Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics

Helena Martins, Kevin Caye, Keurcien Luu, Michael G.B. Blum, Olivier François

May 20, 2016

Université Grenoble-Alpes, Centre National de la Recherche Scientifique, TIMC-IMAG UMR 5525, Grenoble, 38042, France.

Running Title: Identifying outlier loci in admixed and in continuous populations

Keywords: Inference of Population Structure, Geographic Variation, Genome Scans for Selection, Admixture.

Corresponding Author: Olivier François

Université Grenoble-Alpes,
TIMC-IMAG, UMR CNRS 5525,
Grenoble, 38042, France.
+334 56 52 00 25 (ph.)
+334 56 52 00 55 (fax)
olivier.francois@imag.fr
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 paper, 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, by re-analyzing genomic data from European lines of the plant species *Arabidopsis thaliana*, and by re-analyzing human genomic data from the population reference sample, POPRES.
1 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, 1909). Identifying genomic regions that have been the targets of natural selection is one of the most important challenges 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 & Krakauer, 1973; Beaumont & Nichols, 1996; Akey et al., 2002; Weir et al., 2005). For selectively neutral SNPs, $F_{ST}$ is determined by genetic drift, which affects 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 require 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 & 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 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 & 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 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 (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 their most probable ancestral genetic group (Engelhardt & 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 & Durand, 2010; Wright, 1943).

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 (Holsinger & Weir, 2009). 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 et al., 2014; Caye et al., 2016). 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., 2016; Dufreney-Frebourg et al., 2016; Galinsky et al., 2016; Luu et al., in prep.).
2 \(F\)-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, \(\ell\), 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. Assuming \(K\) predefined subpopulations, the fixation index, \(F_{ST}\), can be calculated according to S. Wright’s definition as follows (Wright, 1951)

\[
F_{ST} = 1 - \frac{H_S}{H_T},
\]

where \(H_S = \sum_{k=1}^{K} n_k f_k (1 - f_k) / n\), \(H_T = f(1 - f)\), \(n_k\) is the sample size, \(f_k\) is the allele frequency in subpopulation \(k\), and \(f\) is the allele frequency in the total population.

A new definition of \(F_{ST}\). A classical definition of the fixation index, \(F_{ST}\), corresponds to the proportion of the variance in sampled allele frequency explained by ancestral population structure (or population indicators)

\[
F_{ST} = R^2 = \frac{\sigma_T^2 - \sigma_S^2}{\sigma_T^2}
\]

where \(\sigma_T^2\) is the total variance and \(\sigma_S^2\) is the residual variance (Holsinger & Weir, 2009). This definition of \(F_{ST}\), which uses a linear regression framework, can be extended to models with admixed individuals. Suppose that a population contains admixed individuals, and the source populations are unknown. Assume that the individual ancestry coefficients, \(Q\), and the ancestral population frequencies, \(F\), are estimated from an ancestry estimation algorithm, such as \textit{structure} (Pritchard et al., 2000). 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 at random at the moment of the admixture event. Let \(f_k\) be the allele frequency in ancestral population \(k\). Omitting the locus subscript \(\ell\), 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 modeling the parental gametes. The conditional distribution of \( x_1 \) (resp. \( x_2 \)) is such that \( \text{prob}(x_1 = 1|\text{Anc}_1 = k) = f_k \) (Anc is an integer value between 1 and \( K \) representing the hidden ancestry of each gamete).

The sampled allele frequency is defined as \( x = y/2 \) (\( x \) in \( 0, 1/2, 1 \)). Thus the expected value of the random variable \( x \) is given by the following formula

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

where \( q_k = \text{prob}(\text{Anc} = k) \). The total variance of \( x \) is equal to

\[ 2\sigma^2_T = 2\text{Var}[x] = f(1 - f). \]

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 residual variance, \( \sigma^2_S \), we consider that the two gametes originate from the same ancestral population. Assuming Hardy-Weinberg equilibrium in the ancestral populations, the residual variance can be computed as follows

\[ 2\sigma^2_S = \sum_{k=1}^{K} q_k f_k(1 - f_k), \]

and the formula for \( F_{ST} \) becomes

\[ F_{ST} = 1 - \frac{\sum_{k=1}^{K} q_k f_k(1 - f_k)}{f(1 - f)}. \]  \hspace{1cm} (1)

The above definition of \( F_{ST} \) for admixed populations is obviously related to the original definition of Wright’s fixation index, and it is also related to Long’s formula which assumes known ancestral populations (Long, 1991). The estimates of the sample sizes, \( n_k \), can be obtained by setting \( n_k = nq_k \), and the sampled allele frequencies are replaced by their ancestral allele frequencies. For haploid organisms, for which the genotype is coded 0 or 1, the definition of \( F_{ST} \) follows
the same developments. In this case, the definition of $F_{ST}$ for a structured population corresponds to the squared correlation, $R^2$, in the regression of observed allele frequencies on subpopulation indicators.

**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 et al., 2014). Frichot’s algorithm assumes that the sampled genotype frequencies can be modelled by a mixture of ancestral genotype frequencies as follows

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

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

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

Frichot’s matrix factorization algorithm is much faster than the Monte-Carlo algorithm implemented in structure, and the estimates computed by this method weaken the Hardy-Weinberg equilibrium assumptions made by other methods. As a result of the above equations, estimates of $Q$ and $F$ matrices obtained by matrix factorization algorithms could replace those obtained by the program structure advantageously for large SNP data sets (Wollstein & Lao, 2015).

**Population differentiation tests.** The regression framework developed in the previous paragraph leads to a direct approximation of the distribution of $F_{ST}$ under the null-hypothesis of random mating. We define the squared $z$-scores as follows
\[ z^2 = (n - K) \frac{F_{ST}}{1 - F_{ST}}. \]

Then by classical arguments for regression models, 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 (Sokal & Rohlf, 2012). In genome scans for selection, we assume that \( n \) is large enough to approximate the distribution of squared \( z \)-scores as a \( \chi^2 \) distribution with \( K - 1 \) degrees of freedom

\[ z^2 \sim \chi^2(K - 1). \]

To calibrate the null-hypothesis, we adopt an empirical null-hypothesis testing approach which evaluates the level of population differentiation expected at selectively neutral SNPs (François et al., 2016). Following GWAS approaches, the test calibration is achieved after computing the genomic inflation factor, defined by the median of the squared \( z \)-scores divided by the median of a \( \chi^2 \) distribution with \( K - 1 \) degrees of freedom (genomic control, Devlin & Roeder (1999)).

### 3 Simulation experiments and data sets

**Simulated data sets.** We simulated of 10,000 unlinked SNPs for 200 individuals based on Wright’s two-island models. Each island corresponded to an ancestral population prior to admixture. Two distinct scenarios were considered. In the first scenario, the proportion of loci under selection was equal to 5% whereas this proportion was equal to 10% in the second scenario. To mimic genetic variation at outlier loci, we used that a locus with reduced levels of gene flow is expected to have an increased \( F_{ST} \) value (Bazin et al., 2010; Caye et al., 2016). Thus, we assumed that adaptive SNPs had a migration rate smaller than the migration rate at selectively neutral SNPs (\( 4Nm = 20, 15, 10, 5 \) for neutral SNPs and \( 4Nm_s = 0.1, 0.25, 0.5, 1 \) for adaptive SNPs). A total number of 32 different data sets were generated by using the computer program ms (Hudson, 2002).
Using ancestral populations from two-island models, a sample from a unique continuous population was created by simulating admixture of the two populations. The model of admixture was based on a longitudinal gradient of ancestry (Durand et al., 2009). Geographic coordinates \((x_i, y_i)\) were created for each individual from two Gaussian distributions centered around two centroids put at distance 2 on the longitudinal axis (standard deviation [SD] = 1). As it happens in a secondary contact zone, we assumed that the admixture proportions had a sigmoidal shape across geographic 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).

**Computer programs** We performed genome scans for selection using three methods: snmf (Frichot et al., 2014), tess3 (Caye et al., 2016), padapt (Luu et al., in prep.; Duforet-Frebourg et al., 2016). 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., 2016). For snmf and for tess3, we used \(K = 2\) ancestral populations.

This value of \(K\) corresponds to the minimum of the cross-entropy criterion when \(K\) was varied in the range 1 to 6. 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 the \(F_{ST}\) statistics according to formula (1). We compared the results of snmf and tess3 with the program padapt (Luu et al., in prep.). The test statistic of this new version of padapt is the Manhalobis distance relative to the \(z\)-scores obtained after regressing the SNP frequencies on the \(K - 1\) principal components. We used padapt with the first principal component. As for snmf and for tess3, test calibration in padapt was
based on the computation of the genomic inflation factor (genomic control). 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 2 clusters and a chi-squared distribution with one degree of freedom after recalibration of the null-hypothesis. 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 the analysis.

**Candidate lists.** False Discovered Rate (FDR) control algorithms, as described by Storey & Tibshirani (2003), were applied after the recalibration of the test significance values, yielding lists of outlier loci. Before applying FDR control methods, the histograms of test significance values were checked to display uniformly distributed random variables when the null hypothesis is correct.

**Real data sets.** To provide an application of our method to natural populations, we reanalyzed data from the model plant organism *Arabidopsis thaliana*. 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). 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 & Schmitt, 2006). Ecotypes from Northern Scandinavia were not included in the data (14 ecotypes representing a divergent genetic cluster in the original data set). We applied genome scans for selection using *snmf*, *tess3* and *pcadapt* with $K = 2$ ancestral populations and one principal component. In addition, we analyzed human genetic data for 1,385 European individuals genotyped at 447k SNPs (Nelson et al., 2008). We applied genome scans for selection using *snmf* and *pcadapt* with $K = 2$ ancestral populations and one principal component.

## 4 Results

**Simulations of admixed individuals.** We evaluated the performances of genome scans using tests based on *snmf*, *tess3*, *pcadapt*, and $F_{ST}$, in the presence of admixed individuals. Considering $q$-values thresholds 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 1 for data sets with 5% of outliers, Figure S1 for data sets with 10% of outliers).

These observations confirmed that the use of genomic inflation factors leads to overly conservative
tests (François 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 4 tests in each simulation scenario. As we
expected from the simulation process, the tests had higher power when the relative levels of the
selection intensity were higher. For $4Nm = 5$ and $4Nm_s = .1,.25,.5$, and 1, the power of the
tests for snmf, tess3, pcadapt was close to 27% (expected FDR equal to $\alpha = 0.1$, Figure 2A
for data sets with 5% of outliers, Figure S2A for data sets with 10% of outliers). 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 was in the range 40% - 45% for snmf,
tess3, pcadapt and equal to 26% for the $F_{ST}$ test (Figure 2B (5% of outlier loci), Figure 2B
(10% of outlier loci)). For $4mN \geq 15$, corresponding to the highest selection rates, the power was
approximately equal to 50% for all methods considered (Figure 2C and D ((5% of outlier loci),
Figure 2C and D (10% of outlier loci)). The relatively low power values confirmed that the tests
were conservative, and non-neutral loci were difficult to detect. To provide an upper bound on
the power of an outlier test 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 (Figure 2). 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. This
experiment confirmed that the use of approaches that estimate ancestry coefficients is appropriate
when no subpopulation can be predefined.

**Biological data analysis**

**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, ecotypes from northern Scandinavia, which were grouped together 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 2 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 Central, Eastern Europe and Southern Sweden. For implementing genome scans for selection, we used 2 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 used in the snmf and tess3 tests is around 0.09 (see François et al. 2016). For the three methods, the Manhattan plots exhibited peaks at the same chromosome positions (Figure 3). For an expected FDR level equal to 1%, the Storey and Tibshirani algorithm resulted in a list of 572 chromosome positions for the snmf tests, 882 for the tess3 tests (Figure 3). The test based on PCA was more conservative. The difference between the tests could be attributed to the estimation of the genomic inflation factor which differs for PCA methods (see Venn diagrams in Figure S3).

Table 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 three methods. Figure 4 displays a Manhattan plot for the plant genome showing the main outlier loci detected by our genome scans for selection. For chromosome 1, the list contains SNPs in the gene AT1G80680 involved in resistance against 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 acclimatatory responses (Mühlénbock 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 & Kang, 2005), and in the gene AT5G20610 which is involved in response to blue light (DeBlasio et al., 2005).

**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. For an expected FDR equal to 10%, we found 205 outlier loci using `snmf` tests, and 165 outlier loci with `pcadapt` (Figure 5). For chromosome 2, the most important 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 6).

### 5 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 et al., 2005; Chen et al., 2007; Alexander et al., 2009; Raj et al., 2014; Frichot et al., 2014; Caye et al., 2016). 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 & Cockerham, 1984; Holsinger & Weir, 2009).

The estimator for $F_{ST}$ presented here is related to the estimator proposed by Long (1991) for population data. Long’s estimator is obtained from the variance of allele frequencies with respect
to their expectations based on an admixture model, that enable 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.

Assuming that a large number of SNPs are genotyped across multiple populations, the calibration of statistical tests of neutrality did 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 where 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. 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).

Our reanalysis of European A. thaliana data 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 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). Statistics extending the definition of $F_{ST}$ were also proposed for PCA (Hao et al., 2016; Duforet-Frebourg et al., 2016; Galinsky 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 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 & Hesser, 1971; Adams & Ward, 1973; Tang 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 the new methods proposed 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 (Mathieson et al., 2015).

While only minor differences between our results of genome scans with 4 methods were observed, the results might be still sensitive to the algorithm used to estimating the ancestry matrices. Wollstein & Lao (2015) performed an extensive comparison of 3 recently proposed ancestry estimation methods, admixture, faststructure, snmf (Alexander & Lange, 2011; Raj et al., 2014; Frichot 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 et al. (2016).

References


pathways in *Arabidopsis thaliana* identified by a combined analysis of mRNA co-expression 
patterns, promoter motifs and transcription factors. *BMC Genomics* 8, 1.

Chen C, Durand E, Forbes F, François O (2007). Bayesian clustering algorithms ascertaining 
spatial population structure: A new computer program and a comparison study. *Molecular 

ENZYME13, is essential for growth and development of *Arabidopsis*. *The Plant Cell* 23(6), 
2247–2262.

Chevalier C, Nafati M, Mathieu-Rivet E, et al. (2011). Elucidating the functional role of endoredu-

Darwin C. *On The Origin of Species by Means of Natural Selection* John Murray, London, UK, 
1909.


Dufourt-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2016). Detecting genomic signatures 
of natural selection with principal component analysis: Application to the 1000 Genomes data. 
*Molecular Biology and Evolution* 33(4), 1082–1093.


Engelhardt BE, Stephens M (2010). Analysis of population structure: A unifying framework and 
novel methods based on sparse factor analysis. *PLoS Genetics* 6(9), e1001117. doi: 10.1371/jour-
nal.pgen.1001117.


6 Figures and Tables

**Figure 1.** Observed false discovery rates. The tests are based on (A) **snmf**, (B) **tess3**, (C) **pcadapt**, (D) **F_{ST}**. Sixteen data sets containing 5% of outlier loci were used in each panel.
Figure 2. Power values of snmf, tess3, padapt (Factor methods) and classical $F_{ST}$ tests with assignment and prior to admixture. All data sets contained 5% of outlier loci. Considering an expected FDR of $\alpha = 0.1$: (A) Power values for the case $4N_{m} = 5$. The $F_{ST}$ test based on assignment of individuals to their most likely cluster failed to detect outlier loci. (B) Power values for the case $4N_{m} = 10$. (C) Power values for the case $4N_{m} = 15$. (D) Power values for the case $4N_{m} = 20$. 
Figure 3. Manhattan plots of minus log10(p-values) for the *A. thaliana*. Considering the tests using: (A) snmf, (B) tess3 and (C) pcadapt.
Figure 4. Manhattan plot of minus log10(p-values) for the *A. thaliana*. Candidate loci detected by genome scans for selection are colored in red (expected FDR level of 1%).
Figure 5. Manhattan plots of minus log10(p-values) for the 22 chromosomes of the POPRES data set. Considering the tests using: (A) snmf and (B) pcadapt.
Figure 6. Manhattan plot of minus log10(p-values) for the POPRES data. Candidate loci detected by genome scans for selection are colored in red (expected FDR level of 10%)
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Position (kb)</th>
<th>Gene</th>
<th>Unknown</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>132130</td>
<td>AT1G01340</td>
<td>Salt tolerance</td>
<td>Guo et al. (2008)</td>
</tr>
<tr>
<td></td>
<td>490925</td>
<td>AT1G02410</td>
<td>Plant growth and pollen germination</td>
<td>Radin et al. (2015)</td>
</tr>
<tr>
<td></td>
<td>2191723</td>
<td>AT1G07140(SIRANBP)</td>
<td>Encodes a putative ran-binding protein</td>
<td>Wang et al. (2008)</td>
</tr>
<tr>
<td></td>
<td>10779171</td>
<td>AT1G80470</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26503961</td>
<td>AT1G70340</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29516980</td>
<td>AT1G78450</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>30324008</td>
<td>AT1G80680</td>
<td>Defense response</td>
<td>Roth &amp; Wiermer (2012)</td>
</tr>
<tr>
<td>2</td>
<td>7995729</td>
<td>AT2G18440 (AtGUT15)</td>
<td>Encodes a noncoding RNA</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>2048905</td>
<td>AT3G06580 (GAL1)</td>
<td>Galactose metabolic process</td>
<td>Wang et al. (2008)</td>
</tr>
<tr>
<td></td>
<td>3772311</td>
<td>AT3G11920</td>
<td>Cell redox homeostasis</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5476074</td>
<td>AT3G16170 (AAE13)</td>
<td>Fatty acid biosynthetic process</td>
<td>Chen et al. (2011)</td>
</tr>
<tr>
<td></td>
<td>15095731</td>
<td>AT3G10150</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>18362443</td>
<td>AT3G49530</td>
<td>Response to cold</td>
<td>Chawade et al. (2007)</td>
</tr>
<tr>
<td>5</td>
<td>644279</td>
<td>AT5G02830</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6950660</td>
<td>AT5G20570 (ATRIBX1)</td>
<td>Protein ubiquitination</td>
<td>Ascencio-Ibáñez et al. (2008)</td>
</tr>
<tr>
<td></td>
<td>695628</td>
<td>AT5G20580</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>695628</td>
<td>AT5G20590</td>
<td>Cell wall organization or biogenesis</td>
<td>Xin et al. (2007)</td>
</tr>
<tr>
<td></td>
<td>6967960</td>
<td>AT5G20610</td>
<td>Response to blue light</td>
<td>Dellasio et al. (2005)</td>
</tr>
<tr>
<td></td>
<td>697071</td>
<td>AT5G20620 (UBQUITIN 4)</td>
<td>Cellular protein modification process</td>
<td>Sun &amp; Callis (1997)</td>
</tr>
<tr>
<td></td>
<td>8506476</td>
<td>AT5G24770</td>
<td>Defense response</td>
<td>Catlin et al. (2015)</td>
</tr>
<tr>
<td></td>
<td>8773789</td>
<td>AT5G255280</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>8823283</td>
<td>AT5G25400</td>
<td>Carbohydrate transport</td>
<td>Wang et al. (2008)</td>
</tr>
<tr>
<td></td>
<td>10856791</td>
<td>AT5G28830</td>
<td>Unknown</td>
<td></td>
</tr>
<tr>
<td></td>
<td>26161831</td>
<td>AT5G65460 (KAC2)</td>
<td>Photosynthesis</td>
<td>He et al. (2005)</td>
</tr>
<tr>
<td></td>
<td>26176021</td>
<td>AT5G65480</td>
<td>Unknown</td>
<td>Wang et al. (2008)</td>
</tr>
<tr>
<td></td>
<td>26225832</td>
<td>AT5G65530 (GTE7)</td>
<td>Defense response</td>
<td>Wang et al. (2008)</td>
</tr>
</tbody>
</table>

Table 1. List of 33 candidate SNPs for European *A. thaliana* lines in the 10% top hits, based on a combination of the three methods.
Figure S1. **Observed false discovery rates.** The tests are based on (A) *snmf*, (B) *tess3*, (C) *pcadapt*, (D) $F_{ST}$. Sixteen data sets containing 10% of outlier loci were used in each panel.
Figure S2. Power values of snmf, tess3, pcdapt (Factor methods) and classical $F_{ST}$ tests with assignment and prior to admixture. All data sets contained 10% of outlier. Considering for a expected FDR of $\alpha = 0.1$: (A) Power values for the case $4Nm = 5$. The $F_{ST}$ test based on assignment of individuals to their most likely cluster failed to detect outlier loci. (B) Power values for the case $4Nm = 10$. (C) Power values for the case $4Nm = 15$. (D) Power values for the case $4Nm = 20$. 
Figure S3. Venn diagrams. Intersection between the lists of loci obtained for each method applied to the *A. thaliana* data set. The pcsadapt tests turned out to be more conservative.