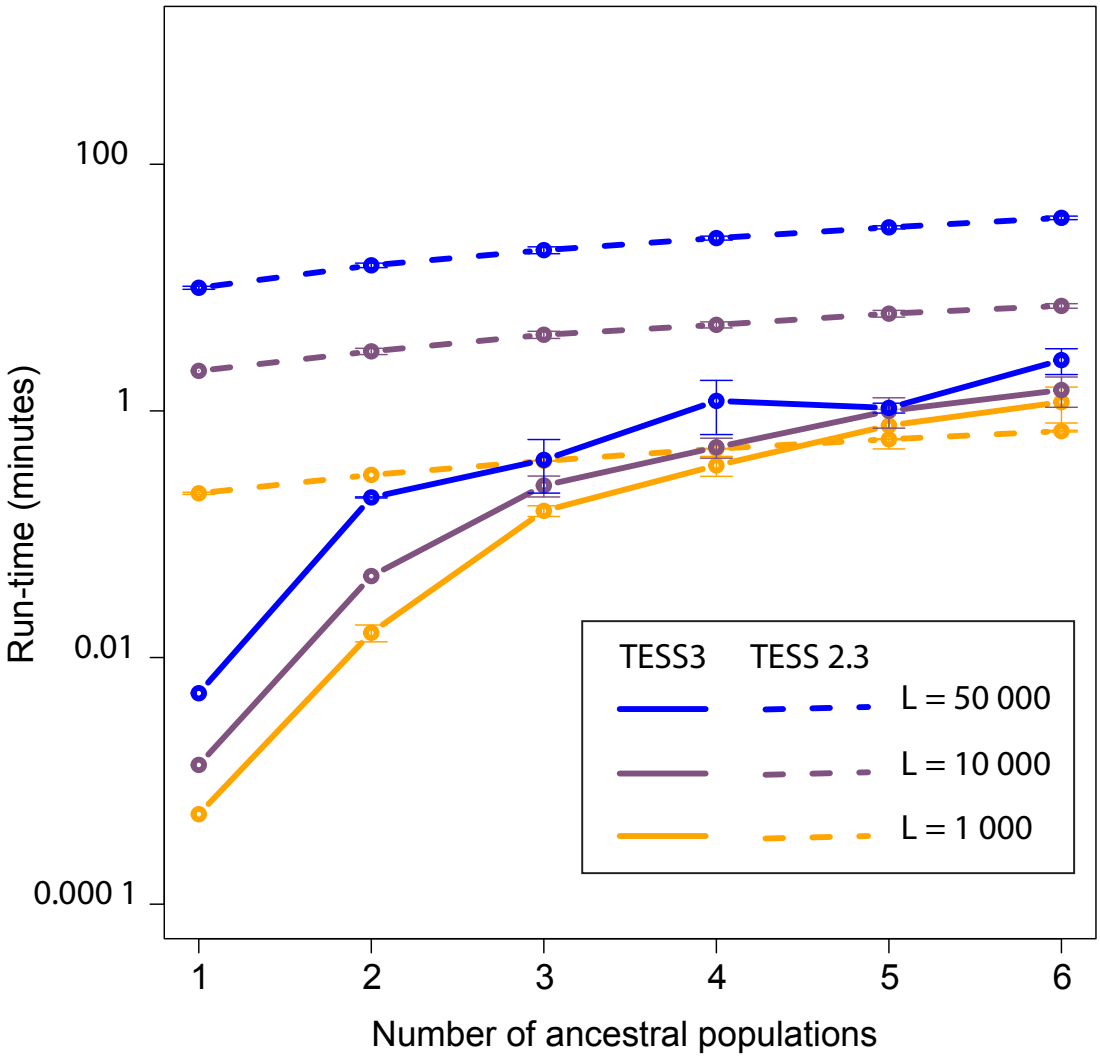


Figure 3.2 – **Run-times for `tess3` and `TESS 2.3`.** The number of ancestral population ranged between $K = 1$ and 6. Run-times were expressed in unit of minutes.

FDR	**Power**			
	After admixture	Before admixture	After admixture	Before admixture
0.05	**0.61**	0.63	**0.20**	0.26
0.10	**0.63**	0.66	**0.23**	0.29
0.15	**0.64**	0.67	**0.25**	0.32
0.20	**0.65**	0.69	**0.26**	0.33

Data set 1: $m_s/m = 0.005$. Data set 2: $m_s/m = 0.05$.

Table 3.1 – Power to reject neutrality of `tess3` outlier tests for two simulated data sets.

3.4.4 Biological data analysis

We applied `tess3` to a genomic data set of 170 European lines of *Arabidopsis thaliana* (216k SNPs). The cross-entropy curve exhibited a change in curvature for K = 3-4 clusters. For $K = 3$, the western cluster grouped all lines from the British Isles, France and Iberia. The eastern cluster grouped all lines from Central, Eastern Europe and Southern Sweden. Fourteen northern Scandinavian accessions were grouped into a separate population (Figure 3.3A). Those results were consistent with those obtained with `TESS 2.3`. The average run-time of `tess3` was about 5 minutes whereas each `TESS 2.3` run took about 2 hours. Then we performed a genome scan for selection based on population differentiation in the three ancestral populations detected by `tess3`. The genomic inflation factor was equal to $\lambda = 15.0$. The histogram of corrected p-values provided evidence that confounding errors were correctly removed (Figure 3.4A). The Manhattan plot exhibited islands of strong differentiation around positions 8,510 kb, 6,944 kb, 6,969 kb and 26,155 kb in the chromosome 5 (Figure 3.3]B). The top hits in the candidate list corresponded to genic SNPs. In particular, we discovered genes involved in defense response (*VSP1*), and in photoperiodism, flowering and root development (*WAV2*) (Mochizuki et al., 2005). The derived allele in the *VSP1* gene was present at high frequency in Eastern Europe and it was almost absent from Western Europe and Northern Scandinavia. The derived allele in the *WAV2* gene was present at high frequency in the Iberian peninsula and at low frequency in Eastern Europe and Northern Scandinavia (Figure 3.4B).

3.5 Discussion

A fundamental objective of evolutionary biology is the evaluation of the distribution of genetic variation among populations in geographic space. During the last few years, high-throughput sequencing technologies have allowed population geneticists to make fast progress in this direction. The access to extensive data have opened the door to a deeper understanding of the spatial distribution of adaptive and nonadaptive genetic variation in model and non-model organisms (Manel, Joost, et al., 2010). This transition from population genetics to population and ecological genomics is

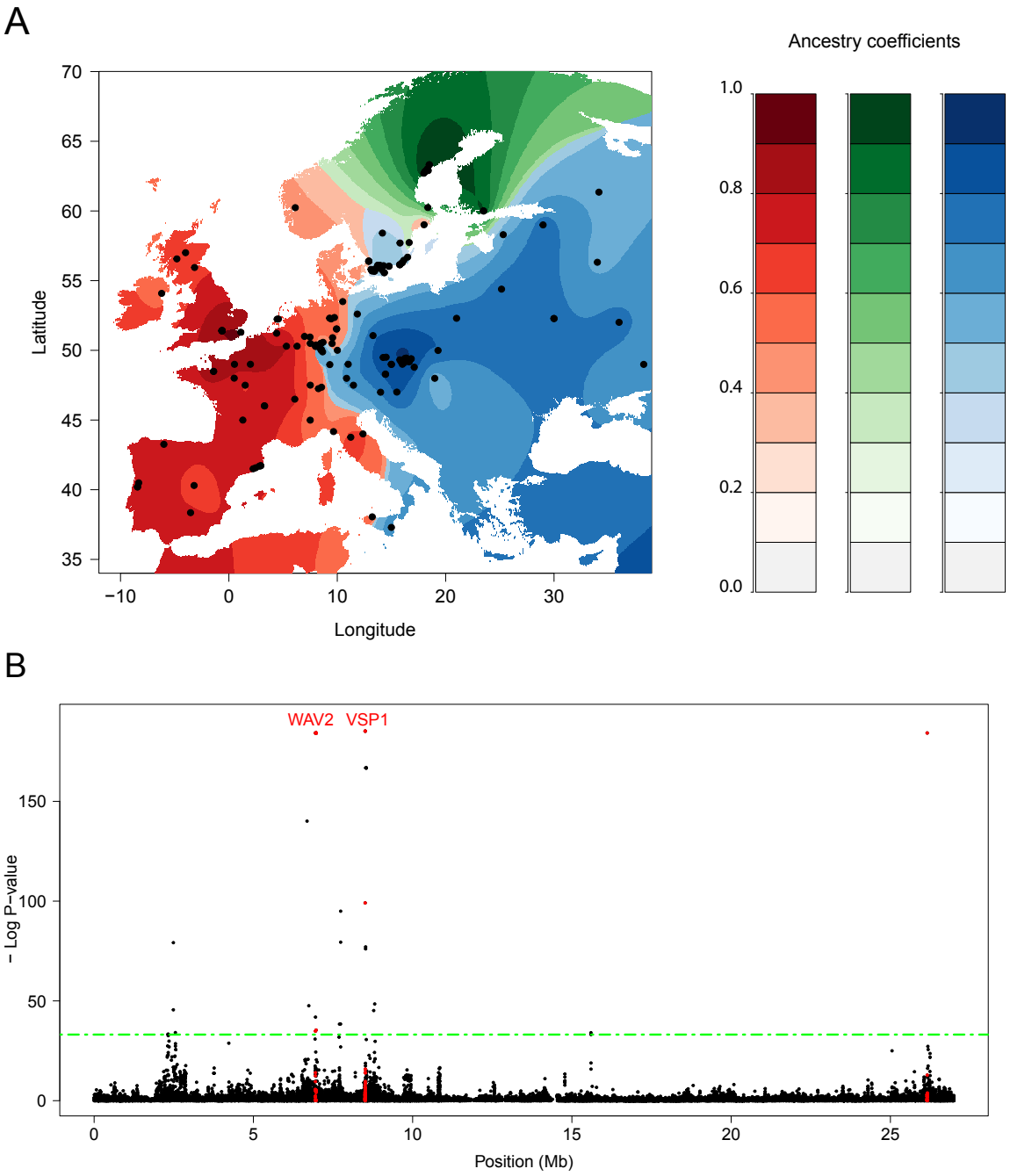


Figure 3.3 – **Results of the** *Arabidopsis thaliana* **data analysis with `tess3`.** A) Geographic maps of ancestry coefficients using $K = 3$ ancestral populations. B) Manhattan plot of $\log_{10}(p\text{-values})$ for the plant chromosome 5. The horizontal line corresponds to an expected FDR value of $q = 10^{-30}$.

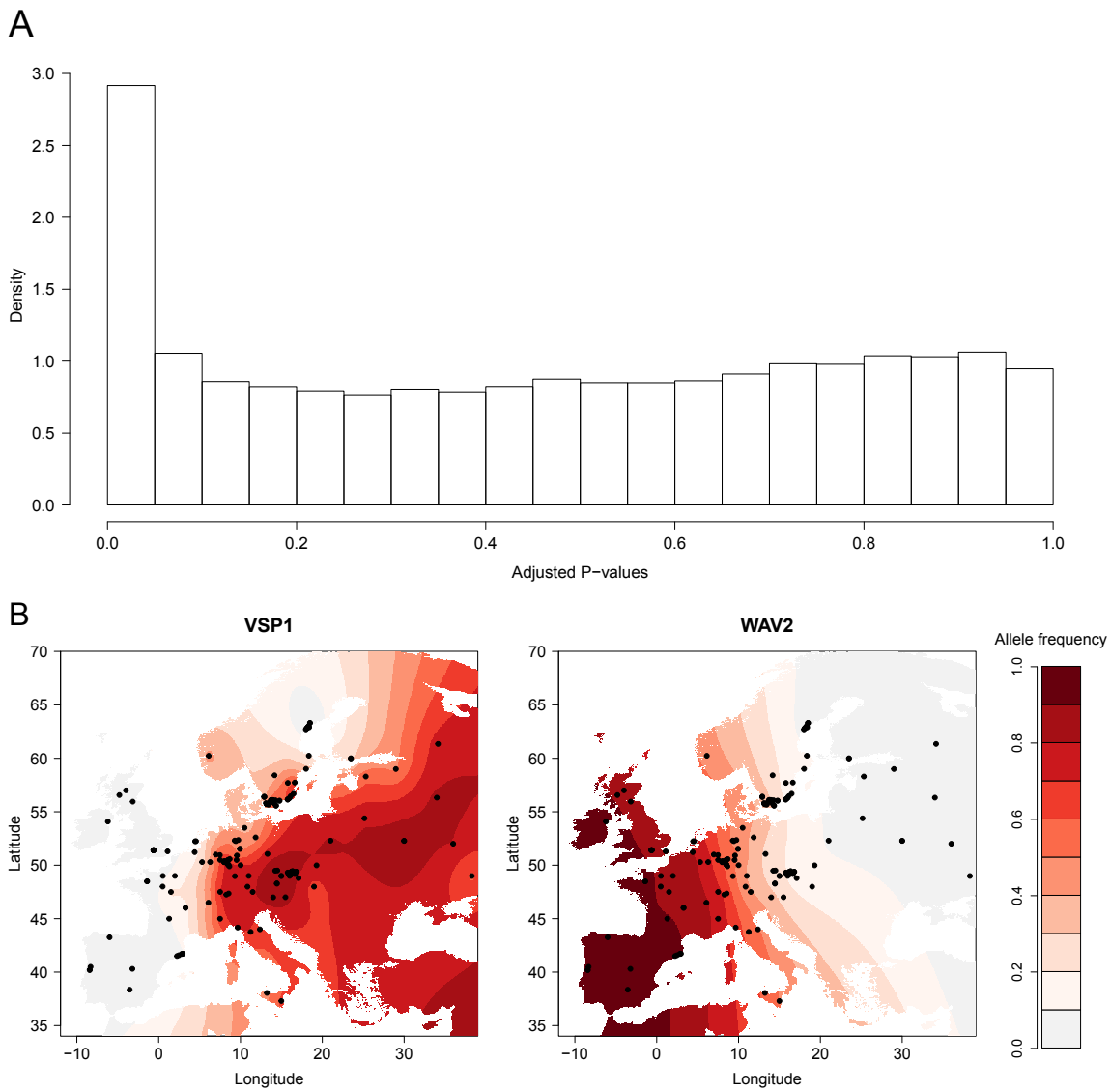


Figure 3.4 – **Candidate SNPs from a genome scan of the** *A. thaliana* **5th chromosome.** A) Histogram of adjusted *p*-values. B) Spatial distribution of allele frequency for two top-hit SNPs located in the *VSP1* and *WAV2* genes.

accompanied by a revolution of the principles and methods used to analyze the influence of landscape features on genetic variation. This revolution is made possible thanks to the availability of fast computing programs than can deal with high dimension and heterogeneity in the data.

By combining matrix factorization and spatial statistical methods, the computer program `tess3` enabled fast analysis of geographic and genome-wide patterns of genetic variation from large genomic data sets. In coalescent simulations of individuals with known ancestry, `tess3` produced accurate estimates of ancestry coefficients and ancestral allele frequencies. `tess3` results were statistically similar to those obtained with the Bayesian clustering program `TESS` 2.3, but `tess3` was about 30 times faster than `TESS` 2.3 when used with $K = 5$ ancestral populations and 50k binary loci. Though Bayesian approaches might be preferable for genotypic matrices of moderate dimensions, `tess3` generally outperformed `TESS` 2.3 when more than a few thousands of markers were used.

A novelty of `tess3` is the identification of outlier loci from the genotypic matrix. An important property of `tess3` outlier tests is that they do not require predefined populations, and that can be applied to individual sampling designs. Based on the estimations of the ancestral allele frequency matrix, the `tess3` algorithm computes a population differentiation statistic estimating a fixation index for each locus. If local adaptation favors a particular allele in some ancestral populations, the population differentiation statistic at that locus will be larger than at loci that are selectively neutral. Outliers in the distribution of the population differentiation statistic are usually considered as loci potentially targeted by local selection (Holsinger and Weir, 2009). In addition, the program output allows population geneticists to determine candidate loci based on classical FDR control algorithms.

The study of European lines of *A. thaliana* illustrated the main steps of analysis using `tess3`. These steps can be summarized as follows: 1) Identifying the number of clusters using the cross-validation criterion, by launching multiple runs of the program for each value of K, 2) Displaying maps of ancestry coefficients using `R` scripts provided with the program, 3) Performing a genome scan for selection based on ancestral allele

frequency differentiation statistics. Results for *A. thaliana* suggested that clinal variation occurs along an East-West gradient separating two ancestral populations in Central Europe. Those results were in very good agreement with previous findings using `TESS` 2.3, although these findings were obtained with a different set of markers (François, Blum, et al., 2008). A genome scan for selection revealed contrasted patterns among European lines of *A. thaliana* and provided evidence of a substantial role for natural selection in shaping the genome-wide variation of the plant species in Europe.

To conclude, the computer program `tess3` provides a major update of the `TESS` program enabling rapid ancestry coefficient estimation and genome scans for adaptive alleles. While preserving the accuracy of `TESS` 2.3, the least-squares algorithms of `tess3` ran substantially faster than the Bayesian algorithms of `TESS` when analyzing large population genomic data sets.

3.6 Data Accessibility

Installing `tess3`. Source codes, installation files and program documentation are available from Github

`https://github.com/cayek/tess3`.

The Atwell et al. (2010) data used in this study are publicly available from the following link:

`https://github.com/Gregor-Mendel-Institute/atpolydb`

3.7 Acknowledgements

This work has been partially supported by the LabEx PERSYVAL- Lab (ANR-11-LABX-0025-01) funded by the French program Investissement d'avenir. Helena Martins acknowledges support from the "Ciências sem Fronteiras" scholarship program from the Brazilian government. Olivier François acknowledges support from Grenoble INP and from the "Agence Nationale de la Recherche" (project AFRICROP ANR-13-BSV7-0017).

3.8 Appendix

This section provides a detailed description of the `tess3` algorithm. The first step of the algorithm builds a nearest-neighbor graph based on the geographic coordinates of the sampling sites. The number of neighbors in the graph was set to represent 5% of total connections. Then, the program runs a least-squares minimization algorithm. In this approach, the estimates of Q and G are obtained after solving the following constrained least-squares problem (Cai et al., 2011)

$$(\hat{Q}, \hat{G}) = \arg\min \mathrm{LS}(Q, G)\,,$$

where

$$\mathrm{LS}(Q, G) = \|\tilde{X} - QG\|_{\mathrm{F}}^2 + \alpha \sum_{s_i \sim s_j} w_{ij} \|Q_{i\cdot} - Q_{j\cdot}\|^2\,, \tag{3.1}$$

and Q and G are non-negative matrices such that, for all i and ℓ, we have

$$\sum_{k=1}^{K} Q_{ik} = 1 \qquad \sum_{j=0}^{p} G_{i\ell}(j) = 1\,.$$

In this equation, $\|M\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix M, $\|V\|$ is the Euclidean norm of a vector V, α is a non-negative *regularization parameter*. The summation on the right-hand side of the second term runs over all pairs of sites, $s_i \sim s_j$, sharing an edge in the nearest-neighbor graph. The quantity w_{ij} is a weight that decreases with geographic distance between sampling sites as follows

$$w_{ij} = \exp(-d(s_i, s_j)^2 / \bar{d^2})\,, \tag{3.2}$$

where d is the Euclidean distance, and $\bar{d}$ is the average distance computed over the neighboring sites in the sample. More specifically, the weight of an edge in the nearest-neighbor graph is related to the Laplace-Beltrami operator on a manifold (Belkin and Niyogi, 2003). In the algorithm, the regularization parameter α is equal to $c \times nL(p+1)/\sum w_{ij}$. The default value of c is 0.1%.

Least squares minimization is performed using the Alternating Non-negativity-constrained Least Squares (ANLS) algorithm with the active set (AS) method following

the approach used in the computer program `sNMF` (Frichot, Mathieu, et al., 2014; Kim and Park, 2011). The ANLS-AS algorithm starts with the initialization of the Q matrix, and then computes a non-negative matrix G that minimizes the quantity

$$\mathrm{LS}_1(G) = \|X - QG\|_F^2 .$$

The obtained solution is normalized so that its entries satisfy the probabilistic constraints for genotypic frequencies. Given G, the Q-matrix is computed after minimizing the following quantity

$$\mathrm{LS}_2(Q) = \left\| \begin{pmatrix} \mathrm{Vec}(\tilde{X}^T) \\ 0 \end{pmatrix} - \begin{pmatrix} \mathrm{Id} \otimes G^T \\ \sqrt{\alpha}\, \Gamma \otimes \mathrm{Id} \end{pmatrix} \mathrm{Vec}(Q^T) \right\|_F^2 ,$$

where $\mathrm{Vec}(\tilde{X})$ denotes the vectorization of the matrix $\tilde{X}$ formed by stacking the columns of $\tilde{X}$ into a single column vector, Γ is the Cholesky decomposition of the graph Laplacian associated with the weights of the graph (Chung, 1997), Id is the identity matrix, and $\otimes$ is a symbol for the Kronecker product. Iterations are stopped when the relative difference between two successive values of $\mathrm{LS}(Q, G)$ is lower than a tolerance threshold of ϵ. The default value for ϵ equals 10^{-7}.

Chapter 4

Influence of linkage disequilibrium and LD-pruning methods in genome scans for selection

Abstract

Linkage disequilibrium, defined as the non-random association of alleles at two or more loci, is essential in population genetics studies and can, for example, provide clues about past events and potential response to selection, and carry important information about population history. Although LD is widely used to provide insight into evolutionary history and for mapping genes in humans and other species, it remains a confounding factor in genome-wide association studies. Considering this problem, investigating how to deal with LD is a strategy for quality control and quality assurance (QC/QA) for genotypic data. Finding genetic signatures of local adaptation is of great interest for many population genetic studies. In this study, we investigated the effects of linkage disequilibrium and LD-pruned methods in genome scans for selection. We conducted simulations of data with three different levels of recombination rate and in the presence or absence of admixed individuals. Trying to correct the impact of linkage disequilibrium in our data, we applied an LD-pruned method using the toolset `PLINK`. Then, we compared the results of tests for selection using a new F_{ST} approach applied with the `snmf` software, with `pcadapt` and the classic F_{ST} statistic. We showed that LD could influence the results of analysis on data with admixed individuals and increase the false discovery rate. Thus, our findings

reiterate the importance of LD investigation in genome scans for selection and, that pruning data is necessary when studying a population that consists of overlapping populations.

4.1 Introduction

During the last years, denser single-nucleotide polymorphisms (SNPs) have been used in genome-wide association studies. Dense genotyping can introduce Linkage Disequilibrium (LD) (Snelling et al., 2017). LD, defined as the non-random association of alleles at two or more loci, is essential in population genetic studies because of many factors. LD can, for example, provide clues about past events and potential response to selection (Hedrick, 2011; Qanbari et al., 2010). As well, when it occurs throughout the genome, LD can carry information about population history, breeding system and geographic patterns of subdivision. When observed in genomic regions, LD reflects the history of natural selection, gene conversion and forces that cause gene-frequency evolution (Slatkin, 2008).

Although LD is widely used to provide insight on evolutionary history and for mapping genes in humans and other species, it remains a confounding factor in kinship and heritability estimation and in principal component analysis (PCA) (Charles et al., 2014). Elevated levels of LD can make some regions of the genome to be overrepresented in the principal components (PCs), distorting the estimation of population substructures (Abdellaoui et al., 2013). Moreover, the problems caused by LD in PCA, the assumption that markers are in linkage equilibrium (LE) may cause apparent over-sharing of multipoint identity by descent (IBD) among affected sibs resulting in false-positive evidence for linkage (Huang et al., 2004). Besides, estimation of narrow-sense heritability, h^2, can be highly sensitive to uneven linkage disequilibrium (LD) between SNPs: contributions to h^2 are overestimated from causal variants in regions of high LD and are underestimated in regions of low LD (Speed et al., 2012). Considering these problems, investigating how to deal with LD is a strategy for quality control and quality assurance (QC/QA) for genotypic data (Laurie et al., 2010).

When considering genome scans for selection, Reed et al. (2015) suggest the use of linkage disequilibrium pruning methods as a way to eliminate redundancy in the data and to reduce the influence of chromosomal artefacts (Laurie et al., 2010). LD pruning methods sequentially scan the genome for nearby SNPs in linkage disequilibrium, performing pairwise thinning based on a given threshold of correlation (Privé et al., 2017). All pairs of SNPs are compared with each other in a moving window. If one pair of markers inside the window is in LD greater than the specified threshold, the SNP with higher minor allele frequency (MAF) is kept. If the two MAFs are identical, the first SNP is kept. LD pruning methods control the quality of genotypic data for IBD analysis and PCA, for ancestry filtering, and results in large computational saving (Laurie et al., 2010). LD pruning can be carried out by using the toolset `PLINK` (Purcell et al., 2007).

The goal of our study is to analyse the influence of LD and LD- pruning methods in genome scans for selection in admixed individuals. For that, we performed a test for selection in genetic data using an F_{ST} definition based on the computation of ancestry coefficients and ancestral allele frequencies. We computed our statistic using the program `snmf`. Scans for selection were performed in simulated data with different levels of linkage disequilibrium, comparing results in the presence and absence of admixed individuals. A LD-pruning method was applied to exclude LD of the data. To analyse LD-pruning method can influence the number of a false regions highlight, we conducted our genome scan for selection on the pruned data sets. We compared our approach with genome scans using the software `pcadapt` and the standard F_{ST} statistic.

4.2 Methods and Materials

4.2.1 Statistical approaches for local adaptation scans

Methods to identify loci under selection

In the following paragraphs, we will describe the three methods considered in this study to identify loci under selection in genome data in the presence of linkage disequilibrium.

Definition of F_{ST} with admixed individuals. The first applied method is a new F_{ST} test that makes viable the application of F_{ST} statistic in genomic scans for populations containing admixed individuals, and for which no subpopulations can be defined a priori. Next, we will give a brief description of this method that was presented in Chapter 2.

Suppose that a population contains admixed individuals, and the population source is unknown. Considering K, the number of ancestral populations, the individual ancestry coefficients, Q, and the ancestral population frequencies, F, obtained from an ancestry estimation algorithm such as `structure`, are used to compute single-locus estimates of a population differentiation statistic F_{ST}, as follows

$$F_{ST}^{Q} = 1 - \frac{\sum_{k=1}^{K} q_k f_k (1 - f_k)}{f(1 - f)}, \tag{4.1}$$

where q_k is the average value of the k^{th} ancestry coefficient over all individuals in the sample, $q_k = \sum_{i=1}^{n} q_{ik}/n$, f_k is the ancestral allele frequency in population k at the locus of interest, and $f = \sum_{k=1}^{K} q_k f_k$ (Martins et al., 2016).

The locus-specific statistics are used to perform statistical tests of neutrality at each locus, by comparing the observed values to their expectations from the genome-wide background. In this framework, the test is based on the z^2-score statistic defined as follows,

$$z^2 = (n - K) F_{ST}/(1 - F_{ST}). \tag{4.2}$$

Assuming random mating at the population level, we have

$$z^2/(K - 1) \sim F(K - 1, n - K)$$

,

where $F(K - 1, n - K)$ is the Fisher distribution with $K - 1$ and $n - K$ degrees of freedom. Also, we assume that the sample size is large enough to approximate the distribution of squared z-scores as a chi-squared distribution with $K - 1$ degrees of freedom (Martins et al., 2016).

The statistic described above is implemented using scripts in R language and the computer programs `LEA` and `snmf`, with which the matrices Q and F are obtained.

`pcadapt` test statistic. The second method, used to perform a genomic scan for selection in genome data in the presence of LD, is the test statistic implemented by the latest version of the software `pcadapt` (Luu et al., 2017a). In the following paragraphs, we present this statistic.

Assuming that n is the number of individuals, p the number of genetic markers and G the genotype matrix that is composed of n lines and p columns, Luu et al. (2017a) considered multiple linear regressions, regressing each of the p SNPs on the K principal components as follows

$$G_j = \sum_{k=1}^{K} \beta_{jk} X_k + \epsilon_j, ; j = 1, \ldots, p, \tag{4.3}$$

where β_{jk} is the regression coefficient corresponding to the j-th SNP regressed by the k-th principal component, and ϵ_j is the residual vector. The result of the regression analysis for the j-th SNP, is summarised by returning a vector of z-scores $z_j = (z_{j1}, \ldots, z_{jK})$ where z_{jk} corresponds to the z-score obtained when regressing the j-th SNP by the k-th principal component.

To look for outliers based on the vector of z-scores, Luu et al. (2017a) considered a classical approach in multivariate analysis for outlier detection. The test statistic is a Mahalanobis distance D_j defined as

$$D_j^2 = (z_j - \bar{z})^T \Sigma^{-1} (z_j - \bar{z}), \tag{4.4}$$

where Σ is the $(K \times K)$ covariance matrix of the z-scores and $\bar{z}$ is the vector of the K z-score means (Maronna and Zamar, 2002).

Standard F_{ST} statistic. The third method applied in this chapter is based on a standard F_{ST} statistic. This statistic is based on the Wright's definition of F_{ST}, that corresponds to the amount of variance in allele frequency that can be explained by population structure (Wright 1951).

Consider a two-allele locus and define K as the number of populations, p_i as the allelic frequency of reference in population i, and $\bar{p}$ as the frequency of the reference allele across all populations. We calculate F_{ST} for a specific locus as follows,

$$F_{ST} = \frac{\frac{1}{K-1}\sum_{i=1}^{K}(p_i - \bar{p})^2}{\bar{p}(1-\bar{p})}, \tag{4.5}$$

where $\bar{p} = \frac{1}{K}\sum_{i=1}^{K} p_i$.

Following equation 4.5, F_{ST} can also be related to the genetic variance due to population structure. A classical definition for F_{ST} corresponds to the proportion of the genetic variation in sampled allele frequency, and it is defined as

$$F_{ST} = \frac{\sigma_S^2}{\sigma_T^2}, \tag{4.6}$$

where, σ_T^2 is the variance of the allelic state in the total population, and σ_S^2 is the variance in the frequency of the allele between different subpopulations (Weir, 1996). Therefore, F_{ST} value can be calculated using analysis of variance (ANOVA) of allele frequencies. `ANOVA` is a statistical method that tests whether the means of two or more groups are equal and can therefore be used to assess the degree of differentiation between populations (Holsinger and Weir, 2009).

Considering the standard population genetic theory, F_{ST}-statistics can be transformed into squared z-scores and p-values can be computed using a chi-squared distribution with $K-1$ degrees of freedom (see Equation 4.2, considered for the method described in paragraph 4.2.1).

Genomic Inflation Factor

In genome scans, there are confounding factors that inflate values of the test statistic and that could lead to an excess of false positives. To reduce the number of false positives, a calibration of the null-hypothesis is performed by using a genomic control method to adjust the test statistic for background levels of population structure (François, Martins, et al., 2016). For F_{ST}^{Q} and for standard F_{ST} statistics, we computed a genomic inflation factor value, defined by the median of the squared z-scores divided

by the median of a chi-squared distribution with $K-1$ degrees of freedom, where K is the number of ancestral populations (genomic control, Devlin & Roeder (1999)). For the `pcadapt` statistic, we considered the calibration provided by the software, that divide Mahalanobis distances by a genomic inflation factor, defined by Luu et al. (2017b) as the median of the Mahalanobis distances divided by the median of the chi-square distribution with K degrees of freedom, where K is the number of principal components. We used the Benjamini-Hochberg algorithm to control the false discovery rate after recalibrating the null-hypothesis (Benjamini and Hochberg, 1995).

4.2.2 Simulated datasets

Unadmixed population simulation model

We simulated genetic data considering a two-population model in which the populations evolved under migration-drift equilibrium (Wright's 2-island model). We made use of the computer program `ms` to perform simulations of neutral and selected loci (Hudson, 2002). In Wright's models, there is a global migration rate of individuals, which, in principle, could be estimated by using neutral markers, and an effective migration rate that reflects the action of selection to filter migrants having not well-adapted genotypes (Bazin et al., 2010). In this case, Petry (1983) defined the effective migration rate (m_s) which can be expressed from a relation between the migration rate for the neutral model (m), the strength of selection (s) and the recombination rate (r), as follows

$$m_s = m\frac{r}{s+r}.$$

Thus, we can define m_s/m and s/r as:

$$\frac{m_s}{m} = \frac{r}{s+r} \quad \text{and} \quad \frac{s}{r} = \frac{m}{m_s} - 1.$$

To create datasets with different strength of selection (s), we varied in the effective migration rate (m_s) and the neutral migration rate (m) values. Figure 4.1 shows the m_s/m ratio as a function of the strength of selection for various values of r. Note that the strength of selection (s) is inversely proportional to the value of m_s/m.

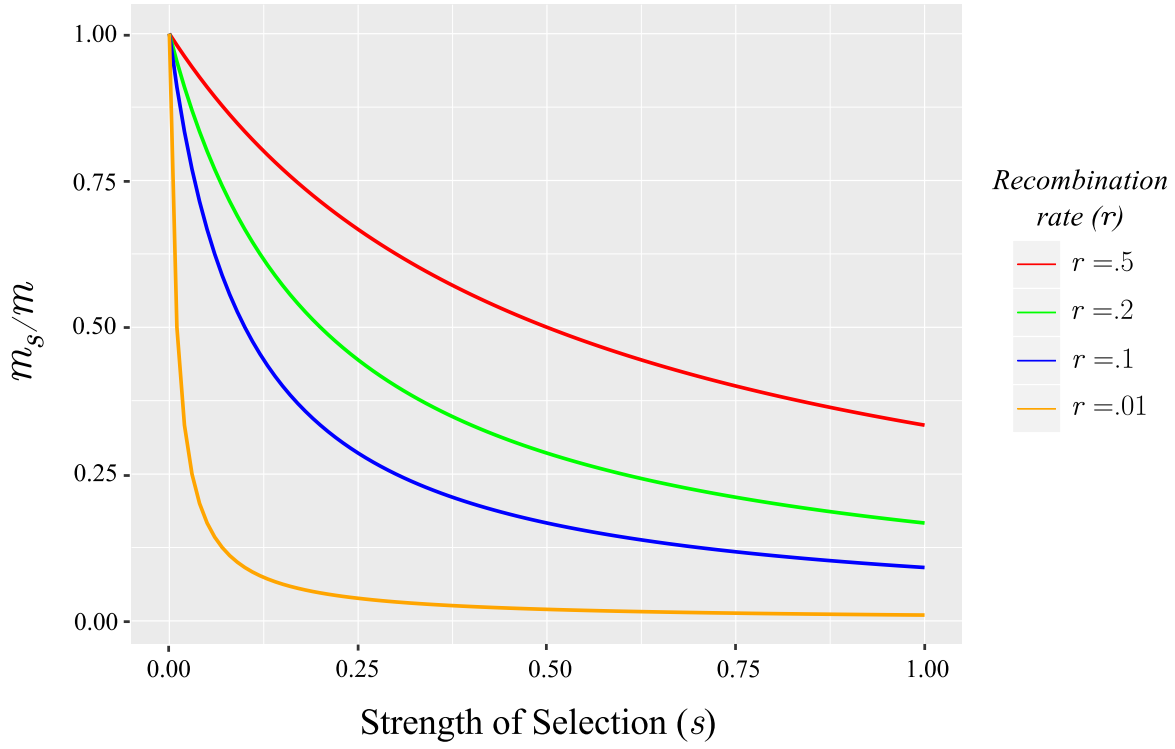


Figure 4.1 – The relation between effective migration and migration for the neutral model (m_s/m) as a function of the strength of selection (s) for various values of r.

In our simulation model, we varied the m_s/m value by assuming $m_s = 1$ and $m = 2$, 5, 10, 15, 20 and 30. So, m_s/m values were equal to 0.5, 0.2, 0.1, 0.06, 0.05 and 0.03. The simulation model reproduces the reduced levels of diversity and the increased levels of differentiation expected under hard selective sweeps occurring at one particular chromosomal segment in one ancestral population. Thus, varying m_s/m and the recombination rate r, six datasets were created for low, medium and high levels of LD and different intensities of selection. In total, we considered eighteen datasets. In the next paragraphs, we will explain in more details the use of the `ms` software to generate simulated datasets. The description and values of the `ms` parameters considered in our models are found in Table 4.1.

Creating datasets using the `ms` software. With the `ms` software, the default command line to build simulated data under a two-island model and with cross-over (recombination) is

$$\texttt{ms nsam nreps -t } \theta \texttt{ -r } \rho \texttt{ nsites -I npop n1 n2 } [4N_0m],$$

where `nsam` is the number of copies of the locus in each sample, and `nreps` is the number of independent samples to generate. The third parameter, $\theta = (4N_0\mu)$, is the mutation parameter where μ is the mutation rate for the entire locus. To have a two-island model, we added the switch `-I` followed by the number of subpopulations, `npop`, and `n1` and `n2`, that indicate the number of chromosomes sampled from each subpopulation. Considering one symmetric island model, we entered the migration parameter $4N_0m$, where m is the migration rate and N_0 is the diploid population size (Hudson, 2004).

The recombination parameter is $\rho = 4N_0r$, where r is the recombination rate between the ends of the segment being simulated. Considering `nsites` as the number of base pairs in the locus, and C as the cross-over probability between adjacent base pairs, the recombination rate among the two ends of the locus is $r = C(\texttt{nsites} - 1)$. To generate simulated data evolving one chromosomal segment we assumed an effective population size of $N_0 = 10^6$ for each ancestral population and a mutation rate per bp M equal to 10^{-9}. The datasets are formed from the union of neutral loci and outlier

loci segments. For neutral segments, we considered 250100 as the number of base pairs in the locus (`nsites`).

Considering that the physical scale of linkage disequilibrium increases as the recombination rate decreases, we varied the value of r to simulate different levels of LD (Andolfatto & Wall, 2003). For a hight level of LD, we set $C_h = 10^{-9}$ and $r_h = 10^{-9}(250100 - 1) \cong 2.5 \times 10^{-4}$. For a medium level of LD, we used $C_m = 10^{-8}$ and $r_m = 10^{-8}(250100 - 1) \cong 2.5 \times 10^{-3}$. For a low level of LD, we considered $C_l = 10^{-7}$ and $r_l = 10^{-7}(250100 - 1) \cong 2.5 \times 10^{-2}$. Subsequently, as $\rho = 4N_0 r$, we used $\rho_h = 10^3$, $\rho_m = 10^4$ and $\rho_l = 10^5$ to generate neutral segments with respectively high, medium and low level of linkage disequilibrium. Finally, with $M = 10^{-9}$, we used a value μ of neutral mutation rate equal to

$$\mu = M \times (\texttt{nsites} - 1) = 10^{-9} \times (250100) = 0.25 \times 10^{-3}.$$

This led to, $\theta = 4N_0\mu = 1000.4$.

The description and values of parameters assumed for simulating neutral segments can be found in Table 4.1.

The following `ms` command is an example for one of our neutral simulations,

```
ms 200 1 -t 1000.4 -r 1000 250100 -I 2 100 100 20
```

Selected segments are created by considering `nsites`= 2501, 10% of number of sites for neutral loci. For recombination rates we used, $r_h = 2.5 \times 10^{-6}$, for a high level of LD, $r_m = 2.5 \times 10^{-5}$, for a medium level of LD and $r_l = 2.5 \times 10^{-4}$, for a low level of LD. In this case, the values of ρ used to create datasets with different levels of LD, were $\rho_h = 10$, for high LD, $\rho_m = 10^2$, for medium LD and $\rho_l = 10^3$, for low LD. Using $M = 10^{-9}$ as the mutation rate per bp, we set a value μ of mutation rate equal to

$$\mu = M \times (\texttt{nsites} - 1) = 10^{-9} \times (2500) \cong 0.25 \times 10^{-5}$$

.

The description and values of parameters used for simulating outlier segments can be found in Table 4.1.

Therefore, we considered $\theta = 10.004$ for the diversity of a selected segment. The follows `ms` command is an example for one of our outlier segments,

```
ms 200 1 -t 10.004 -r 10 2501 -I 2 100 100 1.
```

Admixed population simulation model

Using the unadmixed datasets described in previous paragraphs, we generated admixed genotypes. Our model for admixture was based on a gradual variation of ancestry proportions across geographic space. Geographic coordinates (x_i, y_i) are created for each individual from Gaussian distributions centred around two centroids put at a distance two on a longitudinal axis (Durand et al. 2009) (Figure 4.2). As it happens in a secondary contact zone, we assume that the ancestry proportions had a sigmoidal shape across space (Barton & Hewitt 1985),

$$p(x_i) = \frac{1}{1 + e^{-x_i}}.$$

For each individual, we assumed that each allele originated in the first ancestral population with probability $p(x_i)$ and in the second ancestral population with probability $1 - p(x_i)$ (Durand et al. 2009).

In those scenarios, individuals at each extreme of the geographic range were representative of their population of origin, while individuals at the centre of the range shared intermediate levels of ancestry in the two ancestral populations (Caye et al., 2016).

4.2.3 Software parameter definitions

In this subsection we described the parameters considered to perform genome scans for selection with the program: `snmf` (Frichot, Mathieu, et al., 2014) and `pcadapt` (Luu et al., 2017b; Duforet-Frebourg et al., 2015).

For the `snmf` program, we assumed $K = 2$ ancestral populations. The value of K was chosen based on the cross-entropy criterion. The cross-entropy criterion is based

Table 4.1 – Definition of parameters used to generate simulated dataset using `ms`

Parameter `ms`	**Description**	**Neutral Model**	**Outlier Model**
`nrep`	Number of independent samples to generate	1	1
`nsam`	Number of copies of the locus in each sample	200	200
`nsites`	Number of base pair in the locus	250100	2501
`npop`	Number of subpopulation	2	2
`n1`	Number of locus for population 1	100	100
`n2`	Number of locus for population 2	100	100
N_0	Diploid population size	10^6	10^6
θ	Mutation parameter: $\theta = 4N_0\mu$	10^{-9}	10^{-9}
ρ_h	Recombination parameter: $\rho_h = 4N_0 r_h$ (high level of LD)	10^3	10
ρ_m	Recombination parameter: $\rho_m = 4N_0 r_m$ (medium level of LD)	10^4	10^2
ρ_l	Recombination parameter: $\rho_l = 4N_0 r_l$ (low level of LD)	10^5	10^3
r_h	Recombination rate between the ends of the segment being simulated (high level of LD) : $r_h = C_h$(`nsites`-1)	2.5×10^{-4}	2.5×10^{-6}
r_m	Recombination rate between the ends of the segment being simulated (medium level of LD): $r_m = C_m$(`nsites`-1)	2.5×10^{-3}	2.5×10^{-5}
r_l	Recombination rate between the ends of the segment being simulated (low level of LD): $r_l = C_l$(`nsites`-1)	2.5×10^{-2}	2.5×10^{-4}
μ	mutation rate for the entire locus	0.25×10^{-5}	0.25×10^{-3}
C_h	Cross-over probability between adjacent base pairs (high level of LD)	10^{-9}	10^{-9}
C_m	Cross-over probability between adjacent base pairs (medium level of LD)	10^{-8}	10^{-8}
C_l	Cross-over probability between adjacent base pairs (low level of LD)	10^{-7}	10^{-7}
M	mutation rate per bp	10^{-9}	10^{-9}
m	migration rate	$m = 2;\ 5;\ 10;\ 15;$ 20 and 30	$m_s = 1/2; 1/5; 1/10;$ $1/15; 1/20$ and $1/30$

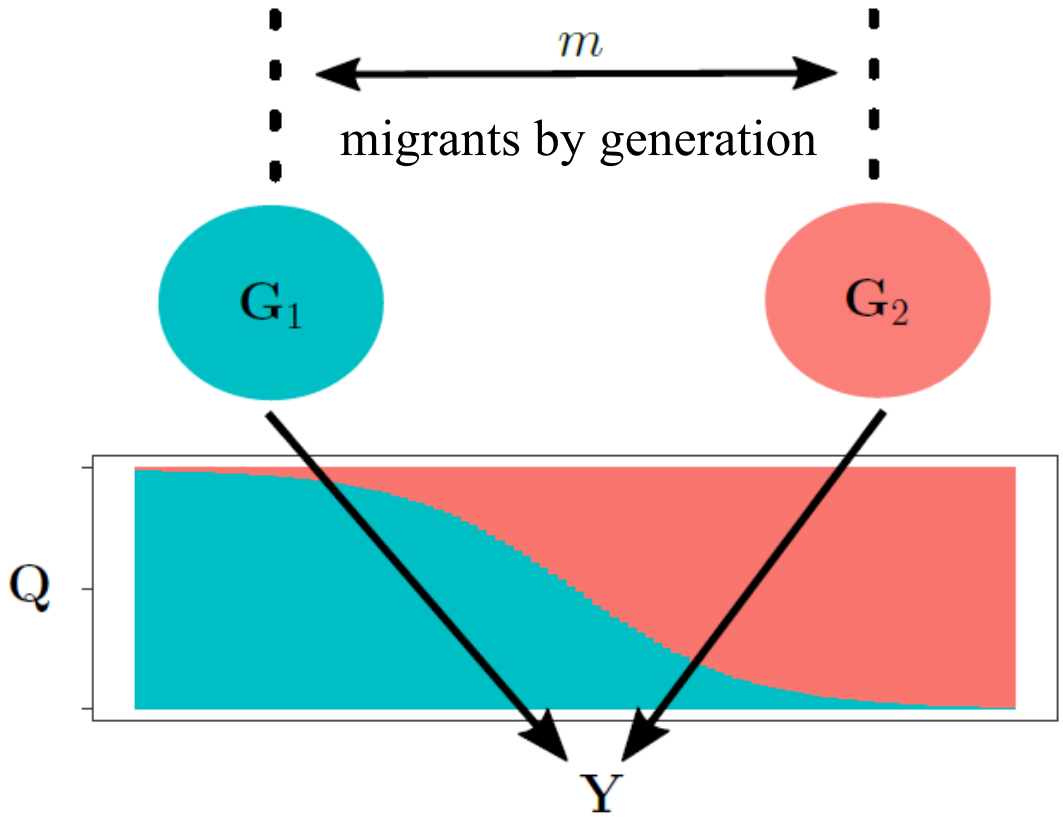


Figure 4.2 – **Simulation of spatially mixed genotypes**. The source populations are simulated using the `ms` program to obtain the frequency matrices of genotypes G_1 and G_2. The Y matrix is generated by pulling genes from both source populations with probabilities generated along a longitudinal gradient and stored in the matrix Q (Caye, 2017).

on the prediction of a fraction of masked genotypes (matrix completion), and on the cross-validation approach. Using the option `entropy=TRUE` in `snmf` and running the program for $K = 1 - 10$, we plotted the values of the cross-entropy criterion for each K. The value for which the function plateaus or increases was our estimate of K. The other parameters of `snmf` were set at their default internal values. Each run was replicated five times, and the run with the lowest cross-entropy value was selected for computing the F_{ST}^{Q} statistics according to formula 4.1. For `pcadapt`, we used $K - 1$ principal components. The statistical test underlying `pcadapt` and the standard F_{ST} statistic were explained in subsection 4.2.1.

4.2.4 LD-decay analysis

The standard measures of linkage disequilibrium, D and r^2, are respectively equivalent to the covariance and the correlation between alleles at two different loci. Consider two diallelic loci l and m, with alleles A and a at the locus l and alleles B and b at the locus m. In this case, there are four possible haplotypes: AB, Ab, aB and ab; with respective frequencies: p_{AB}, p_{Ab}, p_{aB} and p_{ab}. Therefore, the linkage disequilibrium coefficient between A and B, D_{AB}, can be calculated as follows

$$D_{AB} = p_{AB} - p_A p_B,$$

where p_A is the allele frequency of A and p_B is the allele frequency of B.

Assume X^l (X^m, for loci m) is the random variable equal to 1 when an individual carries the allele A at locus l (allele B at locus m) and 0 otherwise. Then, $D_{AB} = Cov(X^l, X^m)$. By definition, we set

$$r^2 = \frac{D_{AB}^2}{p_A(1-p_A)p_B(1-p_B)} = Cor(X^l, X^m)^2,$$

with $D_{AB}^2 = Cov(X^l, X^m)^2$, $Var(X^l) = p_A(1-p_A)$ and $Var(X^m) = p_B(1-p_B)$.

Usually, r^2 decays with the physical distance between loci and is analysed to determine the density of markers to use in whole genome association methods (Stram, 2004). To estimate the r^2 coefficient, we applied the `PLINK` software with its default

command line. We calculated the average r^2 value of each 1 kb region. LD decay was estimated for all datasets within a range of 100 kb.

4.2.5 LD-pruning method

When analysing genome-wide genetic variants, we should consider that some regions of the genome may be over represented in the PCs due to elevated levels of linkage disequilibrium (LD), diluting the genome-wide patterns that reflect ancestry differences. In this case, specific levels of LD can even result in analysis that only shows genetic variation at a specific locus (Abdellaoui et al., 2013). If not corrected, this bias could lead to a large number of SNPs hitchhiking and, consequently, increase FDR in genome scans for selection (Lucotte et al., 2016). To correct for such LD effects, we used a LD-pruning method.

Pruning is an algorithm that sequentially scans the genome for nearby SNPs in linkage disequilibrium, performing pairwise thinning based on a given threshold of correlation (Privé et al., 2017). All pairs of SNPs are compared with each other in a moving window. If one pair of markers inside the window is in LD greater than the specified threshold of correlation, the SNP with higher minor allele frequency (MAF) is kept. If the two MAFs are identical, the first SNP is kept. To perform the LD-based SNP pruning method we used the toolset `PLINK` (Purcell et al., 2007). In PLINK, we adopted the following command line:

```
plink -file DataX-indep-pairwise 50 5 0.2 -out prune.DataX,
```

where the window size in SNPs was equal to 50, the number of SNPs to shift the window at each step was 5 and the r^2 threshold was 0.2. After the pruning step, we thinned the datasets by keeping the markers in approximate linkage equilibrium with each other, for the case where the r^2 threshold was 0.2. We performed genome scans for selection in the pruned data using the `snmf` and `pcadapt` programs, and the standard F_{ST} statistic. The simulations were conducted using $K = 2$ ancestral populations in the `snmf` software and one principal component with the `pcadapt`

program.

4.2.6 False Discovery Rate (FDR) value by region

To verify the effect of submitting the data to an LD-pruning method, we used a False Discovery Rate (FDR) statistic. Since the number of false outliers ("false" regions of the genome) found in pruned data is reduced, we considered an FDR value observed by regions (FDR_w). We will define FDR_w in the next paragraphs.

For all simulated dataset, a list of possible outlier loci (candidate list) was obtained using the false discovery rate (FDR) control algorithms, proposed by Storey and Tibshirani (2003). We considered an expected FDR value of 10%. Let l be the list of candidates obtained and t the list of truly selected SNPs; the observed FDR value is obtained by using the following relation:

$$FDR = \frac{FP}{TP + FP},$$

where FP (False Positives) is the number of SNPs in l that are not in t and TP + FP is the length of l (TP meaning True Positives). The FDR value varies between zero and one, where zero is a considered as the lowest level of FDR and one as the highest level of FDR.

The statistical power of the tests can be calculated as follows

$$Power = \frac{TP}{TP + FN}, \tag{4.7}$$

where TP (True Positives) is the number of SNPs in l that are in t and TP + FN is the length of t (FN meaning False Negatives).

In our simulated datasets, the truly selected SNPs were positioned at the end of the genotype matrix Y. For example, consider a dataset with 5000 SNPs where 500 SNPs are under selection. The first 4500 columns of Y represented neutral SNPs while the last 500 columns represented genuinely selected SNPs. To obtain the value of FDR_w, we considered a window (region of the genome) based on the number of true positives. For example, in a dataset of 5000 SNPs where 500 SNPs are true positives,

we have ten windows of 500 SNPs, where the last window contains the truly selected loci.

Consider the windows $W = W_1, W_2, ..., W_z$, where z is the number of truly selected loci. Using the F_{ST} values computed in the equation 4.1, we obtained the p-values (p) relative to each SNP in Y. For each window W_i, we calculated $\tilde{p}_i = \min(p_j, j \in W_i)$, the p-values representative of the window. Then, we considered a new set of p-values $\tilde{S} = \tilde{p_1}, \tilde{p_2}, ..., \tilde{p_z}$. After, considering the p-values in $\tilde{S}$, we applied the Storey and Tibshirani's FDR control algorithm with an expected FDR value of 10%. So, we obtained a new list of candidates, called candidate windows.

Consider l_w the list of candidate windows and t_w as the list of true positive windows. The FDR_w can be obtain as follow

$$FDR_w = \frac{FP_w}{TP_w + FP_w},$$

where FP_w (False Positives windows) is the number of windows in l_w that are not in t_w and $TP_w + FP_w$ is the length of l_w (TP_w meaning True Positives windows). Since, in our simulated datasets, we have a single true positive window, TP_w is equal to one and, FDR_w can be defined as

$$FDR_w = \frac{FP_w}{1 + FP_w}.$$

Therefore, FDR_w will be equal to 0.5 when two windows (including the truly one) are selected, 0.66 when three windows are selected, 0.75 when four windows are selected and between 0.8 and 1 when more five windows are selected. Considering that the value of FDR_w depends only on the false positives windows, we used the FP_w parameter to analyse the accuracy of the tests for selection. Another consideration is that when all loci in the true positive window are highlighted for the test for selection, the number of false negatives will be equal to zero and thus, the test power equal to 1 (Equation 4.7).

Taking in account that FDR_w does not follow the same scale as the FDR, and considering that FDR_w equal to 0.5, 0.66 and 0.7 are relatively low levels of false discovery rates, we opted for dividing the observed FDR_w values by two. Thus, FDR_w

could follow a scale closer to standard FDR.

4.3 Results

4.3.1 LD decay in simulated datasets

To analyse the effects of linkage disequilibrium in our tests for selection, we studied datasets before and after admixture with three levels of recombination rate (ρ).

Considering data before and after admixture, Figure 4.3 shows the LD Decay analysis for these levels. We noticed that the level of linkage disequilibrium, represented as the mean of r^2, was larger when the data sets are admixed. As well, the analysis showed that the linkage disequilibrium increased as the recombination rate (ρ) decreased. For data before admixture, we obtained a mean r^2 between 0.05 and 0.3 for the low LD datasets, a mean r^2 between 0.1 and 0.4 for medium LD and a mean r^2 between 0.2 and 0.5 for high LD. For datasets after admixture, the values found for r^2 were slightly higher for the three levels of LD. For the datasets with low LD, mean r^2 varied between 0.1 and 0.35 sets, for the medium LD case, between 0.2 and 0.45 and between 0.3 and 0.6 for high LD for the high LD case.

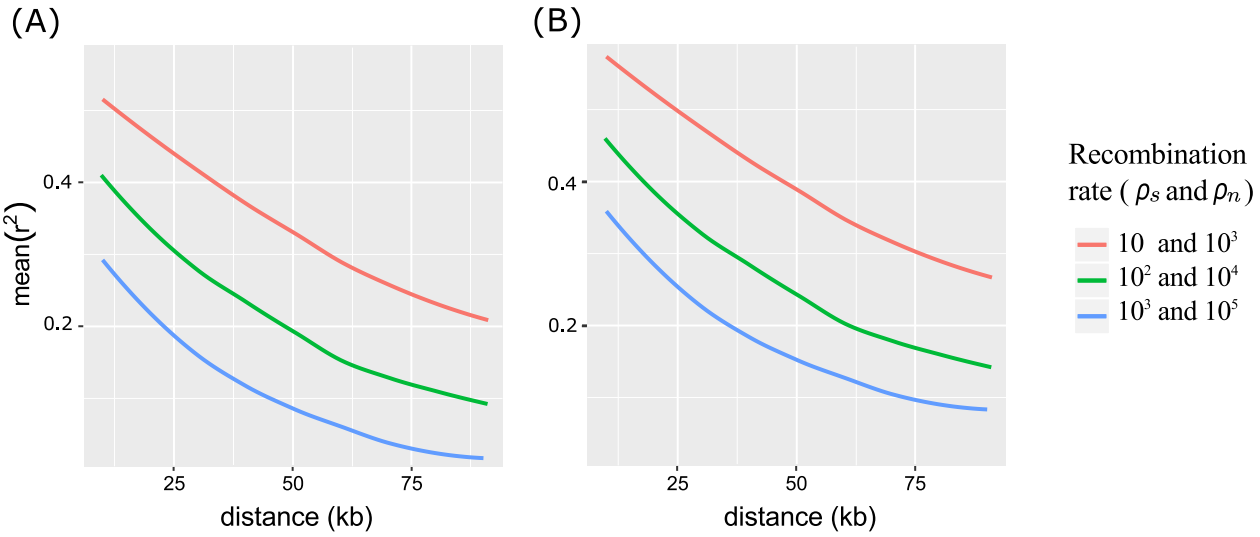


Figure 4.3 – Decay of LD (r^2) with distance between pairs of SNPs in the simulated datasets for different values of recombination rate (ρ) before (A) and after admixture (B). The recombination rate in selected and neutral regions are denoted ρ_s and ρ_n, respectively.

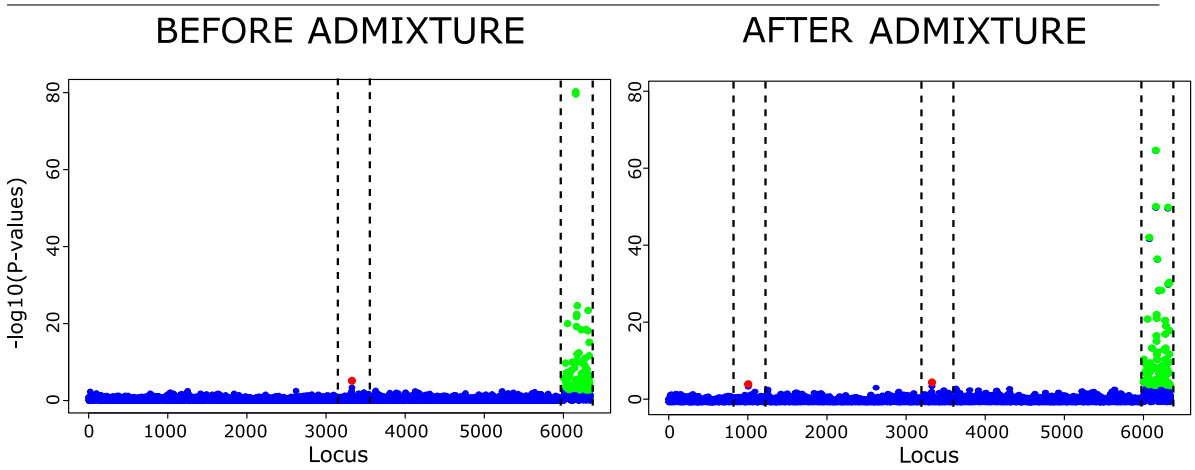


Figure 4.4 – **Manhattan plot for a simulated datasets before and after admixture**: The false positive loci, considering an expected FDR of 10%, are shown in red. Truly selected SNPs are shown in green. The Manhattan plot shows the scan for selection performed using the `snmf` software in a data set containing around 5% of outliers and $m_s/m = 0.05$.

4.3.2 Scans for selection

We performed genome scans for selection using tests based on two factors methods, `snmf` and `pcadapt`, and the standard F_{ST} (using `ANOVA`), in the datasets with different levels of LD (low, medium and high level). The simulated datasets contained around 5% of outlier loci. Figure 4.4, displays Manhattan plots for a test for selection using `snmf`, for a dataset with $m_s/m = 0.05$, before and after admixture. The outlier loci detected by the genome scans for selection based on the new F_{ST} statistic, F_{ST}^{Q} (Equation 4.1), applying with the software `snmf`, are shown in red. As in the case where simulated data do not contain LD, tests for selection in data sets after admixture presented a more significant number of loci outside the true outlier region (larger FDR). This results shows an increase in the complexity of the data when including admixed individuals (Figure 4.4 and Figure 4.5).

We estimated the false positive windows (FP_w) and observed FDR_w values for all methods and levels of LD. Figure 4.5 shows the mean values for FP_w, when we considered data before and after admixture, in the tests using the factor methods, `pcadapt` and `snmf`. The observed FDR_w and FPD_w values were computed considering an expected FDR of 10%. As shown in Figure 4.5, the number of false positive windows

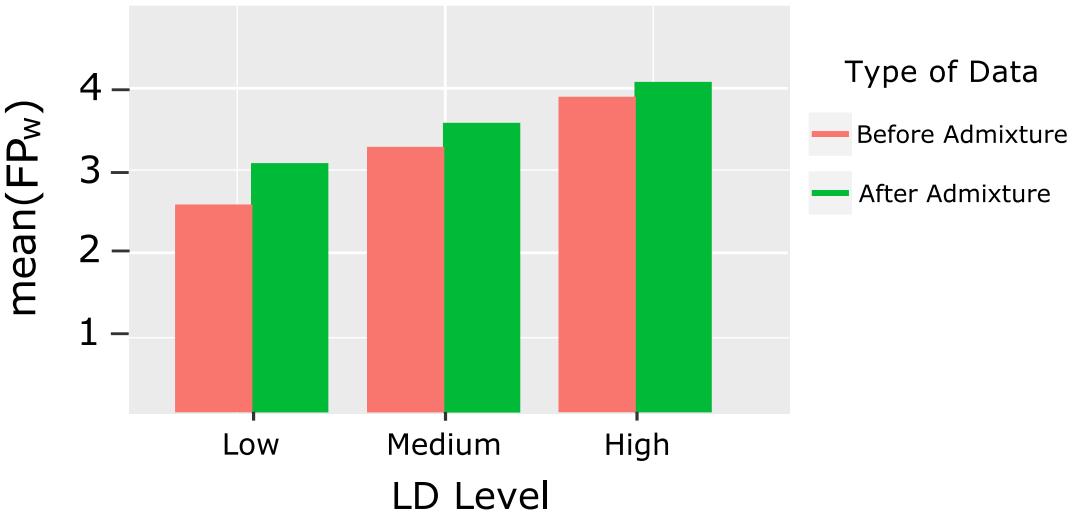


Figure 4.5 – Barplot of mean values for FP_w for data before and after admixture. The values were computed considering an expected FDR of 10% in the tests using `pcadapt` and `snmf` (pooled results).

LD level	**FDRw**	
	Before admixture	**After admixture**
low	0.31	0.33
medium	0.35	0.37
high	0.36	0.38

Table 4.2 – Mean values of observed FDR_w for data after and before admixture in tests with low, medium and high level of LD in a data sets with 5% of outliers. The values of FDR_w were divided by 2.

was greater when LD had larger levels. For a low level of LD in the data before admixture, the mean FP_w value was equal to 2.7 windows, while for the same LD level the mean FP_w was equal to 3.1 windows after admixture. Considering the data with a medium level of LD, the mean FP_w value was 3.3 windows for data before admixture and 3.6 for data after admixture. For the test in data with high level of LD, the mean FP_w value for data before admixture was equal to 3.9 windows. It remained smaller than the mean FP_w value for the data after admixture that was equal to 4.1 windows. The number of false positive windows, FDR_w was lower when LD was low (Table 4.2).

A comparison of FP_w values for the tests with `pcadapt`, `snmf` and `ANOVA` can be found in Figure 4.6. The results were similar for the tests with `pcadapt` and `snmf`.

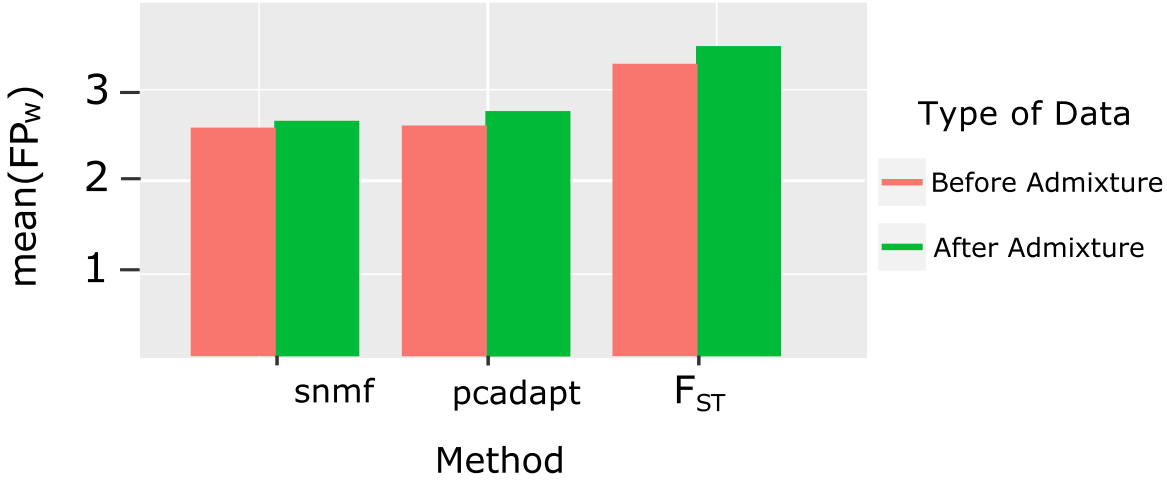


Figure 4.6 – Barplot of mean values for FP_w for data before and after admixture in the tests using `snmf`, `pcadapt` and standard F_{ST}, using `ANOVA`. The values were computed considering an expected FDR of 10% and a low level of LD.

However, as we presumed, the standard F_{ST} statistic, applied using `ANOVA`, had a less efficient performance when compared with the factor methods (`snmf` and `pcadapt`), especially in the case after admixture.

4.3.3 Pruned datasets

Next, we applied an LD-pruning method using `PLINK` to simulated datasets. For the data with low level of LD, 1900 SNPs were kept after pruning, for the data with a medium level of LD, we kept 1200 SNPs, and for the data with high level of LD, we kept 800 SNPs. Considering all simulated data and three different levels of LD (low, medium and high), we obtained the mean number of false positive windows (FP_w) and the mean of observed FDR_w value for genome scans for selection using the statistic F_{ST}^{Q} in the `snmf` program. Figure 4.7 shows the mean values of FP_w for the tests with admixed individuals before and after pruning. To compute these values, we considered an expected FDR value of 10%. For a low level of LD before pruning, the mean FP_w value was equal to 3.1 windows, while for the same LD level, the mean FP_w was equal

to 2.5 windows after pruning. Considering the data with a medium level of LD, the mean FP_w value was 3.6 windows before pruning and 2.4 after pruning. For the test in data sets with high level of LD, the mean FP_w value before pruning was equal to 4.1 windows, larger than the mean FP_w value after pruning that was equal to 2.9 windows.

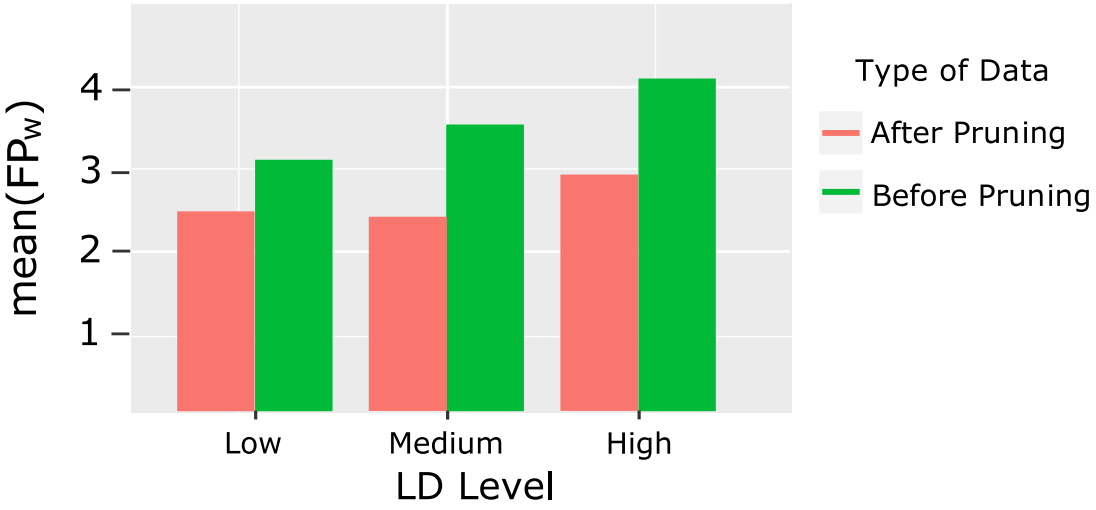


Figure 4.7 – Barplot of mean values for FDR_w for data before and after pruning. The values where computed considering an expected FDR of 10% in the tests using `PCAdapt` and `snmf`.

For the three levels of LD, the mean FDR_w observed in the test with data after pruning was lower than the mean FDR_w observed in the test with data before pruning (Table 4.3). Fewer regions were signalled as under selection outside the outlier region after pruning, what increased the precision of our test (Figure 4.8).

We compared FP_w values in tests for selection on the pruned data with `pcadapt`, `snmf` and standard F_{ST} (`ANOVA`) (Figure 4.9). Again, the results for the tests with `pcadapt` and `snmf` were similar. The standard F_{ST} statistic, had lower performance when compared with the factor methods (`snmf` and `pcadapt`).

4.4 Discussion

Population genetic studies usually assume linkage equilibrium between genetic markers. This assumption can cause overestimation of multipoint identity by descent (IBD) sharing and induces false positives in linkage analysis (Huang et al., 2004). In

LD level	FDRw	
	Before pruning	**After pruning**
low	0.33	0.31
medium	0.37	0.35
high	0.39	0.36

Table 4.3 – Mean values of observed FDR_w for data after and before pruning in tests with low, medium and high level of LD in a data sets with 5% of outliers. The values of FDR_w were divided by 2.

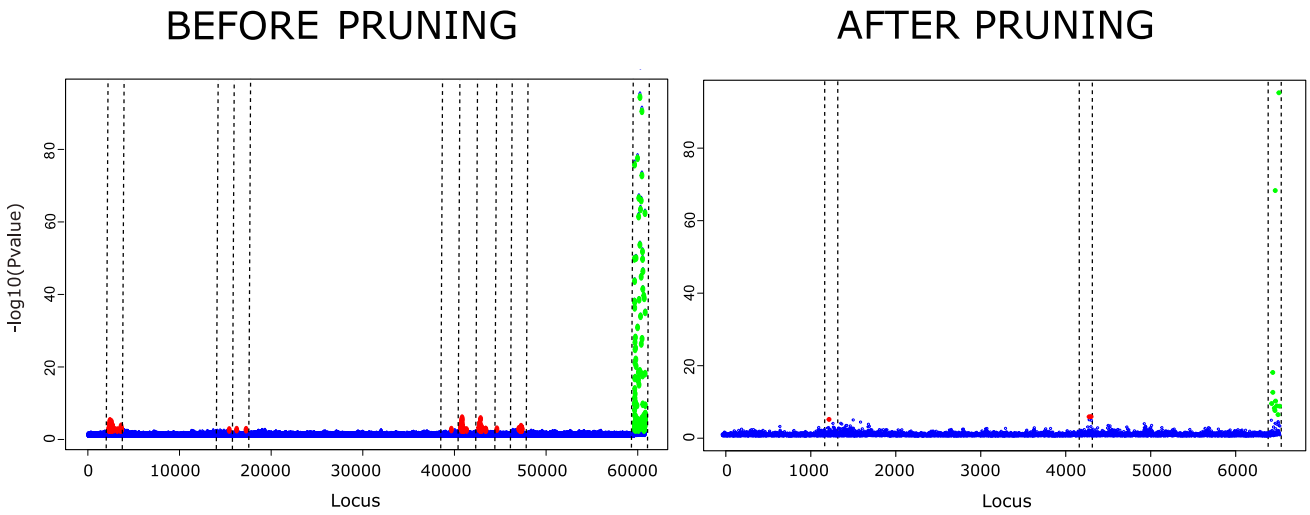


Figure 4.8 – Manhattan plot for data before and after pruning. The loci candidates are showing in red. Truly selected SNPs are shown in green. Less regions are signalled as under selection outside the outlier region in the data after pruning.

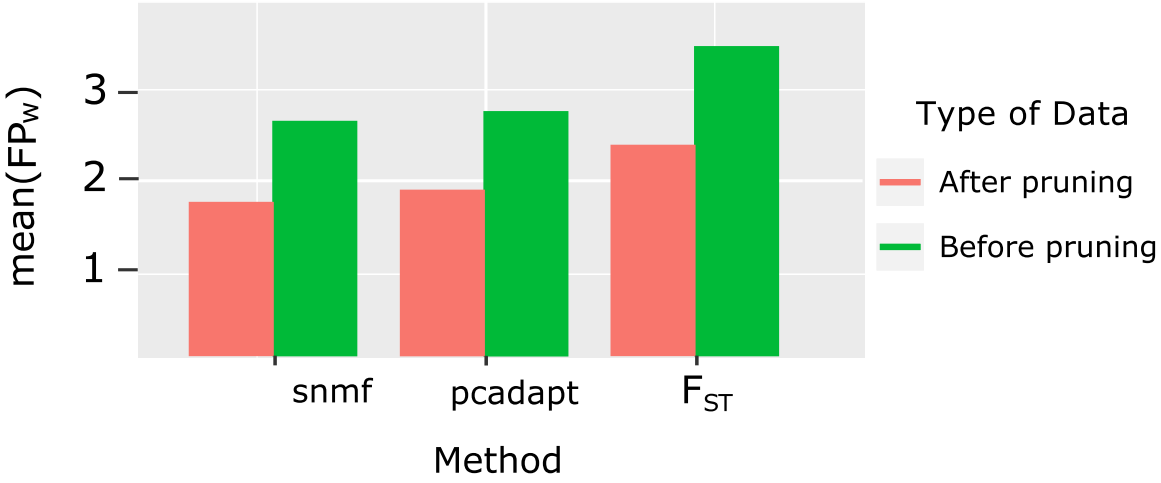


Figure 4.9 – Barplot of mean values for FP_w for data before and after pruning in the tests using `snmf`, `pcadapt` and standard F_{ST}, using `ANOVA`. The values were computed considering an expected FDR of 10% and a low level of LD.

addition, some regions of the genome may be overrepresented in the PCs due to linkage disequilibrium (Abdellaoui et al., 2013). Therefore, LD should be taken into account before to conducted genome-wide association studies (Reed et al., 2015; Laurie et al., 2010). In this study, we focused on analyse possible effects of linkage disequilibrium in genome scans for selection. To correct LD effects, we applied an LD-SNP pruned method, keeping only markers that are not in LD.

We conducted simulations in data with three different levels of recombination rate. We measured linkage disequilibrium in the data using the r^2 coefficient. We determined the physical distance between loci through LD Decay plots. In both admixed and un-admixed individuals, r^2, the value used to measure for linkage disequilibrium, was proportional to admixture in the data. Thus, we can conclude that r^2 tends to increase when individuals in the population are admixed. This conclusion is consistent with earlier studies that prove that admixture of populations creates linkage disequilibrium among loci that are not in LD in parental populations and alter the extent of LD for loci that are in LD in the parental populations (Du et al., 2007). When looking

at recombination rate, our LD decay plots showed that more recombination resulted in less linkage disequilibrium on simulated data. That was justified because when a recombination occurs between two loci, it tends to reduce the dependence between the alleles carried at those loci and thus reduce LD (Li and Stephens, 2003).

Following our goal of analysing the effects of LD in genome scans for selection, we applied scans in data before and after admixture with low, medium and high recombination rate. We calculated false discovery values observed by regions (FDR_w) and number of false positive windows (FP_w) to compare the precision of genome scans in the data with low, medium and high level of linkage disequilibrium. With the test results, we could conclude that LD increased complexity of data and resulted in false positive rates. In this way, we can indeed conclude that linkage disequilibrium can affect the outcome of tests for selection, showing that it should be examined before performing the scans. The simulations were conducted using three different methods, a new F_{ST} statistic for continuous populations and applied with `snmf` software,`pcadapt` and standard F_{ST}, applied using `ANOVA`. As showed in Martins et al. (2016) for data without LD, standard F_{ST} had a less efficient performance for data with linkage disequilibrium. This shows that factor methods are still the best option in genome scan for selection in admixed data even when the considered data have LD.

To complete our investigations about the impact linkage disequilibrium in genome scan for selection we applied an LD pruning method in simulated data using the toolset `plink`. Scans for selection were conducted in the pruned datasets. As a result of data filtering, we observed a significant decrease on the number of loci signalled as under selection outside the region contained truly outliers after pruning. We concluded that LD pruning improves our scans for selection. Our analysis shows that, as in Abdellaoui et al. (2013), pruning data is necessary when studying a population that consists of overlapping subpopulations.

In conclusion, our findings reiterate the importance of adjusting for LD since it changes the results of analysis on data with admixed individuals and can increase the false discovery rate. As a perspective for this work, we should consider improving the pruned method to eliminate loci in LD, while keeping in mind the risk of discarding

potentially informative markers from the analysis.

Chapter 5

Conclusions and Perspectives

Adaptation to local environment triggers modifications in allele frequencies and enables maintenance of genetic variation within and among populations by spatial and temporal variation in selection intensities (Nei, 2005). By screening the genome for differences in allele frequencies among populations, genome scans for selection attempt to identify genomic regions that exhibit signatures of diversifying selection (Storz, 2005; Vitti et al., 2013; Tiffin and Ross-Ibarra, 2014; Haasl and Payseur, 2016).

During the last years, many approaches have been developed to identify loci under selection. Usually, these approaches focus on examining F_{ST} values, which are related to statistical analysis of variance (Feng et al., 2015). Methods based on F_{ST} require individuals to be assigned to predefined populations. However, when the background levels of F_{ST} are weak and when populations are genetically homogeneous (Waples and Gaggiotti, 2006) or when the samples contain admixed individuals, defining subpopulations is not trivial (Pritchard et al., 2000).

Considering the above situation, we presented a method to identify loci under selection based on an extension of the F_{ST} statistic. This extension is based on ancestry coefficients and can be applied in continuous and admixed populations. We performed genome scans for selection using our new statistic in simulated data and genomic data from European lines of the plant species *Arabidopsis thaliana*, and in human genomic data from the population reference sample, `POPRES`.

To further explore the statistical methods that identify local adaptation in admixed populations, we included spatial data to compute ancestry coefficients and allele

frequencies in the software `tess3`. A genome scan for selection using our new F_{ST} statistic and `tess3` provided evidence of a substantial role for natural selection in shaping the genome-wide variation of the plant species *A. thaliana* in Europe.

In the last part of this thesis, we investigated the effects of linkage disequilibrium and considered an LD-pruning method in genome scans for selection. In this study, we reiterated the importance of adjusting for LD in genomic data since it can change the results of analysis on data with admixed individuals and can increase the false discovery rate.

In the next sections we will discuss different perspectives for future research.

5.1 Use of environmental variables in genome scans for selection

In this thesis, we proposed a statistical approach to screen genomes for signatures of diversifying selection, using a new definition of F_{ST} based on the Q (and F) ancestry matrix. By modelling admixed genotypes, our F_{ST}-based genome scan method differs from other approaches because it allows investigation of adaptation in continuous and admixed populations. In the last decade, some portions of various species genome have been identified as targets of selection using F_{ST} statistics. However, it is important to note that other processes than local adaptation can be responsible for the observed patterns in allele frequency or F_{ST}. These include demographic processes such as hierarchical population structure, significant differences in mutation rate across loci and background selection (Edmonds et al., 2004; Edelaar et al., 2011; Kruuk et al., 1999). Therefore ignoring these mechanisms can provoke the discovery of false targets of positive selection. Accounting for processes other than local adaptation requires the introduction of parameters that could capture the effect of those processes, like environmental variables. `BayeScEnv` (Villemereuil and Gaggiotti, 2015) is an example of a method that incorporates environmental information to discriminate between true and false genetic signatures of local adaptation.

`BayeScEnv`, is based on the F-model and extends the software `BayeScan` (Foll and Gaggiotti, 2008)(Subsection 1.4.5) by incorporating environmental data so as to

explicitly consider local adaptation scenarios. Villemereuil and Gaggiotti (2015) assume that genetic differentiation at individual loci is influenced by three type of effects; genomewide effects due to demography, a locus-specific effect due to local adaptation caused by the focal environmental variable and locus-specific effects unrelated to the focal environmental variable. Thus, Villemereuil and Gaggiotti (2015) focus on three different models to explain genetic structure at individual loci:

M1. Neutral model: β_j, where β_j measures the amount of drift in population j;

M2. Local adaptation model with environmental differentiation E_j: $\beta_j + g_i E_j$, where g_i is the sensitivity of locus i to the environmental differentiation of population j, E_j.

M3. Locus-specific model: $\alpha_i + \beta_j$, where α_i is a parameter specific to marker i, which can account for large differences in mutation rate across loci, allele surfing and background selection.

The models described above are summarised in Figure 5.1.

Like `BayeScEnv`, genome scans used in parallel with environmental data provide clear clues for selective forces and can complement robustness of the final set of loci identified as potentially under selection (Feng et al., 2015). Thus, the inclusion of environmental variables could be considered in future extensions of our F_{ST}-based genome-scan method.

This could be done through estimation of ancestry coefficients using environmental variables of sampled individuals, for example using the software `POPS` (Jay, François, et al., 2015). This program implements Bayesian clustering algorithms based on genetic, geographic and environmental variables. `POPS` assigns individuals or genes to genetic groups after modelling the effects of geography and environment on individual membership and admixture proportions. `POPS` is based on latent regression models that consider individual ancestry as a hidden response variable regressed on geographical and environmental covariates. Considering that `POPS` computes admixture coefficients and allele frequencies, these values can be used to estimate F_{ST} values on our new method to identify loci under selection in populations with admixed individuals

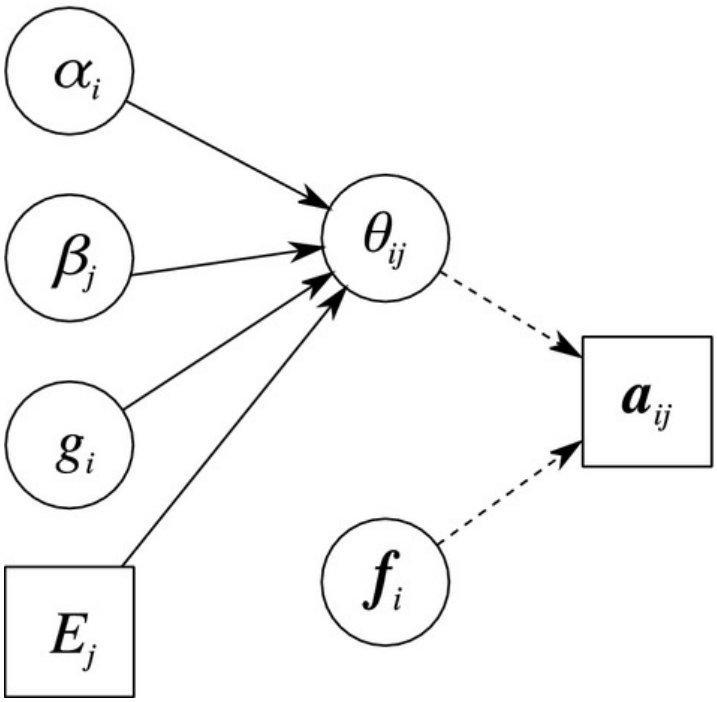


Figure 5.1 – **Directed acyclic graph (DAG) of the `BayeScEnv` model.** Squared nodes denote known quantities (E for environmental data and A for genetic marker data). Circled nodes denote unknown parameters. Plain arrows stand for deterministic relationships, and dashed arrows stand for stochastic relationships. β_j, measures the amount of drift in population j; g_i is the sensitivity of locus i to the environmental differentiation of population j, E_j; α_i is a parameter specific to marker i, such large differences in mutation rate across loci, allele counts $a_i j = a_{ij_1}; \cdots; a_{ij_{Ki}}$ at locus i within population j (where K_i is the number of distinct alleles at locus i) with parameters given by the migrant pool allele frequencies, $f_i = f_{i1}; \cdots; f_{iKi}$, and a population-and-locus-specific parameter of similarity, $\theta_{ij} = (1 - F_{ST}^{ij})/(F_{ST}^{ij})$.

(Chapter 2). Given that Villemereuil and Gaggiotti (2015) conducted simulations on 26 Asian populations from the Human Genome Diversity Panel (HGDP) SNP Genotyping data using `BayeScEnv`, we could use the same data to provide a comparison of the results. We can expect some differences in the results since the statistic used by `BayeScEnv` is based on the `Bayescan` statistic, which is impacted by the presence of admixed individuals (Luu et al., 2017a).

5.2 Spatial Principal Component Analysis

Principal component analysis is a common multivariate approach in population genetics and can be used for investigating spatial genetic patterns. By including spatial information, spatial principal component analysis has been introduced as a way to complement PCA with the objective of gaining more power on the investigation of spatial genetic structure (Montano and Jombart, 2017). Frichot, Schoville, et al. (2012), proposed a method a spatial factor analysis model (`spFA`), which incorporates spatial information in factor analysis in an explicit way.

Consider SNP data for n individuals at L loci, the genotype matrix G, where the entries $G_{i\ell}$ record the number of derived alleles at locus ℓ for individual i, and X_i, the geographic coordinate of individual i. In `spFA`, inference is performed in a factorization model similar to PCA as follows

$$G_{i\ell} = U_i^T V_\ell + \epsilon_{i\ell}, \tag{5.1}$$

where $\epsilon_{i\ell}$ are statistically dependent Gaussian variables with mean zero and with covariance matrix Σ_θ. A radial basis covariance matrix is chosen to model spatial autocorrelation patterns generated by "isolation by distance" (IBD) (Durand et al., 2009). In this way, for all individuals, i and j, Frichot, Schoville, et al. (2012) defined the covariance matrix Σ_θ as follows

$$\Sigma_\theta(i,j) = \exp(-d(X_i, X_j)/\theta), \quad \theta > 0,$$

where $d(X_i, Xj)$ represents the squared Euclidean or great-circle distance between sites with coordinates X_i and X_j. The parameter θ is a scale parameter measured in units of average pairwise distance between geographic sites $\underline{d}$.

To solve the `spFA` model (Equation 5.2), Frichot, Schoville, et al. (2012) used a Cholesky decomposition, $C^T C = \sum_\theta^{-1}$, and established equivalence with the following matrix factorization model

$$\tilde{G}_{i\ell} = \tilde{U}_i^T \tilde{V}_\ell + \epsilon_{i\ell}, \tag{5.2}$$

where $\tilde{G} = CG$, $\tilde{U}^T = CU^T$, $\tilde{V} = V$, and where $\tilde{\epsilon}_\ell$ are statistically independent Gaussian vectors of mean zero and covariance matrix equal to identity.

Considering the matrices $\tilde{G}$ and $\tilde{U}$, one perspective is to compute z-scores as Luu et al. (2017a) in the `pcadapt` statistical approach:

$$z_{i\ell} = \frac{\tilde{U}_{i\ell}^T \tilde{G}_{i\ell}}{\sqrt{\frac{\|\epsilon_{i\ell}\|_2^2}{n-K-1}}},$$

where K is the number of principal components.

As in Luu et al. (2017a), the next step is to use the z-scores values to perform a genome scan for selection considering a classical approach in multivariate analysis for outlier detection. The test statistic is a robust Mahalanobis distance D defined as

$$D_j^2 = (z_j - \bar{z})^T \Sigma^{-1} (z_j - \bar{z}) \tag{5.3}$$

where Σ is the $(K \times K)$ covariance matrix of the z-scores and $\bar{z}$ is the vector of the K z-score means (Maronna and Zamar, 2002).

In this way, spatial information can be used to increase the power of genome scans for selection. As a prospect, we could compare the results of tests for selection using `tess3` and our new F_{ST} statistic, and tests using PCA with geographic variables and Mahalanobis distance. The genome scans for selection could be performed, for example, on genetic data from European lines of the plant species *Arabidopsis thaliana*.

5.3 Detecting signatures of positive selection in real data

In this thesis, we applied our new method of genome scan for selection using genomic data from European lines of the plant species *Arabidopsis thaliana* and human genomic data from the population reference sample, POPRES. The application of our genome scan for selection to *A. thaliana* suggested several new candidate loci involved in, for example, resistance against pathogens, in allowing the maintenance of growth under stress conditions, in response to temperature stress or response to light. With our analysis of POPRES data sets, we provided additional evidence that the signals detected by our F_{ST} genome scan were already present in the populations that were ancestral to modern Europeans.

Domestication is the evolutionary process of genetic adaptation of wild animal populations to environmental conditions created by humans and, it is a relevant application of evolutionary theory (Larson et al., 2014). Comparison of the genomes of domesticated species to their wild founder populations can help to identify the genes underlying differentially selected traits, thereby advancing a fundamental goal of evolutionary biology (Turcotte et al., 2017). Atlantic salmon has been subject to domestication since the aquaculture industry was established in the early 1970s and it has been a target of many investigations (Taranger et al., 2014). Application of genome scans to Atlantic salmon populations has been facilitated by the increased availability of SNPs for Atlantic salmon and the publication of its genome (Lien, Koop, et al., 2016; Houston et al., 2014; Lien, Gidskehaug, et al., 2011). Therefore, recent outlier locus studies use available Atlantic salmon data to compare aquaculture strains with wild populations from the same region as their supposed ancestor population (Liu et al., 2017).

Liu et al. (2017) used a North American Atlantic salmon 6K SNP dataset to locate genome regions of an aquaculture strain (Saint John River) that were highly diverged from that of its supposed wild founder population (Tobique River). In this study, admixed individuals were identified and removed from the data. To detect regions under selection, Liu et al. (2017) used the `Bayescan` software (Feng et al.,

2015)(Subsection 1.4.5). Parallel analyses comparing the aquaculture population with a nearby wild population from the Stewiacke River were conducted to determine whether an overlapping set of outlier loci would be discovered (Liu et al., 2017). Considering that the method of genomic scan presented can handle admixed and continuous populations, a prospect of this PhD thesis is the application of our approach to the North American Atlantic salmon dataset. In addition, the results of our test can be compared with the results of the test done by Liu et al. (2017) using `Bayescan`. This comparison can be interesting since `Bayescan` showed to be impacted by the presence of admixed individuals (Luu et al., 2017a).

5.4 Development of a tool for genome scans for selection in R

During this thesis, genome scans for selection and all analyses were conducted using the `R` language, more specific the package `LEA`. An interesting possibility is the development of an `R` package that could be used as a tool to perform tests for selection in genomic data sets. This package could combine many tests, as our new F_{ST} statistic, `tess3`, `pcadapt`, and others, facilitating the comparison of the results. Also, we could implement an LD-pruning method that can be directly applied in the `R` environment (Privé et al., 2017).

5.5 General conclusion

In this thesis, we presented a new method of genome scan for selection based on an extension of the F_{ST} statistic to admixed and continuous populations. To explore statistical methods to identify local adaptation in admixed population, we calculated F_{ST} values using ancestry coefficients and allele frequencies computed by `tess3`, a spatial statistical program. To complement our work, we investigated the effects of linkage disequilibrium and LD-pruning methods in genome scans for selection. In conclusion, the new statistic makes possible the identification of loci in natural populations with admixed individuals, making room for numerous biological

analyses. In addition, our work allows shifting the perspectives towards the inclusion of environmental variables in scans for selection and for studies in real data with the presence of admixed individuals.

Everything that has already been is
the beginning of what is to come...

João Guimarães Rosa,
Brazilian writer
in *The Devil to Pay in the Backlands*

Appendix

REVIEW: Controlling false discoveries in genome scans for selection

MOLECULAR ECOLOGY

Molecular Ecology (2016) **25**, 454–469 doi: 10.1111/mec.13513

INVITED REVIEWS AND SYNTHESES

Controlling false discoveries in genome scans for selection

OLIVIER FRANÇOIS,* HELENA MARTINS,* KEVIN CAYE* and SEAN D. SCHOVILLE†

Centre National de la Recherche Scientifique, Université Grenoble-Alpes, TIMC-IMAG UMR 5525, Grenoble 38042, France, †Department of Entomology, 637 Russell Laboratories, University of Wisconsin-Madison, 1630 Linden Drive, Madison, WI 53706, USA

Abstract

Population differentiation (PD) and ecological association (EA) tests have recently emerged as prominent statistical methods to investigate signatures of local adaptation using population genomic data. Based on statistical models, these genomewide testing procedures have attracted considerable attention as tools to identify loci potentially targeted by natural selection. An important issue with PD and EA tests is that incorrect model specification can generate large numbers of false-positive associations. Spurious association may indeed arise when shared demographic history, patterns of isolation by distance, cryptic relatedness or genetic background are ignored. Recent works on PD and EA tests have widely focused on improvements of test corrections for those confounding effects. Despite significant algorithmic improvements, there is still a number of open questions on how to check that false discoveries are under control and implement test corrections, or how to combine statistical tests from multiple genome scan methods. This tutorial study provides a detailed answer to these questions. It clarifies the relationships between traditional methods based on allele frequency differentiation and EA methods and provides a unified framework for their underlying statistical tests. We demonstrate how techniques developed in the area of genomewide association studies, such as inflation factors and linear mixed models, benefit genome scan methods and provide guidelines for good practice while conducting statistical tests in landscape and population genomic applications. Finally, we highlight how the combination of several well-calibrated statistical tests can increase the power to reject neutrality, improving our ability to infer patterns of local adaptation in large population genomic data sets.

Keywords: control of false discovery rates, genome scans for selection

Received 1 July 2015; revision received 23 November 2015; accepted 25 November 2015

Introduction

Local adaptation often occurs among species occupying spatially heterogeneous environments, yet we know very little about the conditions in which it occurs or the particular genetic pathways involved, which would provide critical knowledge for how organisms will respond to environmental change (Davis & Shaw 2001; Davis *et al.* 2005; Jump & Penuelas 2005; Jay *et al.* 2012; Schoville *et al.* 2012; Savolainen *et al.* 2013; Fitzpatrick & Keller 2015). Adaptation to local environments triggers modifications in allele frequencies and enables maintenance of genetic variation within and among populations by spatial and temporal variation in selection intensities (Nei 2005). By screening the genome for differences in allele frequencies among populations, genome scans for selection attempt to identify genomic regions that exhibit signatures of diversifying selection (Storz 2005; Vitti *et al.* 2013; Tiffin & Ross-Ibarra 2014; Haasl *et al.* 2015).

Genome scans for divergent selection typically focus on allele frequencies and can be divided into two main approaches: identifying (i) genomic loci that show unu-

Correspondence: Olivier François, Fax: +334 56 52 00 55; E-mail: olivier.francois@imag.fr

sual allele frequency differentiation among populations or (ii) those loci with a strong association between allele frequencies and environmental variables (Savolainen *et al.* 2013). The first group of methods, population differentiation (PD) methods, compares single-locus estimates of a population differentiation statistic with their expectation from a null model of neutral evolution or with the genomewide background (Beaumont & Nichols 1996; Akey *et al.* 2002). If natural selection favours one allele at a particular locus in some populations, the test statistic at that locus will be large compared to loci in which among-population differences are the result of neutral demographic processes. Outliers in the genomewide distribution of the test statistic correspond to loci potentially targeted by selection. One of the best examples of applying genomewide scans is the discovery of several genomic regions containing genes involved in high-altitude adaptation in Tibetan populations (Simonson *et al.* 2010). Tibetans have lived at very high altitudes for thousands of years, and they share physiological traits that enable them to tolerate hypoxia. Contrasting lowland and highland populations, Peng *et al.* (2011) used genome scans to identify strong signals of selective sweeps in two hypoxia-related genes, *EPAS1* and *EGLN1*, that were significantly associated with the body response to oxygen level in highland populations.

The second group of methods, genomewide ecological association (EA) methods, estimate correlations between allele frequencies and one or more ecological variables (Hedrick *et al.* 1976; Joost *et al.* 2007; Hancock *et al.* 2008; Rellstab *et al.* 2015). EA methods rely on the common observation that selection along environmental gradients results in allele frequency clines in spatially distributed populations (Haldane 1948; Berry & Kreitman 1993). Thus, EA methods are more likely to detect gradual changes in allele frequencies linked to spatially varying environments than PD methods (Hancock *et al.* 2010; Schoville *et al.* 2012). EA methods do not require population samples, but instead can be applied to an individual-based design that draws samples from discrete sites in geographic space. A compelling example of EA tests was provided by Hancock *et al.* (2010), who conducted a genomewide scan to identify genetic loci associated with climate in the plant *Arabidopsis thaliana*. They found that nonsynonymous variants were significantly enriched among the loci strongly correlated with climate, suggesting that adaptive alleles were effectively detected, and then used their results to predict relative fitness among a set of geographically diverse ecotypes of *A. thaliana*. Parallel to the development of PD methods, recent research efforts have been devoted to improve EA methods (surveyed by De Mita *et al.* 2013; De Villemereuil *et al.* 2014; Frichot *et al.* 2015; Lotterhos & Whitlock 2015), correcting for confounding effects due not only to population structure, but also to often unobserved additional factors (Frichot *et al.* 2013; Günther & Coop 2013; De Villemereuil & Gaggiotti 2015). Those unobserved factors include uneven sampling designs, genome-sequencing biases, relatedness among individuals, gene interactions that affect phenotypic variation and linkage disequilibrium within genomes, for example.

Despite significant algorithmic improvements, an overlooked question regarding PD and EA methods is how to make decisions about which loci to retain as candidates for further investigations. The same question has been asked for genomewide association studies (GWAS), and led to several important improvements in the design and implementation of these studies. A general answer to this question is closely linked to the correct adjustment of tests for confounding factors, through the use of stringent corrections to the *P*-values across the many thousands of statistical tests performed in association studies (Balding 2006; Pearson & Manolio 2008; Korte & Farlow 2013). Displaying the empirical distribution of the significance values is a common way to show that confounding effects are removed from the analysis and that false discoveries can be controlled (Storey & Tibshirani 2003; McCarthy *et al.* 2008). Assuming that confounding errors are removed, adjusting for multiple comparisons can then be achieved through the application of false discovery rate (FDR) control algorithms (Benjamini & Hochberg 1995; Storey & Tibshirani 2003; Box 1). While corrections for population structure are often included in genome scans for selection, the assessment of test calibration and the correct application of FDR control procedures have received less attention than in GWAS.

In this study, we provide a brief overview of the recent literature on PD and EA methods, and we evaluate the capabilities of these methods to provide correct FDR control procedures. We emphasize that a unified framework is available for most hypothesis testing methods, based on the application of the chi-squared distribution to various test statistics. We review two popular approaches to test calibration: (i) empirical null-hypothesis testing where test corrections are usually carried out on the basis of inflation factors and (ii) the inclusion of random effects to account for confounding errors in statistical models (Devlin & Roeder 1999; Efron 2004; Yu *et al.* 2006). We provide evidence that these approaches are useful for controlling false discoveries in genome scans for selection as well as for combining statistical testing methods. We provide tutorial examples showing how the two approaches can be applied in practice, and how they can be implemented using a few command lines in the

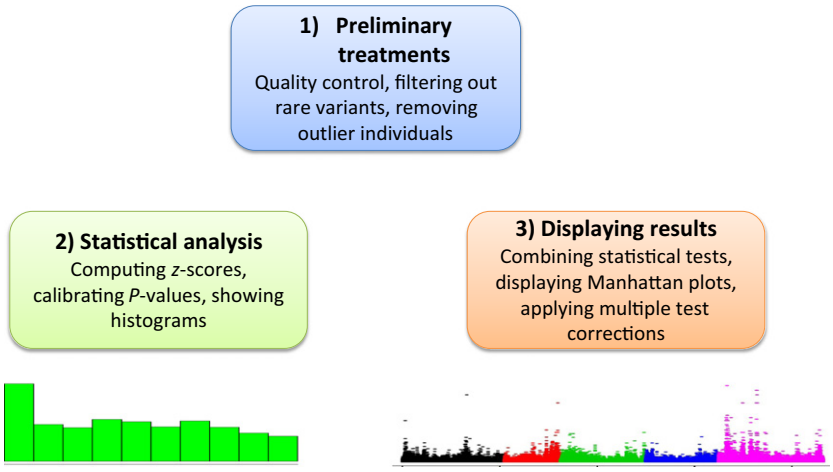


Fig. 1 Main analytical steps in performing genome scans for selection.

Box 1. Glossary

1 *Calibration of P-values*: An algorithmic correction used to ensure that the histogram of *P*-values is flat when the null-hypothesis is true.
2 *Chi-squared (χ^2) distribution*: A probability distribution for most genome scan test statistics under selective neutrality.
3 *False Discovery Rate (FDR)*: The expected proportion of false positives among the list of positive tests.
4 *FDR control algorithm*: Any algorithm utilized to ensure that the expected value of the FDR is lower than a pre-specified level.
5 *Population differentiation (PD) tests*: Tests based on the F_{ST} statistic or variants of this statistic.
6 *Ecological association (EA) tests*: Tests based on the regression of allele frequencies on environmental variables.
7 *Genomewide association studies (GWAS)*: Tests of association between phenotypic and genotypic frequencies or between phenotypes and gene expression levels.
8 *Genomic inflation factor (GIF)*: The median of squared *z*-scores divided by the median of the chi-squared distribution. Here, an estimate of the test statistic at a selectively neutral locus.
9 *Inflation factor*: Any constant used to rescale *z*-scores and recalibrate incorrect *P*-values.
10 *Linear mixed model*: A linear regression model for correlated responses that includes fixed and random effects.
11 *Latent factor mixed model (LFMM)*: A linear mixed model for which the environment is used as a fixed effect, and that includes latent factors.
12 *Power*: The proportion of tests that correctly reject the null hypothesis.
13 *Squared z-score*: Test statistic used in association studies (z^2 follows a chi-squared distribution).

R programming language (see Fig. 1 for a global picture). Finally, we illustrate the use of these calibration methods by providing a controlled list of selected loci from a 230k single-locus polymorphism (SNP) data set of Scandinavian lines of the model plant *Arabidopsis thaliana* (Atwell *et al.* 2010).

FDR control algorithms and genome scans for selection

FDR control algorithms

Given a list of genetic polymorphisms where the hypothesis of selective neutrality is rejected, the FDR is defined as the proportion of false discoveries among the positive tests. Controlling the FDR at level α, for a given probability value α, means that candidate lists are expected to contain less than a proportion α of false positives (Box 2). In genome scans, heuristic methods are often employed to minimize false discoveries, typically in nonmodel organisms where there is much less known about the species evolutionary history and population structure, and there are fewer genetic markers. Heuristic approaches rely on identifying the most significant 'top hits' in the list of putatively selected loci or comparing the output of multiple methods to identify loci that share significance across these approaches (Storz 2005; De Mita *et al.* 2013).

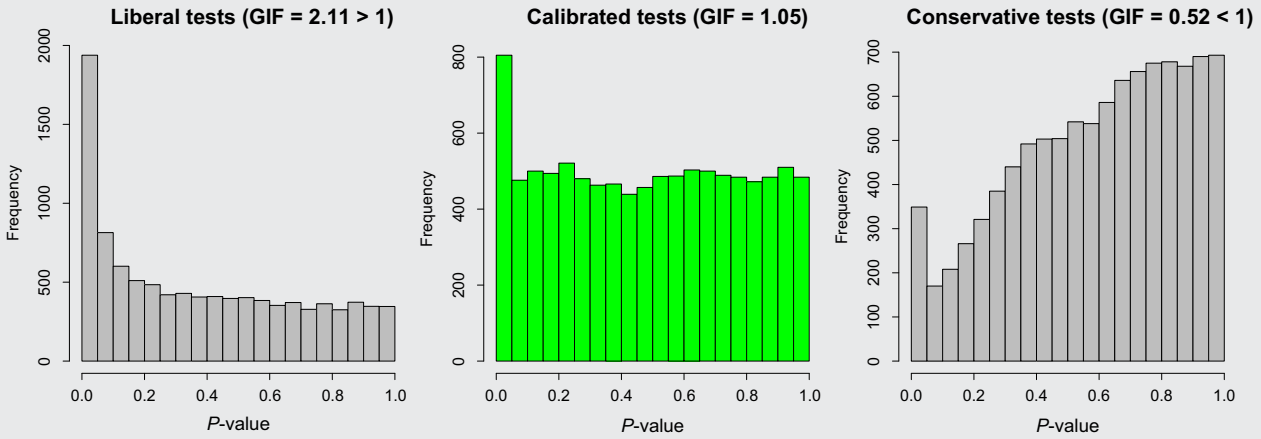
While minimizing the number of false positives is often the primary motivation in genomewide

Box 2. FDR control assumptions

The underlying principle of FDR control algorithms relies on the fact that significance values (*P*-values) corresponding to truly null hypotheses, *i.e.*, selectively neutral loci, are uniformly distributed over the interval (0,1). To see why the uniform distribution assumption is critical here, let us recall that the FDR is the expected value of the ratio F/S where F is the number of false positive tests and S is the number of significant (positive) tests (Storey & Tibshirani 2003). Let L_0 be the total number of truly null hypotheses. To provide control of the FDR at level α, Benjamini & Hochberg (1995) proposed to sort the set of *P*-values and considered the largest value k such that $P_{(k)} \leq k\alpha/L$ (L is the total number of tests). All tested items with *P*-values lower than $P_{(k)}$ were included in the list of discoveries. To compute the expected value of the ratio F/S, let us assume that the random value S is equal to $S = k$. According to the uniform distribution, the expected number of times a truly null hypothesis is rejected is equal to $\mathrm{E}[F|S = k] = L_0 k\alpha/L$ in the Benjamini-Hochberg algorithm. Assuming $k \geq 1$, we have

$$\mathrm{E}[F/S|S = k] = \mathrm{E}[F/k|S = k] = L_0/L \times \alpha \leq \alpha.$$

Under these expectations, we obtain an FDR that is under control. To check that the uniform distribution assumption is correct, standard graphical approaches can be used. These methods display histograms of test *P*-values as in Box Fig. 1.



Box Fig. 1. Histograms of test significance values (*P*-values) prior to the application of FDR control algorithms (artificial data). GIF is the genomic inflation factor for each data set.

studies, overly conservative tests can miss important associations. Approaches that result in overly conservative tests are potentially problematic because, as they inflate the Type II error rate (false negatives), they may lead to biases in the types of selective events identified and the interpretation of functional responses to selection (Williams & Haines 2011). In particular, selection on standing variation or polygenic traits is expected to result in moderate changes in allele frequencies (soft or partial selective sweeps) that are less likely to have the strong significance values and may be difficult to detect depending on the method employed (Teshima *et al.* 2006; De Mita *et al.* 2013). Downstream functional analyses of the significant hits from a pruned list, such as GO term enrichment analyses or KEGG pathway analyses, may then be biased towards a narrow set of genetic pathways (Huang *et al.* 2009). Although Type I errors are not desirable, they may be more easily identified in follow-up studies that involve different sampling strategies or functional validation. Thus, test calibration should aim to provide the largest lists satisfying a properly calibrated α level, a condition not guaranteed by the usual Bonferroni correction method which is overly conservative. When the *P*-values are calibrated, the FDR can be controlled using algorithms as described by Benjamini & Hochberg (1995), Storey & Tibshirani (2003) or Efron (2004) (Box 2).

Table 1 Population differentiation and ecological association tests. References for all the methods listed are provided in the text

	Pop.	Type	Test statistic	Null hypothesis	*d*	Correction method
Lewontin-Krakauer	Yes	PD	$(d-1)F_{ST}/\bar{F}_{ST}$	$chi^2(d-1)$	Number of population	Empirical null
ANOVA	Yes	PD	$(n-d)F_{ST}/(1-F_{ST})$	$chi^2(d-1)$	Number of population	Empirical null
FDIST2	Yes	PD	F_{ST}	Monte Carlo	Number of population	Empirical
FLK	Yes	PD	TF_{LK}	$chi^2(d-1)$	Number of population	Kinship matrix
OUTFLANK	Yes	PD	F_{ST}'	$chi^2(d)$	Maximum Likelihood	Trimming
BAYENV2	Yes	PD/EA	X^TX	$chi^2(d-1)$	Number of population	Kinship matrix
BAYPASS						
PCADAPT	No	PD	$c'h^2$	$chi^2(d)$	Number of PCs	Empirical null
TESS3	No	PD	$(n-d)F_{ST}/(1-F_{ST})$	$chi^2(d-1)$	Number of clusters	Empirical null
SAM	No	EA	z^2-score	$chi^2(d)$	Number of covariate	Empirical null
GLMM	No	EA	z^2-score	$chi^2(d)$	Number of covariate	Covariance matrix
EMMAX	No	–	z^2-score	$chi^2(d)$	Number of covariate	Kinship matrix
LFMM	No	EA	z^2-score	$chi^2(d)$	Number of covariate (1)	Latent factors

Pop indicates the use of population data (otherwise individual data). *Null hypothesis* describes the distribution of the test statistic under the null hypothesis. *d* is the degree of freedom of the chi-squared distribution. *Correction method* summarizes the method used to obtain corrected significance values.

A key condition for applying FDR control algorithms is that the test *P*-values behave as uniformly distributed random variables when the null hypothesis is correct. To check this condition, graphical approaches displaying histograms of *P*-values are very useful (Balding 2006). When performing multiple tests to evaluate loci under selection, a majority of loci are expected to be selectively neutral (H_0), whereas a minority of loci will deviate from the neutral distribution. In this case, the empirical distribution of *P*-values will consist of a mixture of a uniform distribution and a peaky distribution showing a peak close to 0. The uniform distribution corresponds to neutral loci, whereas the interesting loci are found under the peak (Efron 2004). Assessing whether the shape of the empirical distribution of *P*-values is correct is an essential step in GWAS and gene expression analysis (Storey & Tibshirani 2003; Balding 2006; Dudoit & Van der Laan 2007). This, however, is missing in the literature of genome scans for selection.

Genome scans for selection

In scans based on PD methods, the most frequently used statistic to screen the genome is the fixation index, F_{ST}, which can be computed at each locus. The fixation index is directly related to the variance in allele frequency among populations and to the degree of resemblance among individuals within populations (Holsinger & Weir 2009). According to standard population genetic theory, the definition of F_{ST}, which is based on Wright's work, corresponds to the amount of variance in allele frequency that can be explained by population structure (Wright 1951). Thus, estimates of F_{ST} correspond to estimates of correlation coefficients, and their computation can be based on an analysis of variance (ANOVA) of allele frequencies (Weir & Cockerham 1984; Weir 1996). A drawback of a direct application of ANOVA approaches to detecting selection is that they are likely to generate large numbers of false-positive tests. To overcome this problem, Lewontin & Krakauer (1973) pioneered the development of the statistical theory of PD tests for the selective neutrality of polymorphisms by introducing the chi-squared distribution to evaluate the statistical significance of their test (Table 1). Beaumont & Nichols (1996) extended the Lewontin & Krakauer approach, and proposed to detect selected loci using the distribution of neutral F_{ST} conditioned on the expected heterozygosity, while assuming an island model of migration-drift equilibrium (program FDIST2). Similarly, the program DETSEL implemented a method of detecting selection that relies on a model of population divergence by pure random drift (Vitalis , *et al*. 2001, 2003).

Other innovations to the Lewontin & Krakauer approach include F_{LK}, a test which compares patterns of differences in allele frequencies among populations to the values expected under a scenario of neutral evolution (Bonhomme *et al*. 2010). To test selective neutrality, F_{LK} reconstructs a topology modelling population divergence wherein the branch lengths correspond to the amount of genetic drift in each population. More recently, the program BAYENV2 considered a new statistic, X^TX, which evaluates departure from selective neutrality by incorporating predictions from a population genetic model (Günther & Coop 2013). The model of BAYENV2 was improved and re-implemented in the program BAYPASS (Gautier 2015). In contrast to these model-based approaches, some model-free methods rely on the

data to correct the effects of confounding factors empirically. The recently proposed OUTFLANK test used an empirical approach based on a trimming procedure to the evaluate the test *P*-values (Whitlock & Lotterhos 2015). An individual-based empirical method was implemented in TESS3 to perform tests for selection in continuous populations, based on ancestral allele frequency differentiation and spatially varying ancestry coefficients (Caye *et al.* 2015). Similarly, a fast version of PCADAPT implemented an empirical test for selection based on the eigenvalues of a principal component analysis and the communality statistic (h^2, Duforet-Frebourg *et al.* 2015).

The development of EA methods has been much more recent than the development of PD methods (Savolainen *et al.* 2013). Many studies have proposed that EA methods improve the ability to detect adaptation from standing variation (Pritchard *et al.* 2010). Most EA methods use hypothesis testing approaches to identify strong correlations between allele frequencies and an environmental variable (Table 1). Using a standard regression framework, EA methods amount to test the null hypothesis $H_0 : R^2 = 0$ against the alternative hypothesis $H_1 : R^2 > 0$, where R^2 is the proportion of the allele frequency variation explained by the environmental variable computed at each locus. A variety of EA statistical models have been developed (Rellstab *et al.* 2015), including generalized linear models (SAM, Joost *et al.* 2007), generalized linear mixed models (GLMM, Jones *et al.* 2013) or latent factor mixed models (LFMM, Frichot *et al.* 2013). At the exception of the SAM approach, a general feature of EA methods is to use information contained in the genotypic data set to evaluate confounding effects, and eventually achieve test calibration empirically.

A unified testing framework for genome scans for selection

Chi-squared distributions

PD and EA methods have often been considered as two distinct approaches to genome scans for selection (Savolainen *et al.* 2013; Vitti *et al.* 2013). This section argues that a common statistical framework is underlying PD and EA testing methods, and that this framework is useful in applying corrections for confounding errors, and in solving multiple test issues. The common statistical framework for PD and EA testing approaches is based on the use of the chi-squared distribution for rejecting the null hypothesis of selective neutrality at a given locus (Table 1). The connection between PD and EA methods arises because PD methods can be viewed as evaluating the association between allele frequencies and categorical variables encoding population labels (factors) that represent the uncharacterized environment of each population.

However, it is important to note that the test statistic used to reject the null hypothesis and the degrees of freedom differ in each method (Table 1). In PD scans, the degree of freedom is often estimated by the number of population samples (Lewontin & Krakauer 1973; Bonhomme *et al.* 2010; Günther & Coop 2013; Gautier 2015). Whitlock & Lotterhos 2015 proposed estimates based on a maximum-likelihood principle, leading to degrees of freedom less than the actual number of populations, so that their approach to F_{ST} tests accounts for the shared demographic history of the samples. Based on individual ancestry estimation methods, Duforet-Frebourg *et al.* (2015) and Caye *et al.* (2015) estimate the degree of freedom of their tests by the number of principal components or by the number of genetic clusters inferred from the data. The case of EA tests is simpler to describe as the degrees of freedom correspond to the number of environmental predictors (Frichot *et al.* 2015).

The ubiquity of the chi-squared distribution enables a unified treatment of test calibration and FDR control in genome scans for selection, which can be achieved by applying techniques developed for the analysis of GWAS and genomewide gene expression analysis. Suppose that allele frequencies at a particular SNP are significantly associated with some disease. This may occur when a SNP is associated with a confounding factor, which correlates with the GWAS phenotype but is not in the same causal pathway. Just as in genome scans for selection, confounding variables include genetic ancestry, genotyping error, ascertainment bias and epistatic effects (Vilhjálmsson & Nordborg 2013). Correcting the association tests for confounding effects is crucial to the control of false discovery rates, and this is usually based on the chi-squared distribution. A first correction strategy consists of modifying the null hypothesis to match the levels of neutral genetic background variation observed in the data set. This technique is sometimes called genomic control in GWAS, and empirical null hypothesis testing in studies of differential gene expression (Devlin & Roeder 1999; Efron 2004). A second strategy consists of modifying the regression equation to model the effect of various confounding factors directly (Price *et al.* 2006; Yu *et al.* 2006). Those two strategies are detailed in the next paragraphs.

Test correction and inflation factors

Genomic control relies on the introduction of inflation factors (Box 3). Inflation factors are constant values, λ, that are used to rescale the test statistic in order to limit

Box 3. Recalibrating and combining significance values

False positives arise in statistical tests when the null hypothesis (H_0) is misspecified. This phenomenon happens in genome scans because the tests ignore the proportion of variance explained by selectively neutral processes such as past demography, genetic relatedness, population structure and other confounding factors. The presence of confounding factors directly impacts the distribution of the test statistic under neutrality, and indirectly impacts multiple testing correction procedures that assume a uniform distribution of significance values under H_0.

Consider a statistical test based on the chi-squared distribution with D degrees of freedom (Table 1), and denote by z_ℓ^2 the test statistic used to evaluate significance at locus ℓ. Test recalibration is the process by which one builds an empirical null hypothesis from the data, and re-evaluates significance values in a way that accounts for confounding errors. The target of recalibration approaches is to evaluate the expected value of the test statistic at selectively neutral loci. Any estimate of this value, λ, is called an inflation factor. Given an inflation factor, significance values are computed for each locus ℓ as follows (L is the number of loci)

$$p_\ell = \mathrm{P}(\chi_D^2 > z_\ell^2/\lambda), \quad \ell = 1, ..., L,$$

so that λ corrects the inflation of the test statistic z_ℓ^2 at each locus (see Appendix S1, Supporting information for several examples of the use of this formula). A common approach to evaluate the constant λ is a method called genomic control, that estimates the genomic inflation factor, obtained after computing the median of the squared z-scores and dividing this value by the median of the chi-square distribution with D degrees of freedom (Devlin & Roeder 1999). A more general way to evaluate inflation factors is using the local FDR method developed by Efron (2004), for example implemented in the R program FDRTOOL (Strimmer 2008).

Similar scaling approaches can be applied to calculate the significance values resulting from the combination of several methods, each testing the same null hypothesis. For example, a robust version of the Stouffer method is based on the median of z-scores obtained for each method at each locus (Whitlock 2005, see examples in Appendix S1 and S2, Supporting information). In the case of independent tests, the scaling factor for the median is equal to $\sqrt{\pi m/2}$, where m is the number of methods used. In the case of dependent tests, inflation factors can be used to determine the scaling factor, so that significance values resulting from the combination of tests have a flat distribution under H_0.

inflation due to population structure and confounding factors. The goal of the rescaling procedure is to define a modified test statistic leading to a flat histogram for the significance values. As the approaches listed in Table 1 rely on the chi-squared distribution, their test statistics will be designated as squared z-scores in the rest of our study. For chi-squared tests, the rescaled statistic is z_ℓ^2/λ where z_ℓ is the score computed at locus ℓ, and the degree of freedom of the test is left unchanged. This technique has been called an empirical null-hypothesis testing approach by statisticians because it modifies the base-line null hypothesis, $H_0 : z_\ell^2 = 1$, and replaces it by a new null hypothesis, $H_0 : z_\ell^2 = \lambda$, in which λ is estimated from the data.

For PD tests, empirical null-hypothesis tests correspond to testing the null hypothesis $H_0 : F_{ST} = \theta$, where θ is the level of population differentiation expected at selectively neutral SNPs. For EA tests, the modified null hypothesis corresponds to $H_0 : R^2 = \theta$, where θ is the proportion of the genetic variation explained by the environmental variable at selectively neutral SNPs (Wang *et al.* 2013). Using the correspondence between z-scores and correlation coefficients, one can show that the tested value, θ, is linked to the inflation factor by a simple relationship

$$\lambda = (n - d)\frac{\theta}{1 - \theta},$$

or equivalently,

$$\theta = \frac{\lambda}{n - d + \lambda}.$$

Thus, estimates of λ provide estimates of the level of population differentiation or the proportion of the genetic variation explained by the environmental variable expected at selectively neutral SNPs. In the above equations, d represents the number of populations in an ANOVA test, and $d-1$ is equal to the number of environmental predictors in an EA test (n is the number of individuals in the sample). For example, when the inflation factor is equal to $\lambda = 5$, then F_{ST} can be estimated to be around 2.4% at a selectively neutral SNP in a two-population model where we have $n = 100$ individuals in each sample ($d = 2$). Following GWAS

approaches, an estimate of λ is commonly obtained after computing the genomic inflation factor, defined by the median of the squared *z*-scores divided by the median of a chi-squared distribution with $d-1$ degrees of freedom (Devlin & Roeder 1999).

Linear mixed models

A second approach to adjusting for confounding factors consists of modifying the regression model while keeping the null hypothesis unchanged. In principle, the modification introduces additional factors that represent the sources of error due to each confounding effect. For example, Price *et al.* (2006) have suggested the inclusion of principal components of neutral variation as fixed effects to correct for population stratification in GWAS. Yu *et al.* (2006) and Kang *et al.* (2010) considered mixed models in which random effects account for relatedness among individuals (program EMMAX). Those mixed models also perform well in the presence of genetic linkage and epistasis (Platt *et al.* 2010; Vilhjálmsson & Nordborg 2013). GWAS mixed models specify a similarity matrix for the covariance structure of random effects which is commonly based on kinship coefficients.

In general, corrections using fixed or random effects have proven useful to GWAS in which the proportion of SNPs associated with a particular trait is expected to be very small (Price *et al.* 2010), but they may be inappropriate when associations with ecological biotic or abiotic factors are investigated. For example, estimating principal components or a covariance matrix in a genome scan for selection would require a set of SNPs that are assumed to be truly neutral. This set of control SNPs can be difficult to define in EA analyses. Latent factor mixed models (LFMM) do not require any set of control SNPs, and they attempt to remove the effect of relatedness and genetic linkage when inferring ecological associations using latent factors (Frichot *et al.* 2013). The principle is that *K* random factors are included in the regression model. The number of factors, *K*, can be estimated by applying principal component analysis or ancestry estimation algorithms to the genotypic data. The factors are estimated from the full data set at the same time as the regression coefficients are computed. In addition, LFMM can be used as a PD test when the environmental variable is defined as a categorical variable representing population membership for each individual.

Tutorial examples and data analysis

In this section, we summarize our best practices for conducting statistical analysis of genome scan tests and demonstrate these recommendations in supporting examples. First, scientists will choose one or several testing procedures among those presented in Table 1. The application of a particular testing procedure to the data produces test statistics corresponding to squared *z*-scores for each locus. Equivalently, some computer programs result in significance values instead of squared *z*-scores. For those programs, the significance values can be transformed into squared *z*-scores using the quantile function of the chi-squared distribution (see Supporting information for R code). The degrees of freedom of the chi-squared distribution are indicated in Table 1, fifth column.

For each procedure, one then evaluates inflation factors from the test statistics. This is carried out by computing the genomic inflation factor (Box 2) or using more sophisticated approaches such as Efron's local FDR method. For appropriately calibrated methods, the inflation factor is expected to be close to one, and the histogram of test significance values is expected to display a flat shape (Box 2). Several procedures from Table 1 already include corrections for confounding errors generated by population structure or other unobserved factors, and should be correctly calibrated. If the inflation factor is not close to 1 and the significance values are not evenly distributed (i.e. flat), then we can recalibrate the test *P*-values by applying the formula given in Box 3. It should be noted that assessing inflation factors is always subjective, as the proximity to the value of one will vary, usually by increasing with the sample size and the background levels of F_{ST} or R^2. To be justified in recalibrating the significance values, it is advisable to report inflation factors prior to adjustment and histograms following adjustment.

The final step of analysis could consist of combining well-calibrated significance values resulting from distinct tests (example 1 below), or from distinct runs of a particular method (example 2 and 3 below). In both cases, our meta-analysis approach combines *z*-scores using a robust variant of the Stouffer method (Box 3). In summary, corrections are applied at two stages, first to obtain well-calibrated significance values for each test, and then to account for correlation among tests resulting from distinct methods or program runs. A step-by-step description of the statistical analyses of example 1 and 2 and their corresponding R commands are provided as supplementary files (Appendix S1 and S2, Supporting information). Example 3 is computationally more intensive and would take a few hours to reproduce. A short simulation study of the power of tests in described in Appendix S3 (Supporting information).

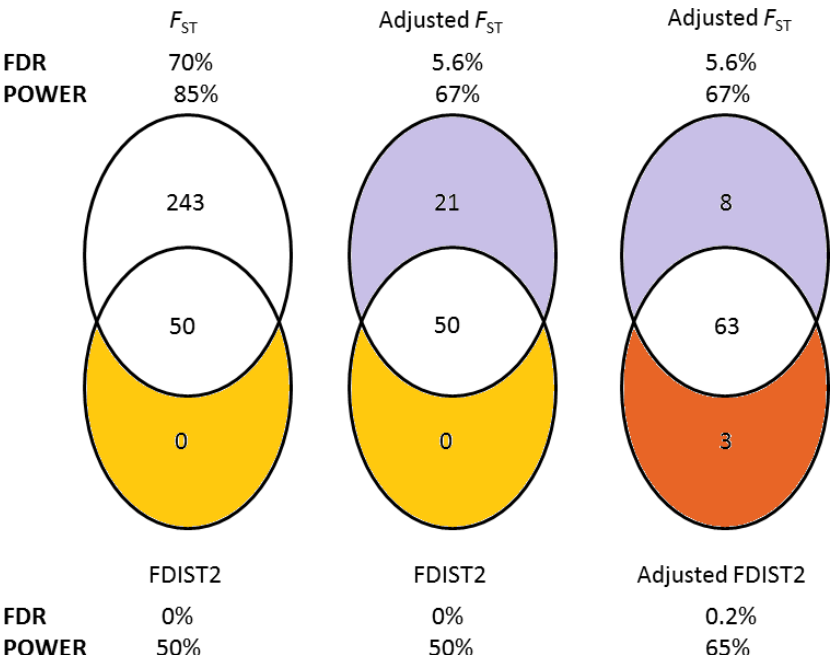


Fig. 2 Summary of simulated data analysis for example 1. Venn diagrams, observed FDR and power for four genome scan tests: F_{ST} (ANOVA), FDIST2, adjusted F_{ST}, adjusted FDIST2. Adjusted F_{ST} and adjusted FDIST2 are tests for which the significance values were recalibrated using the genomic inflation factor. Observed FDR and power were computed for an expected level of FDR of 10%. The numbers in the circles represent the number of positive tests for each method. See Appendix S1 (Supporting information) for an extended description of the data and all the processing steps that generated the Figure.

Combining statistical models (example 1): two tests are better than one

We considered a two-population model in which the populations evolved under migration-drift equilibrium (Wright's 2-island model). We used the computer program MS to perform coalescent simulations of neutral and selected SNP loci (Hudson 2002). One hundred diploid organisms were genotyped in each population. The justification for the use of Wright's models to simulate selection is that there is an overall migration rate of individuals, which, in principle, could be estimated using neutral markers, and an effective migration rate that reflects action of selection to filter out a fraction of migrants having maladapted genotypes (Bazin *et al.* 2010). In this simulation model, effective migration rates can vary across loci. Individual genotypes consisted of 1000 unlinked SNPs, and the proportion of loci under selection was set to 10%. To create the data, we used an overall migration rate of $4Nm = 20$ at 900 truly neutral loci and $4Nm_s = 0.1$ at 100 truly adaptive loci. We performed statistical tests using two methods: the computer program FDIST2 and F_{ST} computed from an ANOVA approach. For FDIST2, a target F_{ST} value of 5% was used for the computation of significance values.

The FDIST tests were conservative, whereas the ANOVA tests were liberal (Appendix S1, Supporting information). For a level of FDR of 10%, the Benjamini-Hochberg algorithm led to observed FDRs equal to 0.0 (FDIST2) and 0.70 (ANOVA). The candidate loci obtained from the FDIST2 tests were included in the ANOVA list, providing little additional insight into potentially selected loci, and the Venn diagram was highly unbalanced (Fig. 2). We then applied corrections to the poorly calibrated tests in order to change the significance threshold of the null hypothesis (Box 3). First, we used genomic control to correct the ANOVA tests, and the histogram of significance values had the desired shape (Appendix S1, Supporting information). After correction, the false positives were nearly all removed from the list of ANOVA discoveries (observed FDR of 5.6%), and the power became superior to FDIST2 (Fig. 2). We then applied corrections to FDIST2 to recalibrate significance values. After transforming the *P*-values into squared *z*-scores, we estimated an inflation factor equal to $\lambda = 0.4$ (Appendix S1, Supporting information). Recalibration clearly increased the power of FDIST2 tests, and the Venn diagram became balanced (Fig. 2). This result was achieved at the cost of an increased level of FDR (0.015), but this FDR remains below our nominal expected level of 10%.

While there is overlap in the two tests (Fig. 2), each uniquely detects significant loci and their combined results provide a longer list of candidate loci. Using a meta-analysis procedure that combined the *z*-scores from both tests, our testing approach had power equal to 0.73 and the observed FDR remained close to the expected level of 10 per cent. In summary, this example shows that: (i) statistical tests can only be compared when the null-hypothesis is correctly specified, which could be achieved using simple calibration procedures; (ii) the well-calibrated model-free approach (ANOVA) outperformed the model-based approach (FDIST2), even

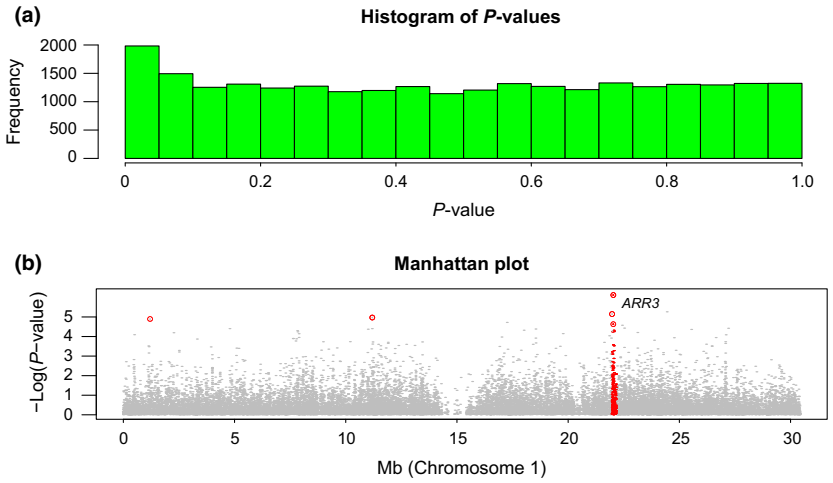


Fig. 3 Adaptation to climate in European lines of *A. thaliana*. (a) Histogram of test significance values resulting from the combination of 5 LFMM models (or runs) using $K = 6$ latent factors, (b) Manhattan plot of minus $\log_10$ (Log) *P*-values for the plant chromosome 1. Candidate loci resulting from the application of the FDR control algorithm are identified with circles (expected FDR level of 7%). See Appendix S2 (Supporting information) for an extended description of this analysis and all the processing steps that generated the Figure.

though the data were simulated under assumptions of the FDIST2 model; and (iii) combining well-calibrated tests increased to power to reject neutrality.

Adaptation to climate in European lines of A. thaliana (example 2)

We examined the utility of the mixed model approach using EA methods in a study of adaptation to climate in 170 European inbred lines of the model plant species *Arabidopsis thaliana* (Atwell *et al.* 2010). *Arabidopsis thaliana* lines were genotyped using a SNP-chip containing 230k SNPs. In our example, the analysis of genetic variation was restricted to the first chromosome with density of one SNP per 1000 bp (26k SNPs). For each of the 170 European lines in the data set, eighteen bioclimatic variables were extracted at 30 arcsecond (1 km^2) resolution from the WorldClim database (Hijmans *et al.* 2005). To test for associations between loci and climate, we summarized the bioclimatic data with the first principal component of the 18 variables.

Our genome scan for selection used five distinct statistical models fitted using the LFMM program as implemented in the R package LEA (Frichot & François 2015). A prior analysis of population genetic structure based on the ancestry estimation program sNMF (Frichot *et al.* 2014) suggested that $K = 6$ clusters (and five principal components) could explain the observed genetic variation in *A. thaliana*. In subsequent analysis of the data with LFMM, $K = 6$ latent factors were used to perform genome scans for selection. The five models corresponded to five distinct runs of the program with distinct initial values. Each model corresponded to a distinct local optimum of the likelihood function and was characterized by a specific set of *z*-scores (Appendix S2, Supporting information).

A meta-analysis of the 5 LFMM models was applied to the *z*-score matrix. The *z*-scores at each locus were combined according to a robust variant of the Stouffer method (Brown 1975; Whitlock 2005). In this robust approach, the mean value of the *z*-scores was replaced by their median value at each locus. The bias created by combining 5 distinct *z*-scores was corrected by the introduction of an inflation factor ($\lambda = 0.78$, Appendix S2, Supporting information), and significance values were computed from the chi-squared distribution with one degree of freedom. The histogram of *P*-values provided evidence that the confounding effects were removed (Fig. 3). A short list of candidate loci was proposed on the basis of the Benjamini-Hochberg algorithm, applied at a level of FDR equal to 7%. The 3 top hits in the resulting list of SNPs identified a protein coding region corresponding to the gene *ARR3* (locus AT1G59940), a type A response regulator gene involved in circadian rhythm, and a gene involved in the cytokinin-activated signalling pathway (Fig. 3).

Local adaptation in Scandinavian lines of A. thaliana (example 3)

Test calibration and FDR control algorithms were also applied to address multiple testing in a genome scan for selection in 52 Scandinavian lines of the model plant species *Arabidopsis thaliana* (Atwell *et al.* 2010; Huber *et al.* 2014). We analysed the five chromosomes with a density of one SNP per 500 bp. The 52 lines were grouped in two genetic clusters showing very low levels of shared ancestry. Each cluster was restricted to a narrow geographic range and corresponded to a geographic region to the north or the south of Scandinavia (Huber *et al.* 2014). The Northern Sweden cluster included 14 individuals, and the Southern Sweden cluster included 38 individuals. Genome scans for selection were performed with ANOVA and with LFMM. Here, LFMMs were used to run a PD test, by considering

population membership as a binary covariate taking the value 0 in the southern sample and 1 in the northern sample. LFMM runs were based on two latent factors ($K = 2$), and resulted in a correctly calibrated test (Fig. S1, Supporting information).

For an expected FDR equal to 0.1%, the Benjamini-Hochberg algorithm resulted in a list of 167 chromosome positions, some of which span the same region due to genetic linkage. The estimate of the inflation factor from the ANOVA test was $\lambda \approx 4.5$, which provided an estimate of the neutral differentiation statistic, $F_{ST} \approx 8\%$. A simulation study comparing ANOVA and LFMM showed that LFMM had power similar to ANOVA tests when the ANOVA tests are optimally calibrated (Appendix S3, Supporting information). We then analysed the top 50 hits from the LFMM runs (Table 2). Several genomic positions were reported in Huber *et al.* 2014), table 4) as corresponding to regions undergoing selective sweeps. The top list contained polymorphisms in genes involved in photosynthesis, response to heat, response to UV, response to freezing and response to light stimulus (AT1G03600, AT2G43130, AT5G07990, AT5G27540, AT5G27630). LFMM also detected mutations in genes involved in root, flower, meristem and xylem growth or development (AT1G04240, AT1G04390, AT3G62160, AT5G07290). Several genes were involved in defence response, response to biotic stimulus, nematode resistance or bacterial immunity (AT5G07390, AT5G11250, AT5G07220, AT3G09980, AT3G12100).

Discussion

Summary of main points

PD and EA methods are widely used to detect signatures of natural selection from population genomic data. Our discussion of statistical tests focused on hypothesis testing methods, and compared approaches based on chi-squared tests. The methods studied in our examples ranged from the simplest to the most elaborate one among a longer list of methods available (Table 1). Thus, the observations reported in our study are representative and applicable to a large category of statistical approaches. The methods rely on a two-stage procedure (Box 4). At the first stage, statistical tests are performed, and the tests return significance values for each locus. This first stage may include a combination of several methods. At the second stage, decisions are made about which loci to retain in a list of candidates potentially under selection. In this study, we argued that the decisions made at the second stage may be incorrect when the histogram of test significance values is not flat under H_0, and test recalibration at the first stage is often necessary.

The two-stage process described above is closely related to the methods employed in genomewide association studies, for which a well-developed literature provides resources to improve FDR control in genome scans for selection (Devlin & Roeder 1999; Storey & Tibshirani 2003; Price *et al.* 2006; Dudoit &

Table 2 Local adaptation in Scandinavian lines of *A. thaliana*. List of loci with annotations among the fifty top 'hits' (expected FDR level of 0.1%, 167 candidate loci)

Chromosome	Position (kb)	Gene	Biological process	References
1	899	AT1G03600	Photosynthesis	
	1145	AT1G04280	Unknown	Huber *et al.*
	1182	AT1G04390	Flower morphogenesis	
	20 009	Intergenic		Huber *et al.*
	20 144	Intergenic		Huber *et al.*
2	9608	AT2G22620	Carbohydrate metabolism	
	17 929	AT2G43130	Response to heat	
	18 256	AT2G44110	Response to biotic stimulus	
	18 679	AT2G45340	Regulation of meristem growth	
5	2265	AT5G07220	Regulation of abiotic stress	
	2298	AT5G07290	Meristem growth	Huber *et al.*
	2561	AT5G07990	Response to UV	
	2564	AT5G08000	Response to heat	
	3591	AT5G11250	Defence response	
	6779	AT5G20080	Unknown	Huber *et al.*
	9726	AT5G27540	Response to freezing	
	9780	AT5G27630	Response to light stimulus	

The last column indicates whether the SNP was reported to be under selection in Huber *et al.* (2014).

Box 4. Summary points

1 Genome scans for selection are two-stage procedures: One first performs statistical tests that return locus significance values, and then makes decisions about which loci to retain as candidates for selection.
2 Based on the histogram of test significance values, statistical test calibration is the key to FDR control.
3 Two approaches are available: (i) empirical null-hypothesis testing, as illustrated by the estimation of inflation factors, and (ii) modelling confounding errors, as illustrated by latent factor models.
4 Combining several well-calibrated statistical tests using the z-score method can increase power to reject neutrality.
5 Following test calibration, candidate loci can be selected on the basis of a classical FDR control algorithm.

Van der Laan 2007). In the GWAS literature, statistical test calibration is the key to control the type I error or the FDR. In our study, two main calibration approaches were applied to genome scans for selection. The first calibration method was based on a technique called empirical null-hypothesis testing and could be implemented by estimating inflation factors. The second calibration method adjusted for confounding errors using mixed models, and it was illustrated by the introduction of latent factors in regression models. In a simulation study, we observed that latent factor models outperformed tests including corrections on ANOVA or GLM for PD and EA methods (Appendix S3, Supporting information). When test calibration is applied, genome scan tests become complementary and can be combined to increase the power to reject neutrality.

Extension to Bayesian methods

Bayesian approaches to detecting selection have historically been considered as alternatives to hypothesis testing methods (Beaumont & Balding 2004; Foll & Gaggiotti 2008; Bazin *et al.* 2010; Coop *et al.* 2010; Günther & Coop 2013; De Villemereuil & Gaggiotti 2015; Gautier 2015). While hypothesis testing methods assess the null hypothesis of selective neutrality using significance values, Bayesian methods evaluate the probability of the null and alternative hypotheses given the data. Bayesian approaches can take advantage of population genetic model predictions, but these methods are not always robust to departure from their underlying model assumptions (Hermisson 2009; Narum & Hess 2011; Lotterhos & Whitlock 2014). Recalibration methods cannot be directly applied to the variety of Bayesian models available, unless these methods propose to compute significance values (Beaumont & Balding 2004; Günther & Coop 2013; Gautier 2015). When recalibration is not possible, a correct application of Bayesian methods requires that the null or the alternative model fit the data. In standard approaches, model fit is usually checked using posterior predictive tests. While model-checking is commonly addressed in applications such as approximate Bayesian computation (Csilléry *et al.* 2010), it can be difficult to address for genome scan algorithms. The difficulty arises as model fit is performed internally, and observations on the fit of the model are not always available to computer program users.

Link to GWAS

By considering genome scans for selection as a category of genomewide association studies, one can draw on a much broader literature that has grappled with the problem of addressing calibration and multiple testing issues. Traditionally, classical GWAS put substantial efforts on avoiding false positives and on improving power to detect true associations. The eventual development of GWAS graphical methods, which display histograms of significance values and use genomic control to reduce the rate of false discoveries, provides a strong parallel to our approaches in this study. However, there are some differences between regression models used in GWAS and genome scans for selection. First, the direct application of GWAS tests to genome scans for selection, for which polygenic effects can be considered to be the rule, leads to overly conservative tests (Frichot *et al.* 2013). Another difference can be explained in terms of generative models (Listgarten *et al.* 2010). Traditional GWAS approaches investigate the association between genetic polymorphisms and specific individual traits or phenotypes. Most phenotypic traits are heritable, and a part of the variation among individuals can be explained by genetic variance components. Most GWAS approaches correct for confounding effects by estimating a genetic similarity matrix, and use this matrix for the covariance of random effects in a mixed model (Yu *et al.* 2006). Unlike phenotypic traits, ecological variables do not follow any mode of inheritance, and generative GWAS models may be inappropriate in this context. Thus, direct applications of GWAS regression models to detecting ecological selection are not straightforward (Yoder *et al.* 2014). The same remarks apply to the closely related fields of phylogenetic regression methods (Grafen 1989; Harvey & Pagel 1991). We believe that statistical frameworks developed in GWAS (or phylogenetic regression methods) could largely benefit PD and EA tests if their generative models are reformulated for these new

applications, for example by incorporating polygenic effects, or by modelling the covariance matrix of the random effects in a way that accounts for the spatial autocorrelation of ecological variables.

Remarks on the power to reject neutrality

Connections between genome scans for selection and chi-squared tests increase our understanding of which factors could influence the power of neutrality tests. In F_{ST}-based tests, the power of tests is maximized when the sampled population sizes are equal, and having uneven sample sizes generally decreases the power to reject neutrality. This property is an immediate consequence of Fisher's ANOVA *F*-statistic which has minimal variance when the sample sizes are equal. In EA tests, the power of tests increases when individuals are statistically uncorrelated. Thus, the tests have increased power when the geographic coverage of the study area is maximal, and when the sampling design does not cluster individuals into groups. These properties indicate that uneven sampling can decrease the power of EA and PD tests to reject neutrality (Lotterhos & Whitlock 2015). In simulations of EA tests, the tests have less power when the direction of the environmental gradient is parallel to the first principal component of genetic variation (Frichot *et al.* 2015, Appendix S3, Supporting information). This property is a very general feature of linear regression tests, and it explains why approaches testing correlation between genetic variation and population structure inherently lack power to detect weak selection. Typical cases include PD approaches based on F_{ST} in which the first principal component of the genetic variation aligns to population groupings (McVean 2009). In this case, PD tests have limited power, and only the hardest sweeps can be detected. In EA tests, the direction of the ecological gradient does not necessarily align to the first principal axis of genetic variation, and EA tests can detect selection on markers having weaker effects on polygenic traits (Pritchard *et al.* 2010). Predicting which tests will lead to the maximal power in a particular model is, however, difficult to do. While test performance is case-specific, combining several well-calibrated tests can decrease the sensitivity to particular models and lead to robust testing approaches with reasonable power to reject neutrality.

Conclusions

Genome scan methods that properly control for a false discovery rate are important to researchers identifying patterns of natural selection and should be considered critically when researchers ascribe a functional interpretation to a list of candidate loci or choose to pursue experimental validation of 'selected' loci. Implementation of test calibration methods based on inflation factors offer an important validation step to ensure that genome scan tests have been properly calibrated. The main remaining challenges for the interpretation of results in genome scans for selection is to ensure that confounding variables, such as complex population structure, uneven sample sizes, linkage disequilibrium and sequencing biases, have been taken into account. Corrections for these confounding effects and proper test calibration are essential before one takes the step to adjust significance values for multiple comparisons. One powerful approach to correct for confounding variables is to specify error structures in mixed models. With the advent of massive genomic data sets, regression methods that include random effects or latent factors may be the most robust way to provide calibration of test *P*-values and control false discoveries in genome scans for selection.

Acknowledgements

We are grateful to Louis Bernatchez for his invitation to write this manuscript, and to Robin Waples for his useful comments. We are also grateful to three anonymous reviewers for their time and efforts in evaluating our manuscript. Olivier François acknowledges support from Grenoble INP, and from the 'Agence Nationale de la Recherche' (project AFRICROP ANR-13-BSV7-0017). Helena Martins acknowledges support from the 'Ciências sem Fronteiras' scholarship program from the Brazilian government. This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissement d'Avenir.

References

Akey JM, Zhang G, Khang K, Jin L, Shriver MD (2002) Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, **12**, 1805–1814.

Atwell S, Huang YS, Vilhjalmsson BJ *et al.* (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature*, **465**, 627–631.

Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, **7**, 781–791.

Bazin E, Dawson KJ, Beaumont MA (2010) Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model. *Genetics*, **185**, 587–602.

Beaumont MA, Balding DJ (2004) Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, **13**, 969–980.

Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. *Proceedings of the Royal Society London B*, **263**, 1619–1626.

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, **57**, 289–300.

Berry A, Kreitman M (1993) Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the East coast of North America. *Genetics*, **134**, 869–893.

Bonhomme M, Chevalet C, Servin B *et al.* (2010) Detecting selection in population trees: the Lewontin and Krakauer test extended. *Genetics*, **186**, 241–262.

Brown M (1975) A method for combining non-independent, one-sided tests of significance. *Biometrics*, **31**, 987–992.

Caye K, Deist T, Martins H, Michel O, François O (2015) TESS3: fast inference of spatial population structure and genome scans for selection. *Molecular Ecology Resources*. doi: 10.1111/1755-0998.12471. in press

Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. *Genetics*, **185**, 1411–1423.

Csilléry K, Blum MG, Gaggiotti OE, François O (2010) Approximate bayesian computation (ABC) in practice. *Trends in Ecology and Evolution*, **25**, 410–418.

Davis MB, Shaw RG (2001) Range shifts and adaptive responses to quaternary climate change. *Science*, **292**, 673–679.

Davis MB, Shaw RG, Etterson JR (2005) Evolutionary responses to changing climate. *Ecology*, **86**, 1704–1714.

De Mita S, Thuillet A-C, Gay L *et al.* (2013) Detecting selection along environmental gradients: analysis of eight methods and their effectiveness for outbreeding and selfing populations. *Molecular Ecology*, **22**, 1383–1399.

De Villemereuil P, Frichot E, Bazin E, François O, Gaggiotti OE (2014) Genome scan methods against more complex models: when and how much should we trust them? *Molecular Ecology*, **23**, 2006–2019.

De Villemereuil P, Gaggiotti OE (2015) A new F_{ST}-based method to uncover local adaptation using environmental variables. *Methods in Ecology and Evolution*, **6**, 1248–1258.

Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics*, **55**, 997–1004.

Dudoit S, Van Der Laan MJ (2007) *Multiple Testing Procedures with Applications to Genomics*. Springer, New York.

Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2015) Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data. *Molecular Biology and Evolution*, in press.

Efron B (2004) Large-scale simultaneous hypothesis testing: the choice of a null hypothesis. *Journal of the American Statistical Association*, **99**, 96–104.

Fitzpatrick MC, Keller SR (2015) Ecological genomics meets community-level modelling of biodiversity: mapping the genomic landscape of current and future environmental adaptation. *Ecology Letters*, **18**, 1–16.

Foll M, Gaggiotti OE (2008) A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics*, **180**, 977–993.

Frichot E, François O (2015) LEA: an R package for landscape and ecological association studies. *Methods in Ecology and Evolution*, **6**, 925–929.

Frichot E, Schoville SD, Bouchard G, François O (2013) Testing for associations between loci and environmental gradients using latent factor mixed models. *Molecular Biology and Evolution*, **30**, 1687–1699.

Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014) Fast and efficient estimation of individual ancestry coefficients. *Genetics*, **196**, 973–983.

Frichot E, Schoville SD, De Villemereuil P, Gaggiotti OE, François O (2015) Detecting adaptive evolution based on association with ecological gradients: orientation matters!. *Heredity*, **115**, 22–28.

Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. *Genetics*, **195**, 205–220.

Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. *Genetics*, **201**, 1555–1579.

Grafen A (1989) The phylogenetic regression. *Philosophical Transaction of the Royal Society of London Series B, Biological Sciences*, **326**, 119–157.

Haasl RJ, Payseur BA (2015) Fifteen years of genome-wide scans for selection: trends, lessons and unaddressed genetic sources of complication. *Molecular Ecology*. doi: 10.1111/mec.13339

Haldane JBS (1948) The theory of a cline. *Journal of Genetics*, **48**, 277–284.

Hancock AM, Witonsky DB, Gordon AS *et al.* (2008) Adaptations to climate in candidate genes for common metabolic disorders. *PLoS Genetics*, **4**, e32.

Hancock AM, Witonsky DB, Ehler E *et al.* (2010) Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. *Proceedings of the National Academy of Sciences United States of America*, **107**, 8924–8930.

Harvey PH, Pagel MD (1991) *The Comparative Method in Evolutionary Biology*. Oxford University Press, Oxford, UK.

Hedrick PW, Ginevan ME, Ewing EP (1976) Genetic polymorphism in heterogeneous environments. *Annual Review of Ecology and Systematics*, **7**, 1–32.

Hermisson J (2009) Who believes in whole-genome scans for selection. *Heredity*, **103**, 283–284.

Hijmans RJ, Cameron SE, Parra JL, Jones PG, Jarvis A (2005) Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*, **25**, 1965–1978.

Holsinger KE, Weir BS (2009) Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature Review Genetics*, **10**, 639–650.

Huang DW, Sherman BT, Lempicki RA (2009) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, **37**, 1–13.

Huber CD, Nordborg M, Hermisson J, Hellmann I (2014) Keeping it local: evidence for positive selection in Swedish *Arabidopsis thaliana*. *Molecular Biology and Evolution*, **31**, 3026–3039.

Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics*, **18**, 337–338.

Jay F, Manel S, Alvarez N *et al.* (2012) Forecasting changes in population genetic structure of alpine plants in response to global warming. *Molecular Ecology*, **21**, 2354–2368.

Jones M, Forester B, Teufel A *et al.* (2013) Integrating landscape genomics and spatially-explicit approaches to detect loci under selection in clinal populations. *Evolution*, **67**, 3455–3468.

Joost S, Bonin A, Bruford MW *et al.* (2007) A spatial analysis method (SAM) to detect candidate loci for selection: towards a landscape genomics approach to adaptation. *Molecular Ecology*, **16**, 3955–3969.

Jump AS, Penuelas J (2005) Running to stand still: adaptation and the response of plants to rapid climate change. *Ecology Letters*, **8**, 1010–1020.

Kang HM, Sul JH, Service SK *et al.* (2010) Variance component model to account for sample structure in genome-wide association studies. *Nature Genetics*, **42**, 348–354.

Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. *Plant Methods*, **9**, 29.

Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics*, **74**, 175–195.

Listgarten J, Kadie C, Schadt EE, Heckerman D (2010) Correction for hidden confounders in the genetic analysis of gene expression. *Proceedings of the National Academy of Sciences United States of America*, **107**, 16465–16470.

Lotterhos KE, Whitlock MC (2014) Evaluation of demographic history and neutral parameterization on the performance of F_{ST} outlier tests. *Molecular Ecology*, **23**, 2178–2192.

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.

McCarthy MI, Abecasis GR, Cardon LR *et al.* (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, **9**, 356–369.

McVean G (2009) A genealogical interpretation of principal components analysis. *PLoS Genetics*, **5**, e1000686.

Narum SR, Hess JE (2011) Comparison of F_{ST} outlier tests for SNP loci under selection. *Molecular Ecology Resources*, **11**, 184–194.

Nei M (2005) Selectionism and neutralism in molecular evolution. *Molecular Biology and Evolution*, **22**, 2318–2342.

Pearson TA, Manolio TA (2008) How to interpret a genome-wide association study. *Journal of the American Medical Association*, **299**, 1335–1344.

Peng Y, Yang Z, Zhang H *et al.* (2011) Genetic variations in Tibetan populations and high-altitude adaptation at the Himalayas. *Molecular Biology and Evolution*, **28**, 1075–1081.

Platt A, Vilhjálmsson BJ, Nordborg M (2010) Conditions under which genome-wide association studies will be positively misleading. *Genetics*, **186**, 1045–1052.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, **38**, 904–909.

Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, **11**, 459–463.

Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Current Biology*, **20**, R208–R215.

Rellstab C, Gugerli F, Eckert AJ, Hancock AM, Holderegger R (2015) A practical guide to environmental association analysis in landscape genomics. *Molecular Ecology*, **24**, 4348–4370.

Savolainen O, Lascoux M, Merilá J (2013) Ecological genomics of local adaptation. *Nature Reviews Genetics*, **14**, 807–820.

Schoville SD, Bonin A, François O, Lobreaux S, Melodelima C, Manel S (2012) Adaptive genetic variation on the landscape: methods and cases. *Annual Review of Ecology, Evolution and Systematics*, **43**, 23–43.

Simonson TS, Yang Y, Huff CD *et al.* (2010) Genetic evidence for high-altitude adaptation in Tibet. *Science*, **329**, 72–75.

Storey JD, Tibshirani R (2003) Statistical significance for genome-wide studies. *Proceedings of the National Academy of Sciences United States of America*, **100**, 9440–9445.

Storz JF (2005) Using genome scans of DNA polymorphism to infer adaptive population divergence. *Molecular Ecology*, **14**, 671–688.

Strimmer K (2008) fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics*, **24**, 1461–1462.

Teshima KM, Coop G, Przeworski M (2006) How reliable are empirical genomic scans for selective sweeps? *Genome Research*, **16**, 702–712.

Tiffin P, Ross-Ibarra J (2014) Advances and limits of using population genetics to understand local adaptation. *Trends in Ecology and Evolution*, **29**, 673–680.

Vilhjálmsson BJ, Nordborg M (2013) The nature of confounding in genome-wide association studies. *Nature Reviews Genetics*, **14**, 1–2.

Vitalis R, Dawson K, Boursot P (2001) Interpretation of variation across marker loci as evidence of selection. *Genetics*, **158**, 1811–1823.

Vitalis R, Dawson K, Boursot P, Belkhir K (2003) DetSel 1.0: a computer program to detect markers responding to selection. *Journal of Heredity*, **94**, 429–431.

Vitti JJ, Grossman SR, Sabeti PC (2013) Detecting natural selection in genomic data. *Annual Review of Genetics*, **47**, 97–120.

Wang IJ, Glor RE, Losos JB (2013) Quantifying the roles of ecology and geography in spatial genetic divergence. *Ecology Letters*, **16**, 175–182.

Weir BS (1996) *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Associates, Sunderland, USA.

Weir BS, Cockerham CC (1984) Estimating *F*-statistics for the analysis of population structure. *Evolution*, **38**, 1358–1370.

Whitlock MC (2005) Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *Journal of Evolutionary Biology*, **18**, 1368–1373.

Whitlock MC, Lotterhos KE (2015). Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of F_{ST}. *The American Naturalist*, **186**, S24–36. doi: 10.1086/682949

Williams SM, Haines JL (2011) Correcting away the hidden heritability. *Annals of Human Genetics*, **75**, 348–350.

Wright S (1951) The genetical structure of populations. *Annals of Eugenics*, **15**, 323–354.

Yoder JB, Stanton-Geddes J, Zhou P, Briskine R, Young ND, Tiffin P (2014) Genomic signature of local adaptation to climate in *Medicago truncatula*. *Genetics*, **196**, 1263–1275.

Yu J, Pressoir G, Briggs WH *et al.* (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, **38**, 203–208.

O.F. and S.D.S designed the study. O.F., H.M. and K.C. performed the analyses. O.F. and S.D.S. drafted the manuscript. All authors read and approved the final version of the manuscript.

Data accessibility

The data used in our tutorial examples and in supplementary files 1 and 2 have been submitted to Dryad. Their entries are http://datadryad.org/review?doi=doi:10.5061/dryad.78642 The *A. thaliana* data sets used in this study are publicly available from the following link https://github.com/Gregor-Mendel-Institute/atpolydb

Supporting information

Additional supporting information may be found in the online version of this article.

Figure S1 Local adaptation in Scandinavian lines of *A. thaliana*. Top: Histogram of test *P*-values using five runs of LFMM with $K = 2$ latent factors. Bottom: Manhattan plot of minus $\log_{10}$ *P*-values.

Appendix S1 Calibrating and combining results from two statistical models for genome scans for selection.

Appendix S2 Adaptation to climate in European lines of *A. thaliana*.

Appendix S3 Simulation study of genome scans for selection (FDR and power).

Bibliography

Abdellaoui, Abdel et al. (2013). "Population structure, migration, and diversifying selection in the Netherlands". In: *European journal of human genetics* 21.11, p. 1277 (cit. on pp. 88, 101, 111, 112).

Adams, Julian and Richard H. Ward (1973). "Admixture Studies and the Detection of Selection". In: 180.4091, pp. 1137–1143. DOI: 10.1126/science.180.4091.1137 (cit. on p. 60).

Akey, Joshua M et al. (2002). "Interrogating a high-density SNP map for signatures of natural selection". In: *Genome research* 12.12, pp. 1805–1814 (cit. on pp. 33, 38).

Alexander, D. H. and Kenneth Lange (2011). "Enhancements to the ADMIXTURE algorithm for individual ancestry estimation". In: *BMC Bioinformatics* 12.1, p. 246. ISSN: 1471-2105. DOI: 10.1186/1471-2105-12-246. URL: http://dx.doi.org/10.1186/1471-2105-12-246 (cit. on pp. 60, 71).

Alexander, D. H., John Novembre, and K. Lange (2009). "Fast model-based estimation of ancestry in unrelated individuals". In: *Genome Research* 19, pp. 1655–1664. DOI: 10.1101/gr.094052.109 (cit. on pp. 41, 57).

Ascencio-Ibanez, José Trinidad et al. (2008). "Global analysis of Arabidopsis gene expression uncovers a complex array of changes impacting pathogen response and cell cycle during geminivirus infection". In: *Plant physiology* 148.1, pp. 436–454 (cit. on p. 56).

Atwell, Susanna et al. (2010). "Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines". In: *Nature* 465.7298, pp. 627–631 (cit. on pp. 21, 47, 51, 59, 61, 71, 75).

Barton, Nicholas H and Godfrey M Hewitt (1985). "Analysis of hybrid zones". In: *Annual review of Ecology and Systematics* 16.1, pp. 113–148 (cit. on p. 45).

Bazin, Eric, Kevin J Dawson, and Mark A Beaumont (2010). "Likelihood-free inference of population structure and local adaptation in a Bayesian hierarchical model". In: *Genetics* 185.2, pp. 587–602 (cit. on pp. 22, 45, 75, 93).

Beall, Cynthia M, Gary M Brittenham, et al. (1998). "Hemoglobin concentration of high-altitude Tibetans and Bolivian Aymara". In: *American journal of physical anthropology* 106.3, pp. 385–400 (cit. on p. 16).

Beall, Cynthia M, Gianpiero L Cavalleri, et al. (2010). "Natural selection on EPAS1 (HIF2α) associated with low hemoglobin concentration in Tibetan highlanders". In: *Proceedings of the National Academy of Sciences* 107.25, pp. 11459–11464 (cit. on p. 24).

Beaumont, Mark A and David J Balding (2004). "Identifying adaptive genetic divergence among populations from genome scans". In: *Molecular ecology* 13.4, pp. 969–980 (cit. on p. 27).

Beaumont, Mark A and Richard A Nichols (1996). "Evaluating loci for use in the genetic analysis of population structure". In: *Proceedings of the Royal Society of London B: Biological Sciences* 263.1377, pp. 1619–1626 (cit. on pp. 33, 38, 44).

Belkin, Mikhail and Partha Niyogi (2003). "Laplacian eigenmaps for dimensionality reduction and data representation". In: *Neural computation* 15.6, pp. 1373–1396 (cit. on p. 84).

Benjamini, Yoav and Yosef Hochberg (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)*, pp. 289–300 (cit. on pp. 31, 74, 93).

Bersaglieri, Todd et al. (2004). "Genetic signatures of strong recent positive selection at the lactase gene". In: *The American Journal of Human Genetics* 74.6, pp. 1111–1120 (cit. on p. 57).

Bierne, Nicolas, Denis Roze, and John J Welch (2013a). "Pervasive selection or is it...? Why are FST outliers sometimes so frequent?" In: *Molecular ecology* 22.8, pp. 2061–2064 (cit. on p. 59).

— (2013b). "Pervasive selection or is it...? why are FST outliers sometimes so frequent?" In: *Molecular ecology* 22.8, pp. 2061–2064 (cit. on p. 27).

Blumberg, Baruch S and Jana E Hesser (1971). "Loci differentially affected by selection in two American black populations". In: *Proceedings of the National Academy of Sciences* 68.10, pp. 2554–2558 (cit. on p. 60).

Bonhomme, Maxime et al. (2010). "Detecting selection in population trees: the Lewontin and Krakauer test extended". In: *Genetics* 186.1, pp. 241–262 (cit. on p. 27).

Bryc, Katarzyna et al. (2010). "Genome-wide patterns of population structure and admixture in West Africans and African Americans". In: *Proceedings of the National Academy of Sciences* 107.2, pp. 786–791 (cit. on p. 39).

Cai, Deng et al. (2011). "Graph regularized nonnegative matrix factorization for data representation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33.8, pp. 1548–1560 (cit. on pp. 71, 84).

Catinot, Jérémy et al. (2015). "ETHYLENE RESPONSE FACTOR 96 positively regulates Arabidopsis resistance to necrotrophic pathogens by direct binding to GCC elements of jasmonate and ethylene-responsive defence genes". In: *Plant, cell & environment* 38.12, pp. 2721–2734 (cit. on p. 56).

Cavalli-Sforza, Luigi Luca (1966). "Population structure and human evolution". In: *Proceedings of the Royal Society of London. Series B, Biological Sciences* 164.995, pp. 362–379 (cit. on pp. 38, 70).

Cavalli-Sforza, Luigi Luca, Paolo Menozzi, and Alberto Piazza (1994). *The history and geography of human genes*. Princeton university press (cit. on p. 39).

Caye, Kevin et al. (2015). "TESS3: fast inference of spatial population structure and genome scans for selection". In: *Molecular Ecology Resources* 16.2, pp. 540–548. ISSN: 1755-098X. DOI: 10.1111/1755-0998.12471. URL: http://dx.doi.org/10.1111/1755-0998.12471 (cit. on pp. 26, 35, 39, 46, 57).

Charles, Bashira A, Daniel Shriner, and Charles N Rotimi (2014). "Accounting for linkage disequilibrium in association analysis of diverse populations". In: *Genetic epidemiology* 38.3, pp. 265–273 (cit. on p. 88).

Chawade, Aakash et al. (2007). "Putative cold acclimation pathways in Arabidopsis thaliana identified by a combined analysis of mRNA co-expression patterns, promoter motifs and transcription factors". In: *BMC genomics* 8.1, p. 304 (cit. on p. 56).

Chen, Chibiao, Eric Durand, et al. (2007). "Bayesian clustering algorithms ascertaining spatial population structure: a new computer program and a comparison study". In: *Molecular Ecology Resources* 7.5, pp. 747–756 (cit. on p. 57).

Chen, Guo-Bo, Sang Hong Lee, et al. (2016). "EigenGWAS: finding loci under selection through genome-wide association studies of eigenvectors in structured populations". In: *Heredity* 117.1, pp. 51–61 (cit. on pp. 40, 60).

Chen, Hui, Hyun Uk Kim, Hua Weng, et al. (2011). "Malonyl-CoA synthetase, encoded by ACYL ACTIVATING ENZYME13, is essential for growth and development of Arabidopsis". In: *The Plant Cell* 23.6, pp. 2247–2262 (cit. on pp. 56, 70).

Chen, JM Boeger, and Bruce A McDonald (1994). "Genetic stability in a population of a plant pathogenic fungus over time". In: *Molecular Ecology* 3.3, pp. 209–218 (cit. on p. 14).

Chevalier, Christian et al. (2011). "Elucidating the functional role of endoreduplication in tomato fruit development". In: *Annals of Botany* 107.7, pp. 1159–1169 (cit. on p. 55).

Chung, Fan RK (1997). *Spectral graph theory*. 92. American Mathematical Soc. (cit. on p. 85).

Darwin, Charles (1859). "On the origin of species by means of natural selection, or, The preservation of favoured races in the struggle for life." In: DOI: 10.5962/bhl.title.2109. URL: http://dx.doi.org/10.5962/bhl.title.2109 (cit. on p. 38).

DeBlasio, Stacy L, Darron L Luesse, and Roger P Hangarter (2005). "A plant-specific protein essential for blue-light-induced chloroplast movements". In: *Plant Physiology* 139.1, pp. 101–114 (cit. on pp. 55, 56).

Devlin, Bernie and Kathryn Roeder (1999). "Genomic control for association studies". In: *Biometrics* 55.4, pp. 997–1004 (cit. on pp. 32, 33, 43, 74).

Du, Feng-Xing, Archie C Clutter, and Michael M Lohuis (2007). "Characterizing linkage disequilibrium in pig populations". In: *International Journal of Biological Sciences* 3.3, p. 166 (cit. on p. 111).

Duforet-Frebourg, Nicolas et al. (2015). "Detecting genomic signatures of natural selection with principal component analysis: application to the 1000 Genomes data". In: *Molecular biology and evolution* 33.4, pp. 1082–1093 (cit. on pp. 22, 26, 28, 40, 46, 60, 97).

Durand, E. et al. (2009). "Spatial Inference of Admixture Proportions and Secondary Contact Zones". In: *Molecular Biology and Evolution* 26.9, pp. 1963–1973. ISSN: 1537-1719. DOI: 10.1093/molbev/msp106. URL: http://dx.doi.org/10.1093/molbev/msp106 (cit. on pp. 45, 70, 71, 74, 75, 119).

Edelaar, PIM, Pablo Burraco, and IVAN GOMEZ-MESTRE (2011). "Comparisons between QST and FST—how wrong have we been?" In: *Molecular Ecology* 20.23, pp. 4830–4839 (cit. on p. 116).

Edmonds, Christopher A, Anita S Lillie, and L Luca Cavalli-Sforza (2004). "Mutations arising in the wave front of an expanding population". In: *Proceedings of the National Academy of Sciences of the United States of America* 101.4, pp. 975–979 (cit. on p. 116).

Efron, Bradley (2007). "Size, Power and False Discovery Rates". In: *The Annals of Statistics* 35.4, pp. 1351–1377. DOI: 10.1214/009053606000001460. URL: https://doi.org/10.1214/009053606000001460 (cit. on p. 33).

Engelhardt, Barbara E and Matthew Stephens (2010). "Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis". In: *PLoS genetics* 6.9, e1001117 (cit. on p. 39).

Epperson, Bryan K (2003). *Geographical genetics (MPB-38)*. Princeton University Press (cit. on p. 70).

Falush, Daniel, Lucy van Dorp, and Daniel Lawson (2016). "A tutorial on how (not) to over-interpret STRUCTURE/ADMIXTURE bar plots". In: *bioRxiv*, p. 066431 (cit. on p. 59).

Fariello, Maria Inés et al. (2013). "Detecting signatures of selection through haplotype differentiation among hierarchically structured populations". In: *Genetics* 193.3, pp. 929–941 (cit. on p. 27).

Feng, Xiao-Jing, Guo-Fang Jiang, and Zhou Fan (2015). "Identification of outliers in a genomic scan for selection along environmental gradients in the bamboo locust, Ceracris kiangsu". In: *Scientific reports* 5, p. 13758 (cit. on pp. 34, 115, 117, 121).

Foll, Matthieu and Oscar Gaggiotti (2008). "A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective". In: *Genetics* 180.2, pp. 977–993 (cit. on pp. 22, 28, 116).

François, Olivier, Sophie Ancelet, and Gilles Guillot (2006). "Bayesian clustering using hidden Markov random fields in spatial population genetics". In: *Genetics* 174.2, pp. 805–816 (cit. on p. 70).

François, Olivier, Michael GB Blum, et al. (2008). "Demographic history of European populations of Arabidopsis thaliana". In: *PLoS genetics* 4.5, e1000075 (cit. on pp. 59, 83).

François, Olivier and Eric Durand (2010). "Spatially explicit Bayesian clustering models in population genetics". In: *Molecular Ecology Resources* 10.5, pp. 773–784. ISSN: 1755-098X. DOI: 10.1111/j.1755-0998.2010.02868.x. URL: http://dx.doi.org/10.1111/j.1755-0998.2010.02868.x (cit. on pp. 39, 70, 74).

François, Olivier, Helena Martins, et al. (2016). "Controlling false discoveries in genome scans for selection". In: *Molecular Ecology* 25.2, pp. 454–469. ISSN: 0962-1083. DOI: 10.1111/mec.13513. URL: http://dx.doi.org/10.1111/mec.13513 (cit. on pp. 43, 48, 51, 60, 92).

Frichot, Eric and Olivier François (2015). "LEA: An R package for landscape and ecological association studies". In: *Methods in Ecology and Evolution* 6.8. Ed. by BrianEditor O'Meara, pp. 925–929. ISSN: 2041-210X. DOI: 10.1111/2041-

210x.12382. URL: http://dx.doi.org/10.1111/2041-210x.12382 (cit. on pp. 43, 73, 74).

Frichot, Eric, François Mathieu, et al. (2014). "Fast and efficient estimation of individual ancestry coefficients". In: *Genetics* 196.4, pp. 973–983 (cit. on pp. 39, 41, 46, 57, 60, 71, 73, 85, 97).

Frichot, Eric, Sean D Schoville, et al. (2012). "Correcting principal component maps for effects of spatial autocorrelation in population genetic data". In: *Frontiers in genetics* 3, p. 254 (cit. on pp. 119, 120).

Galinsky, Kevin J et al. (2016). "Fast principal-component analysis reveals convergent evolution of ADH1B in Europe and East Asia". In: *The American Journal of Human Genetics* 98.3, pp. 456–472 (cit. on pp. 28, 40, 57, 60).

Griffiths, Anthony JF (2002). *Modern genetic analysis: integrating genes and genomes*. Vol. 1. Macmillan (cit. on p. 13).

Guo, Kun-Mei et al. (2008). "The cyclic nucleotide-gated channel, AtCNGC10, influences salt tolerance in Arabidopsis". In: *Physiologia Plantarum* 134.3, pp. 499–507 (cit. on p. 56).

Haasl, Ryan J and Bret A Payseur (2016). "Fifteen years of genomewide scans for selection: trends, lessons and unaddressed genetic sources of complication". In: *Molecular ecology* 25.1, pp. 5–23 (cit. on p. 115).

Han, Yi et al. (2007). "Evidence of positive selection on a class I ADH locus". In: *The American Journal of Human Genetics* 80.3, pp. 441–456 (cit. on p. 57).

Hancock, Angela M et al. (2011). "Adaptation to climate across the Arabidopsis thaliana genome". In: *Science* 334.6052, pp. 83–86 (cit. on p. 59).

Hao, Wei, Minsun Song, and J. D. Storey (2015). "Probabilistic models of genetic variation in structured populations applied to global human studies". In: *Bioinformatics* 32.5, pp. 713–721. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/btv641. URL: http://dx.doi.org/10.1093/bioinformatics/btv641 (cit. on pp. 40, 60).

Hartl, Daniel L, Andrew G Clark, and Andrew G Clark (1997). *Principles of population genetics*. Vol. 116. Sinauer associates Sunderland (cit. on p. 13).

He, Xin-Jian et al. (2005). "AtNAC2, a transcription factor downstream of ethylene and auxin signaling pathways, is involved in salt stress response and lateral root development". In: *The Plant Journal* 44.6, pp. 903–916 (cit. on p. 56).

Hedrick, Philip (2011). *Genetics of populations*. Jones & Bartlett Learning (cit. on p. 88).

Hewitt, Godfrey (2000). "The genetic legacy of the Quaternary ice ages". In: *Nature* 405.6789, pp. 907–913 (cit. on p. 45).

Hey, Jody and Catarina Pinho (2012). "Population genetics and objectivity in species diagnosis". In: *Evolution* 66.5, pp. 1413–1429 (cit. on p. 13).

Holsinger, Kent E and Bruce S Weir (2009). "Genetics in geographically structured populations: defining, estimating and interpreting F ST". In: *Nature Reviews Genetics* 10.9, p. 639 (cit. on pp. 21, 23, 57, 82, 92).

Horton, Matthew W et al. (2012). "Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel". In: *Nature Genetics* 44.2, pp. 212–216. ISSN: 1546-1718. DOI: 10.1038/ng.1042. URL: http://dx.doi.org/10.1038/ng.1042 (cit. on pp. 38, 55).

Houston, Ross D et al. (2014). "Development and validation of a high density SNP genotyping array for Atlantic salmon (Salmo salar)". In: *BMC genomics* 15.1, p. 90 (cit. on p. 121).

Huang, Qiqing, Sanjay Shete, and Christopher I Amos (2004). "Ignoring linkage disequilibrium among tightly linked markers induces false-positive evidence of linkage for affected sib pair analysis". In: *The American Journal of Human Genetics* 75.6, pp. 1106–1112 (cit. on pp. 88, 109).

Hudson, R. R. (2002). "Generating samples under a Wright-Fisher neutral model of genetic variation". In: *Bioinformatics* 18.2, pp. 337–338. ISSN: 1460-2059. DOI: 10.1093/bioinformatics/18.2.337. URL: http://dx.doi.org/10.1093/bioinformatics/18.2.337 (cit. on pp. 22, 45, 74, 93).

Hudson, Richard R (2004). "ms a program for generating samples under neutral models". In: (cit. on p. 95).

Jay, Flora, Olivier François, et al. (2015). "POPS: A Software for Prediction of Population Genetic Structure Using Latent Regression Models Olivier François Eric Y. Durand". In: *Journal of Statistical Software* 68.9 (cit. on p. 117).

Jay, Flora, Stéphanie Manel, et al. (2012). "Forecasting changes in population genetic structure of alpine plants in response to global warming". In: *Molecular Ecology* 21.10, pp. 2354–2368 (cit. on p. 75).

Jolliffe, Ian T (1986). "Principal Component Analysis". In: *Principal component analysis*. Springer, pp. 115–128 (cit. on p. 28).

Kalinowski, Steven T (2010). "The computer program STRUCTURE does not reliably identify the main genetic clusters within species: simulations and implications for human population structure". In: *Heredity* 106.4, p. 625 (cit. on p. 57).

Kawecki, Tadeusz J and Dieter Ebert (2004). "Conceptual issues in local adaptation". In: *Ecology letters* 7.12, pp. 1225–1241 (cit. on pp. 14, 15).

Kettlewell, Henry Bernard Davies (1956). "A resume of investigations on the evolution of melanism in the Lepidoptera". In: *Proceedings of the Royal Society of London. Series B, Biological Sciences* 145.920, pp. 297–303 (cit. on p. 17).

Kim, Jingu and Haesun Park (2011). "Fast nonnegative matrix factorization: An active-set-like method and comparisons". In: *SIAM Journal on Scientific Computing* 33.6, pp. 3261–3281 (cit. on p. 85).

Kim, Yeon-Ok, Jin Sun Kim, and Hunseung Kang (2005). "Cold-inducible zinc finger-containing glycine-rich RNA-binding protein contributes to the enhancement of freezing tolerance in Arabidopsis thaliana". In: *The Plant Journal* 42.6, pp. 890–900 (cit. on pp. 55, 56).

Kimura, Motoo and George H Weiss (1964). "The stepping stone model of population structure and the decrease of genetic correlation with distance". In: *Genetics* 49.4, p. 561 (cit. on p. 70).

Kliman, Richard, Bob Sheehy, and Joanna Schultz (2008). "Genetic drift and effective population size". In: *Nature Education* 1.3, p. 3 (cit. on p. 15).

Kruuk, LEB et al. (1999). "A comparison of multilocus clines maintained by environmental adaptation or by selection against hybrids". In: *Genetics* 153.4, pp. 1959–1971 (cit. on p. 116).

Landguth, EL et al. (2010). "Quantifying the lag time to detect barriers in landscape genetics". In: *Molecular ecology* 19.19, pp. 4179–4191 (cit. on p. 22).

Lao, Oscar et al. (2008). "Correlation between genetic and geographic structure in Europe". In: *Current Biology* 18.16, pp. 1241–1248 (cit. on p. 38).

Larson, Greger et al. (2014). "Current perspectives and the future of domestication studies". In: *Proceedings of the National Academy of Sciences* 111.17, pp. 6139–6146 (cit. on p. 121).

Laurie, Cathy C et al. (2010). "Quality control and quality assurance in genotypic data for genome-wide association studies". In: *Genetic epidemiology* 34.6, pp. 591–602 (cit. on pp. 34, 88, 89, 111).

Lenormand, Thomas (2002). "Gene flow and the limits to natural selection". In: *Trends in Ecology & Evolution* 17.4, pp. 183–189 (cit. on p. 14).

Levy, Stuart B (1998). "Antimicrobial resistance: Bacteria on the defence: Resistance stems from misguided efforts to try to sterilise our environment". In: *BMJ: British Medical Journal* 317.7159, p. 612 (cit. on p. 16).

Lewontin, R. C. and Jesse Krakauer (1973). "Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms". In: *Genetics* 74.1, pp. 175–195. ISSN: 0016-6731. eprint: http://www.genetics.org/content/74/1/175.full.pdf. URL: http://www.genetics.org/content/74/1/175 (cit. on pp. 21, 26, 27, 33, 38, 44).

Li and Matthew Stephens (2003). "Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data". In: *Genetics* 165.4, pp. 2213–2233 (cit. on p. 112).

Lien, Sigbjørn, Lars Gidskehaug, et al. (2011). "A dense SNP-based linkage map for Atlantic salmon (Salmo salar) reveals extended chromosome homeologies and striking differences in sex-specific recombination patterns". In: *BMC genomics* 12.1, p. 615 (cit. on p. 121).

Lien, Sigbjørn, Ben F Koop, et al. (2016). "The Atlantic salmon genome provides insights into rediploidization". In: *Nature* 533.7602, p. 200 (cit. on p. 121).

Liu, Lei et al. (2017). "A genome scan for selection signatures comparing farmed Atlantic salmon with two wild populations: Testing colocalization among outlier markers, candidate genes, and quantitative trait loci for production traits". In: *Evolutionary applications* 10.3, pp. 276–296 (cit. on pp. 121, 122).

Long, Jeffrey C (1991). "The genetic structure of admixed populations." In: *Genetics* 127.2, pp. 417–428 (cit. on pp. 39, 57, 60).

Lotterhos, Katie E. and Michael C. Whitlock (2015). "The Relative Power of Genome Scans To Detect Local Adaptation Depends on Sampling Design and Statistical Method". In: *Molecular Ecology* 24.5, pp. 1031–1046. DOI: 10.1111/mec.13100. URL: https://doi.org/10.1111/mec.13100 (cit. on pp. 26, 45, 46, 49).

Lu, Yan, Jun Zhu, and Pengyuan Liu (2005). "A two-step strategy for detecting differential gene expression in cDNA microarray data". In: *Current genetics* 47.2, pp. 121–131 (cit. on p. 55).

Lucotte, Elise A et al. (2016). "Detection of allelic frequency differences between the sexes in humans: a signature of sexually antagonistic selection". In: *Genome biology and evolution* 8.5, pp. 1489–1500 (cit. on p. 101).

Luu, Keurcien, Eric Bazin, and Michael GB Blum (2017a). "pcadapt: an R package to perform genome scans for selection based on principal component analysis". In: *Molecular ecology resources* 17.1, pp. 67–77 (cit. on pp. 22, 28, 48, 91, 119, 120, 122).

— (2017b). "pcadapt: an R package to perform genome scans for selection based on principal component analysis". In: *Molecular ecology resources* 17.1, pp. 67–77 (cit. on pp. 40, 46, 93, 97).

Majerus, Michael EN (2009). "Industrial melanism in the peppered moth, Biston betularia: an excellent teaching example of Darwinian evolution in action". In: *Evolution: Education and Outreach* 2.1, pp. 63–74 (cit. on pp. 15, 19).

Malécot, Gustave et al. (1948). "The mathematics of heredity." In: *The mathematics of heredity.* (Cit. on p. 70).

Manel, Stéphanie, Stéphane Joost, et al. (2010). "Perspectives on the use of landscape genetics to detect genetic adaptive variation in the field". In: *Molecular Ecology* 19.17, pp. 3760–3772 (cit. on p. 79).

Manel, Stéphanie, Michael K Schwartz, et al. (2003). "Landscape genetics: combining landscape ecology and population genetics". In: *Trends in ecology & evolution* 18.4, pp. 189–197 (cit. on p. 38).

Maronna, Ricardo A and Ruben H Zamar (2002). "Robust estimates of location and dispersion for high-dimensional datasets". In: *Technometrics* 44.4, pp. 307–317 (cit. on pp. 30, 91, 120).

Martins, Helena et al. (2016). "Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics". In: *Molecular Ecology* 25.20, pp. 5029–5042. ISSN: 0962-1083. DOI: 10.1111/mec.13822. URL: http://dx.doi.org/10.1111/mec.13822 (cit. on pp. 35, 90, 112).

Mathieson, Iain et al. (2015). "Genome-wide patterns of selection in 230 ancient Eurasians". In: *Nature* 528.7583, pp. 499–503 (cit. on p. 60).

Mitchell-Olds, Thomas and Johanna Schmitt (2006). "Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis". In: *Nature* 441.7096, pp. 947–952 (cit. on p. 47).

Mochizuki, Susumu et al. (2005). "The Arabidopsis WAVY GROWTH 2 protein modulates root bending in response to environmental stimuli". In: *The Plant Cell* 17.2, pp. 537–547 (cit. on p. 79).

Montano, Valeria and Thibaut Jombart (2017). "An Eigenvalue test for spatial principal component analysis". In: *BMC bioinformatics* 18.1, p. 562 (cit. on p. 119).

Muhlenbock, Per, Barbara Karpinska, and Stanislaw Karpinski (2007). "Oxidative stress and redox signalling in plants". In: *eLS* (cit. on p. 55).

Nei, Masatoshi (2005). "Selectionism and neutralism in molecular evolution". In: *Molecular biology and evolution* 22.12, pp. 2318–2342 (cit. on p. 115).

Nelson, Matthew R et al. (2008). "The Population Reference Sample, POPRES: a resource for population, disease, and pharmacological genetics research". In: *The American Journal of Human Genetics* 83.3, pp. 347–358 (cit. on pp. 21, 47).

Novembre, John et al. (2008). "Genes mirror geography within Europe". In: *Nature* 456.7219, pp. 274–274. ISSN: 1476-4687. DOI: 10.1038/nature07566. URL: http://dx.doi.org/10.1038/nature07566 (cit. on p. 38).

Okasha, Samir (2006). "Population genetics". In: *The Stanford Encyclopedia of Philosophy (Winter 2016 Edition)* (cit. on p. 13).

Patterson, Nick, Alkes L. Price, and David Reich (2006). "Population Structure and Eigenanalysis". In: *PLoS Genetics* 2.12, e190. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.0020190. URL: http://dx.doi.org/10.1371/journal.pgen.0020190 (cit. on pp. 39, 59).

Petry, Doug (1983). "The effect on neutral gene flow of selection at a linked locus". In: *Theoretical population biology* 23.3, pp. 300–313 (cit. on p. 93).

Pritchard, Jonathan K, Matthew Stephens, and Peter Donnelly (2000). "Inference of population structure using multilocus genotype data". In: *Genetics* 155 (cit. on pp. 34, 39, 40, 57, 70, 115).

Privé, Florian et al. (2017). "Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr". In: *Bioinformatics*, bty185 (cit. on pp. 89, 101, 122).

Puechmaille, Sebastien J (2016). "The program structure does not reliably recover the correct population structure when sampling is uneven: subsampling and new estimators alleviate the problem". In: *Molecular ecology resources* 16.3, pp. 608–627 (cit. on p. 59).

Purcell, Shaun et al. (2007). "PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses". In: *The American Journal of Human Genetics* 81.3 (cit. on pp. 89, 101).

Qanbari, S et al. (2010). "The pattern of linkage disequilibrium in German Holstein cattle". In: *Animal genetics* 41.4, pp. 346–356 (cit. on p. 88).

Radin, Ivan et al. (2015). "The Arabidopsis COX11 homolog is essential for cytochrome c oxidase activity". In: *Frontiers in plant science* 6 (cit. on p. 56).

Raj, Anil, Matthew Stephens, and Jonathan K. Pritchard (2014). "fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets". In: *Genetics* 197.2, pp. 573–589. ISSN: 1943-2631. DOI: 10.1534/genetics.114.164350. URL: http://dx.doi.org/10.1534/genetics.114.164350 (cit. on pp. 41, 57, 60, 71).

Rajjou, Loic et al. (2006). "Proteomic investigation of the effect of salicylic acid on Arabidopsis seed germination and establishment of early defense mechanisms". In: *Plant physiology* 141.3, pp. 910–923 (cit. on p. 56).

Reed, Eric et al. (2015). "A guide to genome-wide association analysis and post-analytic interrogation". In: *Statistics in medicine* 34.28, pp. 3769–3792 (cit. on pp. 89, 111).

Roth, Charlotte and Marcel Wiermer (2012). "Nucleoporins Nup160 and Seh1 are required for disease resistance in Arabidopsis". In: *Plant signaling & behavior* 7.10, pp. 1212–1214 (cit. on p. 56).

Savolainen, Outi, Martin Lascoux, and Juha Merila (2013). "Ecological genomics of local adaptation". In: *Nature Reviews Genetics* 14.11, pp. 807–820 (cit. on pp. 14–16).

Schoville, Sean D et al. (2012). "Adaptive genetic variation on the landscape: methods and cases". In: *Annual Review of Ecology, Evolution, and Systematics* 43, pp. 23–43 (cit. on p. 38).

Slatkin, Montgomery (2008). "Linkage disequilibrium—understanding the evolutionary past and mapping the medical future". In: *Nature Reviews Genetics* 9.6, pp. 477–485 (cit. on p. 88).

Slatkin, Montgomery and Lianne Voelm (1991). "FST in a hierarchical island model." In: *Genetics* 127.3, pp. 627–629 (cit. on p. 22).

Snelling, WM et al. (2017). "Linkage disequilibrium among commonly genotyped SNP variants detected from bull sequence". In: *Animal genetics* 48.5, pp. 516–522 (cit. on p. 88).

Sokal, Robert R and F James Rohlf (1981). "Biometry: the principles and practice of statistics in biological research 2nd edition." In: (cit. on p. 42).

Speed, Doug et al. (2012). "Improved heritability estimation from genome-wide SNPs". In: *The American Journal of Human Genetics* 91.6, pp. 1011–1021 (cit. on p. 88).

Storey, John D and Robert Tibshirani (2003). "Statistical significance for genomewide studies". In: *Proceedings of the National Academy of Sciences* 100.16, pp. 9440–9445 (cit. on pp. 31, 47, 51, 102).

Storz, Jay F (2005). "INVITED REVIEW: Using genome scans of DNA polymorphism to infer adaptive population divergence". In: *Molecular ecology* 14.3, pp. 671–688 (cit. on p. 115).

Sun, Chih-Wen and Judy Callis (1997). "Independent modulation of Arabidopsis thaliana polyubiquitin mRNAs in different organs and in response to environmental changes". In: *The Plant Journal* 11.5, pp. 1017–1027 (cit. on p. 56).

Tang, Hua, Shweta Choudhry, et al. (2007). "Recent genetic selection in the ancestral admixture of Puerto Ricans". In: *The American Journal of Human Genetics* 81.3, pp. 626–633 (cit. on pp. 39, 60).

Tang, Hua, Jie Peng, et al. (2005). "Estimation of individual admixture: Analytical and study design considerations". In: *Genetic Epidemiology* 28.4, pp. 289–301. ISSN: 1098-2272. DOI: 10.1002/gepi.20064. URL: http://dx.doi.org/10.1002/gepi.20064 (cit. on p. 57).

Taranger, Geir Lasse et al. (2014). "Risk assessment of the environmental impact of Norwegian Atlantic salmon farming". In: *ICES Journal of Marine Science* 72.3, pp. 997–1021 (cit. on p. 121).

Tiffin, Peter and Jeffrey Ross-Ibarra (2014). "Advances and limits of using population genetics to understand local adaptation". In: *Trends in ecology & evolution* 29.12, pp. 673–680 (cit. on p. 115).

Turcotte, Martin M et al. (2017). "The eco-evolutionary impacts of domestication and agricultural practices on wild species". In: *Phil. Trans. R. Soc. B* 372.1712, p. 20160033 (cit. on p. 121).

Villemereuil, Pierre and Oscar E Gaggiotti (2015). "A new FST-based method to uncover local adaptation using environmental variables". In: *Methods in Ecology and Evolution* 6.11, pp. 1248–1258 (cit. on pp. 116, 117, 119).

Visser, Mijke, Manfred Kayser, and Robert-Jan Palstra (2012). "HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter". In: *Genome research* 22.3, pp. 446–455 (cit. on p. 57).

Vitti, Joseph J, Sharon R Grossman, and Pardis C Sabeti (2013). "Detecting natural selection in genomic data". In: *Annual review of genetics* 47, pp. 97–120 (cit. on pp. 38, 115).

Wang, Yi et al. (2008). "Transcriptome analyses show changes in gene expression to accompany pollen germination and tube growth in Arabidopsis". In: *Plant physiology* 148.3, pp. 1201–1211 (cit. on p. 56).

Waples, Robin S and Oscar Gaggiotti (2006). "What is a population? An empirical evaluation of some genetic methods for identifying the number of gene pools and their degree of connectivity". In: *Molecular ecology* 15.6, pp. 1419–1439 (cit. on pp. 34, 38, 115).

Weigel, Detlef and Richard Mott (2009). "The 1001 genomes project for Arabidopsis thaliana". In: *Genome biology* 10.5, p. 107 (cit. on p. 38).

Weir, Bruce (1996). *Genetic data analysis II: methods for discrete population genetic data*. Vol.2. Sinauer Associates, 445p. ISBN: 0878939024 (cit. on pp. 24, 43, 74, 92).

Weir, Bruce S, Lon R Cardon, et al. (2005). "Measures of human population structure show heterogeneity among genomic regions". In: *Genome research* 15.11, pp. 1468–1476 (cit. on pp. 33, 38).

Weir, Bruce S and C Clark Cockerham (1984). "Estimating F-statistics for the analysis of population structure". In: *evolution* 38.6, pp. 1358–1370 (cit. on p. 57).

Wollstein, Andreas and Oscar Lao (2015). "Detecting individual ancestry in the human genome." In: *Investigative genetics* 6, p. 7 (cit. on pp. 41, 60, 71).

Wright, Sewall (1943). "Isolation by distance". In: *Genetics* 28.2, p. 114 (cit. on p. 70).

— (1949). "The genetical structure of populations". In: *Annals of Human Genetics* 15.1, pp. 323–354 (cit. on pp. 21, 23, 41).

Wu, Tianyi and Bengt Kayser (2006). "High altitude adaptation in Tibetans". In: *High Altitude Medicine & Biology* 7.3, pp. 193–208 (cit. on pp. 16, 17).

Xin, Zhanguo et al. (2007). "Arabidopsis ESK1 encodes a novel regulator of freezing tolerance". In: *The Plant Journal* 49.5, pp. 786–799 (cit. on p. 56).