

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

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

5 May 20, 2016

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

8
9 Running Title: Identifying outlier loci in admixed and in continuous populations

10
11 **Keywords:** Inference of Population Structure, Geographic Variation, Genome Scans for Selection,
12 Admixture.

13 Corresponding Author: Olivier François

14 Université Grenoble-Alpes,
15 TIMC-IMAG, UMR CNRS 5525,
16 Grenoble, 38042, France.

17 +334 56 52 00 25 (ph.)

18 +334 56 52 00 55 (fax)

19 `olivier.francois@imag.fr`

20

Abstract

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

33 1 Introduction

34 Natural selection, the process by which organisms that are best adapted to their environment
35 have an increased contribution of genetic variants to future generations, is the driving force of
36 evolution (Darwin, 1909). Identifying genomic regions that have been the targets of natural selec-
37 tion is one of the most important challenge in modern population genetics (Vitti *et al.*, 2013). To
38 this aim, examining the variation in allele frequencies between populations is a frequently applied
39 strategy (Cavalli-Sforza, 1966). More specifically, by sampling a large number of single nucleotide
40 polymorphisms (SNPs) throughout the genome, loci that have been affected by diversifying selec-
41 tion can be identified as outliers in the upper tail of the empirical distribution of F_{ST} (Lewontin &
42 Krakauer, 1973; Beaumont & Nichols, 1996; Akey *et al.*, 2002; Weir *et al.*, 2005). For selectively
43 neutral SNPs, F_{ST} is determined by genetic drift, which affects all SNPs across the genome in a
44 similar way. In contrast, natural selection has locus-specific effects that can cause deviations in
45 F_{ST} values at selected SNPs and at linked loci.

46 Outlier tests based on the empirical distribution of F_{ST} across the genome requires that the
47 sample is subdivided into K subsamples, each of them corresponding to a distinct genetic group.
48 For outlier tests, defining subpopulations may be a difficult task, especially when the background
49 levels of F_{ST} are weak and when populations are genetically homogeneous (Waples & Gaggiotti,
50 2006). For example, Europe is genetically homogeneous for human genomes, and it is characterized
51 by gradual variation in allele frequencies from the south to the north of the continent (Lao *et al.*,
52 2008), in which genetic proximity mimics geographic proximity (Novembre *et al.*, 2008). Studying
53 evolution in the field, most ecological studies use individual-based sampling along geographic
54 transects without using prior knowledge of populations (Manel *et al.*, 2003; Schoville *et al.*, 2012).
55 For example, the 1001 genomes project for the plant species *Arabidopsis thaliana* used a strategy
56 in which individual ecotypes were sampled with a large geographic coverage of the native and
57 naturalized ranges (Horton *et al.*, 2012; Weigel & Mott, 2009). One last difficulty with F_{ST} tests
58 arises from the presence of individuals with multiple ancestries (admixture), for which the genome
59 exhibits a mosaic of fragments originating from different ancestral populations (Long, 1991). The
60 admixture phenomenon is ubiquitous over sexually reproducing organisms (Pritchard *et al.*, 2000).

61 Admixture is pervasive in humans because migratory movements have brought together peoples
62 from different origins (Cavalli-Sforza *et al.*, 1994). Striking examples include the genetic history
63 of African American and Mestizo populations, for which the contributions of European, Native
64 American, and African populations had been studied extensively (Bryc *et al.*, 2010; Tang *et al.*,
65 2007).

66 Most of the concerns raised by definitions of subpopulations are commonly answered by the
67 application of clustering or ancestry estimation approaches such as **structure** or principal com-
68 ponent analysis (Pritchard *et al.*, 2000; Patterson *et al.*, 2006). These approaches rely on the
69 framework of factor models, where a factor matrix, the Q -matrix for **structure** and the score
70 matrix for PCA, is used to define individual ancestry coefficients, or to assign individuals their
71 most probable ancestral genetic group (Engelhardt & Stephens, 2010). To account for geographic
72 patterns of genetic variation produced by complex demographic histories, spatially explicit ver-
73 sions of the **structure** algorithm can include models for which individuals at nearby locations
74 tend to be more closely related than individuals from distant locations (François & Durand, 2010;
75 Wright, 1943).

76 In this study, we propose new tests to identify outlier loci in admixed and in continuous
77 populations by extending the definition of F_{ST} to this framework (Long, 1991). Our tests are
78 based on the computation of ancestry coefficient and ancestral allele frequency, Q and F , matrices
79 obtained from ancestry estimation programs. We develop a theory for the derivation of this new
80 F_{ST} statistic, defining it as the proportion of genetic diversity due to allele frequency differences
81 among populations in a model with admixed individuals (Holsinger & Weir, 2009). Then we
82 compute our new statistic using the outputs of two ancestry estimation programs: **snmf** which is
83 used as fast and accurate version of the **structure** algorithm, and **tess3** a fast ancestry estimation
84 program using genetic and geographic data (Frichot *et al.*, 2014; Caye *et al.*, 2016). Using simulated
85 data sets and SNPs from human and plants, we compared the results of genome scans obtained
86 with our new F_{ST} statistic with the results of PCA-based methods (Hao *et al.*, 2016; Duforet-
87 Frebourg *et al.*, 2016; Galinsky *et al.*, 2016; Luu *et al.*, in prep.).

88 2 F -statistics for populations with admixed individuals

89 In this section, we extend the definition of F_{ST} to populations containing admixed individuals,
90 and for which no subpopulations can be defined a priori. We consider SNP data for n individuals
91 genotyped at L loci. The data for each individual, i , and for each locus, ℓ , are recorded into a
92 genotypic matrix Y . The matrix entries, $y_{i\ell}$, correspond to the number of derived or reference
93 alleles at each locus. For diploid organisms, $y_{i\ell}$ is an integer value 0, 1 or 2. Assuming K predefined
94 subpopulations, the fixation index, F_{ST} , can be calculated according to S. Wright's definition as
95 follows (Wright, 1951)

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

96 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
97 in subpopulation k , and f is the allele frequency in the total population.

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

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

101 where σ_T^2 is the total variance and σ_S^2 is the residual variance (Holsinger & Weir, 2009).
102 This definition of F_{ST} , which uses a linear regression framework, can be extended to models
103 with admixed individuals. Suppose that a population contains admixed individuals, and the
104 source populations are unknown. Assume that the individual ancestry coefficients, Q , and the
105 ancestral population frequencies, F , are estimated from an ancestry estimation algorithm, such as
106 **structure** (Pritchard *et al.*, 2000). For diploid organisms, a genotype is the sum of two parental
107 gametes, taking the values 0 or 1. In an admixture model, the two gametes can be sampled
108 either from the same or from distinct ancestral populations. The admixture model assumes that
109 individuals mate at random at the moment of the admixture event. Let f_k be the allele frequency
110 in ancestral population k . Omitting the locus subscript ℓ , a statistical model for an admixed

111 genotype at a given locus can be written as follows

$$y = x_1 + x_2$$

112 where x_1 and x_2 are independent Bernoulli random variables modeling the parental gametes. The
113 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
114 integer value between 1 and K representing the hidden ancestry of each gamete).

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

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

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

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

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

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

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

123 and the formula for F_{ST} becomes

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

124 The above definition of F_{ST} for admixed populations is obviously related to the original def-
125 inition of Wright's fixation index, and it is also related to Long's formula which assumes known
126 ancestral populations (Long, 1991). The estimates of the sample sizes, n_k , can be obtained by
127 setting $n_k = nq_k$, and the sampled allele frequencies are replaced by their ancestral allele frequen-
128 cies. For haploid organisms, for which the genotype is coded 0 or 1, the definition of F_{ST} follows

129 the same developments. In this case, the definition of F_{ST} for a structured population corresponds
130 to the squared correlation, R^2 , in the regression of observed allele frequencies on subpopulation
131 indicators.

132 **Admixture estimates.** While many algorithms can compute the Q and F matrices, our ap-
133 plication of the above definition will focus on ancestry estimates obtained by nonnegative matrix
134 factorization algorithms (Frichot *et al.*, 2014). Frichot’s algorithm assumes that the sampled
135 genotype frequencies can be modelled by a mixture of ancestral genotype frequencies as follows

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

136 where $y_{i\ell}$ is the genotype of individual i at locus ℓ , the Q_{ik} are the ancestry coefficients for
137 individual i , the $G_{k\ell}(j)$ are the ancestral genotype frequencies, and p is the ploidy of the studied
138 organism (δ is the Kronecker delta symbol). For diploids ($p = 2$), the relationship between
139 ancestral allele and genotype frequencies can be written as follows

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

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

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

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

146 **Population differentiation tests.** The regression framework developed in the previous para-
147 graph leads to a direct approximation of the distribution of F_{ST} under the null-hypothesis of
148 random mating. We define the squared z -scores as follows

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

149 Then by classical arguments for regression models, we have

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

150 where $F(K - 1, n - K)$ is the Fisher distribution with $K - 1$ and $n - K$ degrees of freedom (Sokal
151 & Rohlf, 2012). In genome scans for selection, we assume that n is large enough to approximate
152 the distribution of squared z -scores as a chi-squared distribution with $K - 1$ degrees of freedom

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

153 To calibrate the null-hypothesis, we adopt an empirical null-hypothesis testing approach which
154 evaluates the level of population differentiation expected at selectively neutral SNPs (François
155 *et al.*, 2016). Following GWAS approaches, the test calibration is achieved after computing the
156 genomic inflation factor, defined by the median of the squared z -scores divided by the median
157 of a chi-squared distribution with $K - 1$ degrees of freedom (genomic control, Devlin & Roeder
158 (1999)).

159 3 Simulation experiments and data sets

160 **Simulated data sets.** We simulated of 10,000 unlinked SNPs for 200 individuals based on
161 Wright's two-island models. Each island corresponded to an ancestral population prior to admix-
162 ture. Two distinct scenarios were considered. In the first scenario, the proportion of loci under
163 selection was equal to 5% whereas this proportion was equal to 10% in the second scenario. To
164 mimic genetic variation at outlier loci, we used that a locus with reduced levels of gene flow is
165 expected to have an increased F_{ST} value (Bazin *et al.*, 2010; Caye *et al.*, 2016). Thus, we assumed
166 that adaptive SNPs had a migration rate smaller than the migration rate at selectively neutral
167 SNPs ($4Nm = 20, 15, 10, 5$ for neutral SNPs and $4Nm_s = 0.1, 0.25, 0.5, 1$ for adaptive SNPs). A
168 total number of 32 different data sets were generated by using the computer program `ms` (Hudson,
169 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),

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

170 For each individual, we assumed that each allele originated in the first ancestral population with
171 probability $p(x_i)$ and in the second ancestral population with probability $1 - p(x_i)$ (Durand *et al.*,
172 2009).

173 **Computer programs** We performed genome scans for selection using three methods: **snmf** (Fri-
174 chot *et al.*, 2014), **tess3** (Caye *et al.*, 2016), **pcadapt** (Luu *et al.*, in prep.; Duforet-Frebourg *et al.*,
175 2016). A fourth method used the standard F_{ST} statistic where subpopulations were obtained from
176 the assignment of individuals to their most likely genetic cluster. Like for **snmf**, the **tess3** esti-
177 mates of the Q and G matrices are based on matrix factorization techniques. The main difference
178 between the two programs is that **tess3** computes ancestry estimates by incorporating informa-
179 tion on individual geographic coordinates in its algorithm whereas the **snmf** algorithm is closer to
180 **structure** (Caye *et al.*, 2016). For **snmf** and for **tess3**, we used $K = 2$ ancestral populations.
181 This value of K corresponded to the minimum of the cross-entropy criterion when K was varied
182 in the range 1 to 6. The default values of the two programs were implemented for all their internal
183 parameters. Each run of the two programs was replicated five times, and the run with the lowest
184 cross-entropy value was selected for computing the F_{ST} statistics according to formula (1). We
185 compared the results of **snmf** and **tess3** with the program **pcadapt** (Luu *et al.*, in prep.). The test
186 statistic of this new version of **pcadapt** is the Manhanalobis distance relative to the z -scores ob-
187 tained after regressing the SNP frequencies on the $K - 1$ principal components. We used **pcadapt**
188 with the first principal component. As for **snmf** and for **tess3**, test calibration in **pcadapt** was

189 based on the computation of the genomic inflation factor (genomic control). For genome scans
190 based on the F_{ST} statistic where subpopulations are obtained from the assignment of individuals
191 to their most likely genetic cluster, we used 2 clusters and a chi-squared distribution with one
192 degree of freedom after recalibration of the null-hypothesis. Before applying the methods to the
193 simulated data sets, the SNPs were filtered out and only the loci with minor allele frequency
194 greater than 5% were retained for the analysis.

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

199 **Real data sets.** To provide an application of our method to natural populations, we reanalyzed
200 data from the model plant organism *Arabidopsis thaliana*. We analyzed genomic data from 120
201 European lines of *A. thaliana* genotyped for 216k SNPs, with a density of one SNP per 500
202 bp (Atwell *et al.*, 2010). This annual plant is native to Europe and central Asia, and within
203 its native range, it goes through numerous climatic conditions and selective pressures (Mitchell-
204 Olds & Schmitt, 2006). Ecotypes from Northern Scandinavia were not included in the data (14
205 ecotypes representing a divergent genetic cluster in the original data set). We applied genome
206 scans for selection using `snmf`, `tess3` and `pcadapt` with $K = 2$ ancestral populations and one
207 principal component. In addition, we analyzed human genetic data for 1,385 European individuals
208 genotyped at 447k SNPs (Nelson *et al.*, 2008). We applied genome scans for selection using `snmf`
209 and `pcadapt` with $K = 2$ ancestral populations and one principal component.

210 4 Results

211 **Simulations of admixed individuals.** We evaluated the performances of genome scans using
212 tests based on `snmf`, `tess3`, `pcadapt`, and F_{ST} , in the presence of admixed individuals. Considering
213 q -values thresholds between 0.01 and 0.2, we computed observed FDR values for the lists of outlier
214 loci produced by each test. The observed FDR values remained generally below their expected

215 values (Figure 1 for data sets with 5% of outliers, Figure S1 for data sets with 10% of outliers).
216 These observations confirmed that the use of genomic inflation factors leads to overly conservative
217 tests (François *et al.*, 2016). Since similar levels of observed FDR values were observed across the
218 4 tests, we did not implement other calibration methods than genomic control.

219 Next, we evaluated the sensitivity (power) of the 4 tests in each simulation scenario. As we
220 expected from the simulation process, the tests had higher power when the relative levels of the
221 selection intensity were higher. For $4Nm = 5$ and $4Nm_s = .1, .25, .5,$ and 1, the power of the
222 tests for **snmf**, **tess3**, **pcadapt** was close to 27% (expected FDR equal to $\alpha = 0.1$, Figure 2A
223 for data sets with 5% of outliers, Figure S2A for data sets with 10% of outliers). The F_{ST} test
224 based on assignment of individuals to their most likely cluster failed to detect outlier loci (power
225 value equal to 0%). For $4Nm = 10$, the power of the tests was in the range 40% - 45% for **snmf**,
226 **tess3**, **pcadapt** and equal to 26% for the F_{ST} test (Figure 2B (5% of outlier loci), Figure 2B
227 (10% of outlier loci)). For $4mN \geq 15$, corresponding to the highest selection rates, the power was
228 approximately equal to 50% for all methods considered (Figure 2C and D ((5% of outlier loci),
229 Figure 2C and D (10% of outlier loci)). The relatively low power values confirmed that the tests
230 were conservative, and non-neutral loci were difficult to detect. To provide an upper bound on
231 the power of an outlier test in the context of admixed populations, we applied an F_{ST} test to
232 the samples obtained prior to admixture, estimating allele frequencies from their true ancestral
233 populations (Figure 2). For $4Nm = 5$ and 10, the power of the tests for **snmf**, **tess3**, **pcadapt** was
234 similar to the power obtained when we applied outlier tests to the data before admixture. This
235 experiment confirmed that the use of approaches that estimate ancestry coefficients is appropriate
236 when no subpopulation can be predefined.

237 **Biological data analysis**

238 **Arabidopsis data.** We applied **snmf**, **tess3** and **pcadapt** to perform genome scans for selection
239 in 120 European lines of *Arabidopsis thaliana* (216k SNPs). Each ecotype was collected from a
240 unique geographic location, and there were no predefined populations. To study adaptation at the
241 continental scale, ecotypes from northern Scandinavia, which were grouped together by clustering

242 programs, were removed from the original data set of Atwell *et al.* (2010). For **snmf** and **tess3**,
243 the cross-entropy criterion indicated that there are 2 main clusters in Europe, and that finer
244 substructure could be detected as a result of historical isolation-by-distance processes. For $K = 2$,
245 the western cluster grouped all lines from the British Isles, France and Iberia and the eastern
246 cluster grouped all lines from Central, Eastern Europe and Southern Sweden. For implementing
247 genome scans for selection, we used 2 clusters in **snmf** and **tess3**, and one principal component
248 in **pcadapt**. The genomic inflation factor was equal to $\lambda = 11.5$ for the test based on **snmf**, and
249 it was equal to $\lambda = 13.1$ for the test based on **tess3**. The interpretation of these two values is
250 that the background level of population differentiation that was used in the **snmf** and **tess3** tests
251 is around 0.09 (see François *et al.* 2016). For the three methods, the Manhattan plots exhibited
252 peaks at the same chromosome positions (Figure 3). For an expected FDR level equal to 1%,
253 the Storey and Tibshirani algorithm resulted in a list of 572 chromosome positions for the **snmf**
254 tests, 882 for the **tess3** tests (Figure 3). The test based on PCA was more conservative. The
255 difference between the tests could be attributed to the estimation of the genomic inflation factor
256 which differs for PCA methods (see Venn diagrams in Figure S3).

257 Table 1 reports a list of 33 candidate SNPs for European *A. thaliana* lines in the 10% top hits,
258 based on the peaks detected by the three methods. Figure 4 displays a Manhattan plot for the plant
259 genome showing the main outlier loci detected by our genome scans for selection. For chromosome
260 1, the list contains SNPs in the gene AT1G80680 involved in resistance against bacterial pathogens.
261 For chromosome 2, the list contains SNPs in the gene AT2G18440 (*AtGUT15*), which can be used
262 by plants as a sensor to interrelated temperatures, and which has a role for controlling growth
263 and development in response to a shifting environment (Lu *et al.*, 2005). For chromosome 3, the
264 list contains SNPs in the gene AT3G11920 involved in cell redox homeostasis. Fine control of
265 cellular redox homeostasis is important for integrated regulation of plant defense and acclimatory
266 responses (Mühlenbock *et al.*, 2007). For chromosome 4, we found SNPs in the gene AT4G31180
267 (*IBI1*) involved in defense response to fungi. The most important list of candidate SNPs was
268 found in the fifth chromosome. For example, the list of outlier SNPs contained SNPs in the gene
269 AT5G02820, involved in endoreduplication, that might contribute to the adaptation to adverse

270 environmental factors, allowing the maintenance of growth under stress conditions (Chevalier
271 *et al.*, 2011), in the genes AT5G18620, AT5G18630 and AT5G20620 (UBIQUITIN 4) involved in
272 response to temperature stress (Kim & Kang, 2005), and in the gene AT5G20610 which is involved
273 in response to blue light (DeBlasio *et al.*, 2005).

274 **Human data.** We applied the `snmf` and `pcadapt` tests to 1,385 European individuals from the
275 POPRES data set (447k SNPs in 22 chromosomes). We used $K = 2$ ancestral populations in
276 `snmf` and one principal component for PCA. For `snmf`, the genomic inflation factor was equal to
277 $\lambda = 9.0$, indicating a background level of population differentiation around 0.006 between northern
278 and southern European populations. For an expected FDR equal to 10%, we found 205 outlier
279 loci using `snmf` tests, and 165 outlier loci with `pcadapt` (Figure 5). For chromosome 2, the most
280 important signal of selection was found at the lactase persistence gene (*LCT*) (Bersaglieri *et al.*,
281 2004). For chromosome 4, 5 SNPs were found at the *ADH1C* locus that is involved in alcohol
282 metabolism (Han *et al.*, 2007), close to the *ADH1B* locus reported by Galinsky *et al.* (2016). For
283 chromosome 6, a signal of selection corresponding to the human leukocyte antigen (*HLA*) region
284 was identified. For chromosome 15, there was an outlier SNP in the *HERC2* gene, which modulates
285 human pigmentation (Visser *et al.*, 2012) (Figure 6).

286 5 Discussion

287 When no subpopulation can be defined a priori, analysis of population structure commonly
288 relies on the computation of the Q (and F) ancestry matrix obtained through the application of
289 the program `structure` or one of its improved versions (Pritchard *et al.*, 2000; Tang *et al.*, 2005;
290 Chen *et al.*, 2007; Alexander *et al.*, 2009; Raj *et al.*, 2014; Frichot *et al.*, 2014; Caye *et al.*, 2016).
291 In this context, we proposed a definition of F_{ST} based on the Q and F matrices, and we used
292 this new statistic to screen genomes for signatures of diversifying selection. By modelling admixed
293 genotypes, our definition of F_{ST} was inspired by an analysis of variance approach for the genotypic
294 data (Weir & Cockerham, 1984; Holsinger & Weir, 2009).

295 The estimator for F_{ST} presented here is related to the estimator proposed by Long (1991) for
296 population data. Long's estimator is obtained from the variance of allele frequencies with respect

297 to their expectations based on an admixture model, that enable estimating the effect of genetic drift
298 and the effective size of the hybrid population. In order to obtain Long's estimate, multiple locus
299 samples are required from the hybrid population and from all contributing parental populations.
300 For the method proposed in our manuscript, information on ancestral genetic diversity is evaluated
301 with less prior assumptions by the application of ancestry estimation programs.

302 Assuming that a large number of SNPs are genotyped across multiple populations, the calibra-
303 tion of statistical tests of neutrality did not require assumptions about population demographic
304 history. Our simulations of admixed populations provided evidence that the tests based on this
305 new statistic had an increased power compared to tests where we assigned individuals to their
306 most probable cluster. Interestingly, the power of those tests was only slightly lower than stan-
307 dard F_{ST} tests based on the truly ancestral allele frequencies. A comparison of our results for
308 Europeans from the POPRES data sets and the genome-wide patterns of selection in 230 ancient
309 Eurasians provides additional evidence that the signals detected by our F_{ST} were already present
310 in the populations that were ancestral to modern Europeans (Mathieson *et al.*, 2015).

311 Our reanalysis of European *A. thaliana* data provided a clear example of the usefulness of our
312 F_{ST} statistic to detect targets of natural selection in plants. European ecotypes of *Arabidopsis*
313 *thaliana* are continuously distributed across the continent, with population structure influenced by
314 historical isolation-by-distance processes (Atwell *et al.*, 2010; Hancock *et al.*, 2011; François *et al.*,
315 2008). The application of our F_{ST} statistic to the SNP data suggested several new candidate
316 loci involved in resistance against pathogens, in growth and development in response to a shifting
317 environment, in the regulation of plant defense and acclimatory responses, in the adaptation to
318 adverse environmental factors, in allowing the maintenance of growth under stress conditions, in
319 response to temperature stress or response to light.

320 An alternative approach to investigating population structure without predefined populations is
321 by using principal component analysis (Patterson *et al.*, 2006). Statistics extending the definition
322 of F_{ST} were also proposed for PCA (Hao *et al.*, 2016; Duforet-Frebourg *et al.*, 2016; Galinsky
323 *et al.*, 2016). The performances of PCA statistics and our new F_{ST} statistic were highly similar.
324 The small differences observed for the two tests could be ascribed to the estimation of inflation

325 factors to calibrate the null-hypothesis. The idea of detecting signatures of selection in an admixed
326 population has a considerable history and has been explored since the early seventies (Blumberg
327 & Hesser, 1971; Adams & Ward, 1973; Tang *et al.*, 2007). The connection between our definition
328 of F_{ST} and previous works shows that the methods studied in this study, including PCA or
329 ancestry programs, are extensions of classical methods of detection of selection using admixed
330 populations (Long, 1991). Our results allow us to hypothesize that the age of selection detected
331 by PCA and by the new methods proposed is similar. Thus it is likely that the selective sweeps
332 detected by PCA and F_{ST} methods correspond to ancient selective sweeps already differentiating
333 in ancestral populations (Mathieson *et al.*, 2015).

334 While only minor differences between our results of genome scans with 4 methods were ob-
335 served, the results might be still sensitive to the algorithm used to estimating the ancestry matrices.
336 Wollstein & Lao (2015) performed an extensive comparison of 3 recently proposed ancestry esti-
337 mation methods, `admixture`, `faststructure`, `snmf` (Alexander & Lange, 2011; Raj *et al.*, 2014;
338 Frichot *et al.*, 2014), and they concluded that the accuracy of the methods could differ in some
339 simulation scenarios. In practice, it would be wise to apply several methods and to combine their
340 results by using a meta-analysis approach as demonstrated in François *et al.* (2016).

341 References

- 342 Adams J, Ward RH (1973). Admixture studies and the detection of selection. *Science* 180, 1137–
343 1143.
- 344 Akey JM, Zhang G, Zhang K, Jin L, Shriver MD (2002). Interrogating a high-density SNP map
345 for signatures of natural selection. *Genome Research* 12, 1805–1814.
- 346 Alexander DH, Lange K (2011.) Enhancements to the ADMIXTURE algorithm for individual
347 ancestry estimation. *BMC Bioinformatics* 12, 246.
- 348 Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated
349 individuals. *Genome Research* 19, 1655–1664.
- 350 Ascencio-Ibáñez JT, Sozzani R, Lee TJ, *et al.* (2008). Global analysis of *Arabidopsis* gene expres-

- 351 sion uncovers a complex array of changes impacting pathogen response and cell cycle during
352 geminivirus infection. *Plant Physiology* 148, 436–454.
- 353 Atwell S, Huang YS, Vilhjálmsson BJ, *et al.* (2010). Genome-wide association study of 107 phe-
354 notypes in *Arabidopsis thaliana* inbred lines. *Nature* 465, 627–631.
- 355 Barton NH, Hewitt GM (1985). Analysis of hybrid zones. *Annual review of Ecology and Systematics*
356 16, 113–148.
- 357 Bazin E, Dawson KJ, Beaumont MA (2010). Likelihood-free inference of population structure and
358 local adaptation in a Bayesian hierarchical model. *Genetics* 185, 587–602.
- 359 Beaumont MA, Nichols RA (1996). Evaluating loci for use in the genetic analysis of population
360 structure. *Proceedings of the Royal Society of London B: Biological Sciences* 263, 1619–1626.
- 361 Bersaglieri T, Sabeti PC, Patterson N, *et al.* (2004). Genetic signatures of strong recent positive
362 selection at the lactase gene. *The American Journal of Human Genetics* 74, 1111–1120.
- 363 Blumberg BS, Hesser JE (1971). Loci differentially affected by selection in two american black
364 populations. *Proceedings of the National Academy of Sciences* 68, 2554–2558.
- 365 Bryc K, Auton A, Nelson MR, *et al.* (2010). Genome-wide patterns of population structure and
366 admixture in west africans and african americans. *Proceedings of the National Academy of*
367 *Sciences* 107, 786–791.
- 368 Catinot J, Huang JB, Huang PY, *et al.* (2015). ETHYLENE RESPONSE FACTOR 96 positively
369 regulates *Arabidopsis* resistance to necrotrophic pathogens by direct binding to GCC elements
370 of jasmonate-and ethylene-responsive defence genes. *Plant, Cell & Environment* 38, 2721–2734.
- 371 Cavalli-Sforza LL, Menozzi P, Piazza A. *The History and Geography of Human Genes*. Princeton
372 University Press, Princeton, USA, 1994.
- 373 Cavalli-Sforza LL (1966). Population structure and human evolution. *Proceedings of the Royal*
374 *Society of London B: Biological Sciences* 164, 362–379.

- 375 Caye K, Deist TM, Martins H, Michel O, François O (2016). TESS3: Fast inference of spatial
376 population structure and genome scans for selection. *Molecular Ecology Resources* 16, 540–548.
- 377 Chawade A, Bräutigam M, Lindlöf A, Olsson O, Olsson B (2007). Putative cold acclimation
378 pathways in *Arabidopsis thaliana* identified by a combined analysis of mRNA co-expression
379 patterns, promoter motifs and transcription factors. *BMC Genomics* 8, 1.
- 380 Chen C, Durand E, Forbes F, François O (2007). Bayesian clustering algorithms ascertaining
381 spatial population structure: A new computer program and a comparison study. *Molecular*
382 *Ecology Notes* 7, 747–756.
- 383 Chen H, Kim HU, Weng H (2011). Malonyl-coA synthetase, encoded by ACYL ACTIVATING
384 ENZYME13, is essential for growth and development of *Arabidopsis*. *The Plant Cell* 23(6),
385 2247–2262.
- 386 Chevalier C, Nafati M, Mathieu-Rivet E, *et al.* (2011). Elucidating the functional role of endoredu-
387 plication in tomato fruit development. *Annals of Botany* 107(7), 1159–1169.
- 388 Darwin C. *On The Origin of Species by Means of Natural Selection* John Murray, Londonn, UK,
389 1909.
- 390 DeBlasio SL, Luesse DL, Hangarter RP (2005). A plant-specific protein essential for blue-light-
391 induced chloroplast movements. *Plant Physiology* 139, 101–114.
- 392 Devlin B, Roeder K (1999). Genomic control for association studies. *Biometrics* 55, 997–1004.
- 393 Duforet-Frebourg N, Luu K, Laval G, Bazin E, Blum MGB (2016). Detecting genomic signatures
394 of natural selection with principal component analysis: Application to the 1000 Genomes data.
395 *Molecular Biology and Evolution* 33(4), 1082–1093.
- 396 Durand E, Jay F, Gaggiotti OE, François O (2009). Spatial inference of admixture proportions
397 and secondary contact zones. *Molecular Biology and Evolution* 26(9), 1963–1973.
- 398 Engelhardt BE, Stephens M (2010). Analysis of population structure: A unifying framework and
399 novel methods based on sparse factor analysis. *PLoS Genetics* 6(9), e1001117. doi: 10.1371/jour-
400 nal.pgen.1001117.

- 401 François O, Blum MGB, Jakobsson M, Rosenberg NA (2008). Demographic history of Euro-
402 pean populations of *Arabidopsis thaliana*. *PLoS Genetics* 4(5), e1000075. doi: 10.1371/jour-
403 nal.pgen.1000075.
- 404 François O, Durand E (2010). Spatially explicit Bayesian clustering models in population genetics.
405 *Molecular Ecology Resources* 10, 773–784.
- 406 François O, Martins H, Caye K, Schoville SD (2016). Controlling false discoveries in genome scans
407 for selection. *Molecular Ecology* 25, 454–469.
- 408 Frichot E, Mathieu F, Trouillon T, Bouchard G, François O (2014). Fast and efficient estimation
409 of individual ancestry coefficients. *Genetics* 196, 973–983.
- 410 Galinsky KJ, Bhatia G, Loh PR, *et al.* (2016). Fast principal component analysis reveals convergent
411 evolution of ADH1B in Europe and East Asia *The American Journal of Human Genetics* 98(3),
412 456–472.
- 413 Guo KM, Babourina O, Christopher DA, Borsics T, Rengel Z (2008). The cyclic nucleotide-gated
414 channel, AtCNGC10, influences salt tolerance in *Arabidopsis*. *Physiologia Plantarum* 134, 499–
415 507.
- 416 Han Y, Gu S, Oota H, *et al.* (2007) Evidence of positive selection on a class I ADH locus. *The*
417 *American Journal of Human Genetics* 80(3), 441–456.
- 418 Hancock AM, Brachi B, Faure N, *et al.* (2011) Adaptation to climate across the *Arabidopsis*
419 *thaliana* genome. *Science* 334, 83–86.
- 420 Hao W, Song M, Storey JD (2016). Probabilistic models of genetic variation in structured popu-
421 lations applied to global human studies. *Bioinformatics* 32(5), 713–721.
- 422 He XJ, Mu RL, Cao WH, Zhang ZG, Zhang JS, Chen SY (2005). AtNAC2, a transcription factor
423 downstream of ethylene and auxin signaling pathways, is involved in salt stress response and
424 lateral root development. *The Plant Journal* 44, 903–916.
- 425 Holsinger KE, Weir BS (2009). Genetics in geographically structured populations: Defining, esti-
426 mating and interpreting F_{ST} . *Nature Reviews Genetics* 10, 639–650.

- 427 Horton MW, Hancock AM, Huang YS, *et al.* (2012). Genome-wide patterns of genetic variation
428 in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics* 44,
429 212–216.
- 430 Hudson RR (2002). Generating samples under a Wright–Fisher neutral model of genetic variation.
431 *Bioinformatics* 18(2), 337–338.
- 432 Kim YO, Kang H (2005). Cold-inducible zinc finger-containing glycine-rich RNA-binding protein
433 contributes to the enhancement of freezing tolerance in *Arabidopsis thaliana*. *The Plant Journal*
434 42, 890–900.
- 435 Lao O, Lu TT, Nothnagel M, *et al.* (2008). Correlation between genetic and geographic structure
436 in europe. *Current Biology* 18(16), 1241–1248.
- 437 Lewontin R, Krakauer J (1973). Distribution of gene frequency as a test of the theory of the
438 selective neutrality of polymorphisms. *Genetics* 74, 175–195.
- 439 Long JC (1991). The genetic structure of admixed populations. *Genetics* 127(2), 417–428.
- 440 Lu Y, Zhu J, Liu P (2005). A two-step strategy for detecting differential gene expression of cDNA
441 microarray data. *Current Genetics* 47(2), 121–131.
- 442 Luu K, Bazin E, Blum MGB (in prep.) pcadapt: An R package for performing genome scans for
443 selection based on principal component analysis. In preparation.
- 444 Manel S, Schwartz MK, Luikart G, Taberlet P (2003). Landscape genetics: Combining landscape
445 ecology and population genetics. *Trends in Ecology & Evolution* 18, 189–197.
- 446 Mathieson I, Lazaridis I, Rohland N, *et al.* (2015). Genome-wide patterns of selection in 230
447 ancient Eurasians. *Nature* 528, 499–503.
- 448 Mitchell-Olds T, Schmitt J (2006). Genetic mechanisms and evolutionary significance of natural
449 variation in arabidopsis. *Nature* 441, 947–952.
- 450 Mühlenbock P, Karpinska B, Karpinski S (2007). Oxidative stress and redox signalling in plants.
451 *eLS*. doi: 10.1002/9780470015902.a0020135.

- 452 Nelson MR, Bryc K, King KS, *et al.* (2008). The population reference sample, POPRES: A resource
453 for population, disease, and pharmacological genetics research. *The American Journal of Human*
454 *Genetics* 83(3), 347–358.
- 455 Novembre J, Johnson T, Bryc K, *et al.* (2008). Genes mirror geography within Europe. *Nature*
456 456, 98–101.
- 457 Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genetics*
458 2(12), e190.
- 459 Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus
460 genotype data. *Genetics* 155(2), 945–959.
- 461 Radin I, Mansilla N, Rödel G, Steinebrunner I (2015). The *Arabidopsis* COX11 homolog
462 is essential for cytochrome *c* oxidase activity. *Frontiers in Plant Science* 6, 1091.
463 doi:10.3389/fpls.2015.01091.
- 464 Raj A, Stephens M, Pritchard JK (2014). fastSTRUCTURE: Variational inference of population
465 structure in large SNP data sets. *Genetics* 197, 573–589.
- 466 Rajjou L, Belghazi M, Huguet R, *et al.* (2006). Proteomic investigation of the effect of salicylic
467 acid on *Arabidopsis* seed germination and establishment of early defense mechanisms. *Plant*
468 *Physiology* 141(3), 910–923.
- 469 Roth C, Wiermer M (2012). Nucleoporins Nup160 and Seh1 are required for disease resistance in
470 *Arabidopsis*. *Plant Signaling & Behavior* 7(10), 1212–1214.
- 471 Schoville SD, Bonin A, François O, *et al.* (2012). Adaptive genetic variation on the landscape:
472 Methods and cases. *Annual Review of Ecology, Evolution, and Systematics* 43, 23–43.
- 473 Sokal R, Rohlf F. *Biometry: The Principles and Practice of Statistics in Biological Research* (4th
474 edn). W.H. Freeman & Company, New York, NY, 2012.
- 475 Storey JD, Tibshirani R (2003). Statistical significance for genomewide studies. *Proceedings of the*
476 *National Academy of Sciences* 100(16), 9440–9445.

- 477 Sun CW, Callis J (1997). Independent modulation of *Arabidopsis thaliana* polyubiquitin mRNAs
478 in different organs and in response to environmental changes. *The Plant Journal* 11, 1017–1027.
- 479 Tang H, Choudhry S, Mei R, *et al.* (2007). Recent genetic selection in the ancestral admixture of
480 puerto ricans. *The American Journal of Human Genetics* 81(3), 626–633.
- 481 Tang H, Peng J, Wang P, Risch NJ (2005). Estimation of individual admixture: Analytical and
482 study design considerations. *Genetic Epidemiology* 28, 289–301.
- 483 Visser M, Kayser M, Palstra RJ (2012). HERC2 rs12913832 modulates human pigmentation by
484 attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter.
485 *Genome Research* 22(3), 446–455.
- 486 Vitti JJ, Grossman SR, Sabeti PC (2013). Detecting natural selection in genomic data. *Annual*
487 *Review of Genetics* 47, 97–120.
- 488 Wang Y, Zhang WZ, Song LF, *et al.* (2008). Transcriptome analyses show changes in gene expres-
489 sion to accompany pollen germination and tube growth in *Arabidopsis*. *Plant Physiology* 148,
490 1201–1211.
- 491 Waples RS, Gaggiotti O (2006). What is a population? an empirical evaluation of some genetic
492 methods for identifying the number of gene pools and their degree of connectivity. *Molecular*
493 *Ecology* 15, 1419–1439.
- 494 Weigel D, Mott R (2009). The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biology*
495 10(5), 1–5.
- 496 Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005). Measures of human population
497 structure show heterogeneity among genomic regions. *Genome Research* 15(11), 1468–1476.
- 498 Weir BS, Cockerham CC (1984). Estimating F -statistics for the analysis of population structure.
499 *Evolution* 38(6), 1358–1370.
- 500 Wollstein A, Lao O (2015). Detecting individual ancestry in the human genome. *Investigative*
501 *Genetics* 6, 1–12.

502 Wright S (1943). Isolation by distance. *Genetics* 28, 114.

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

504 Xin Z, Mandaokar A, Chen J, Last RL, Browse J (2007). Arabidopsis ESK1 encodes a novel
505 regulator of freezing tolerance. *The Plant Journal* 49, 786–799.

506 6 Figures and Tables

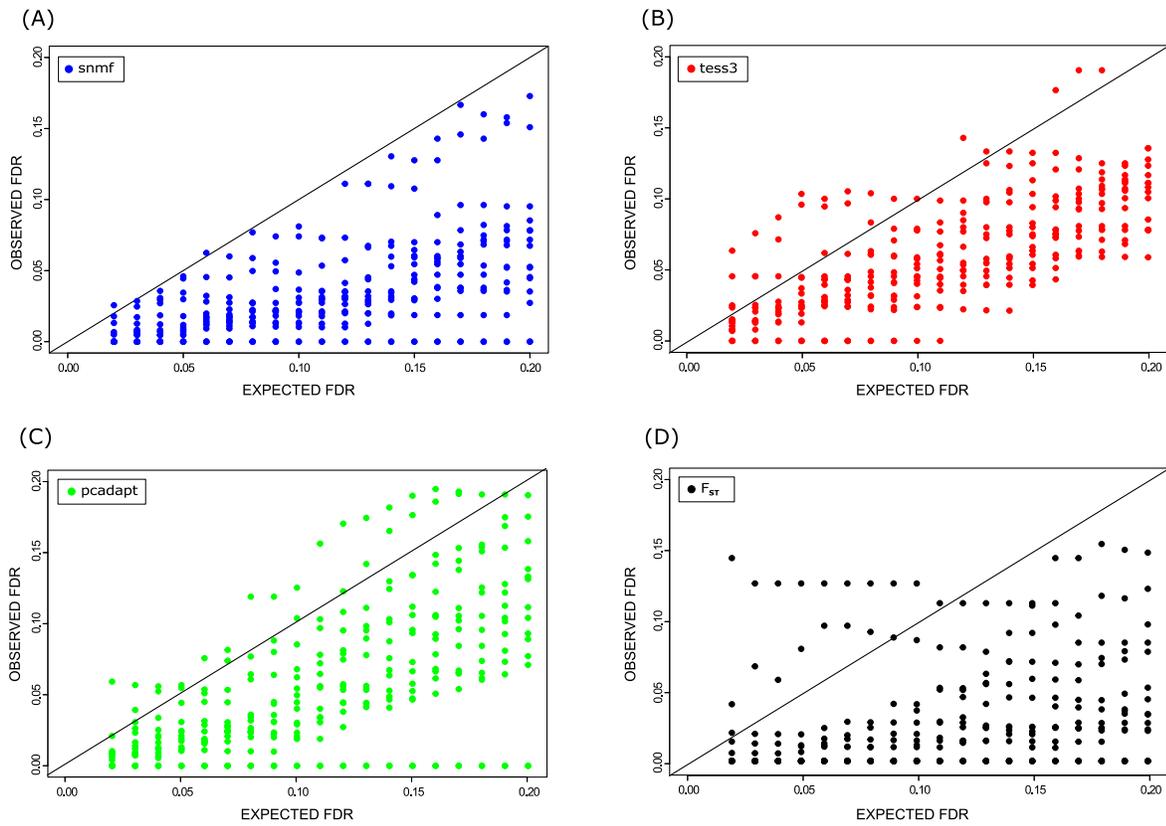


Figure 1. Observed false discovery rates. The tests are based on (A) *snmf*, (B) *tess3*, (C) *pcadapt*, (D) F_{ST} . Sixteen data sets containing 5% of outlier loci were used in each panel.

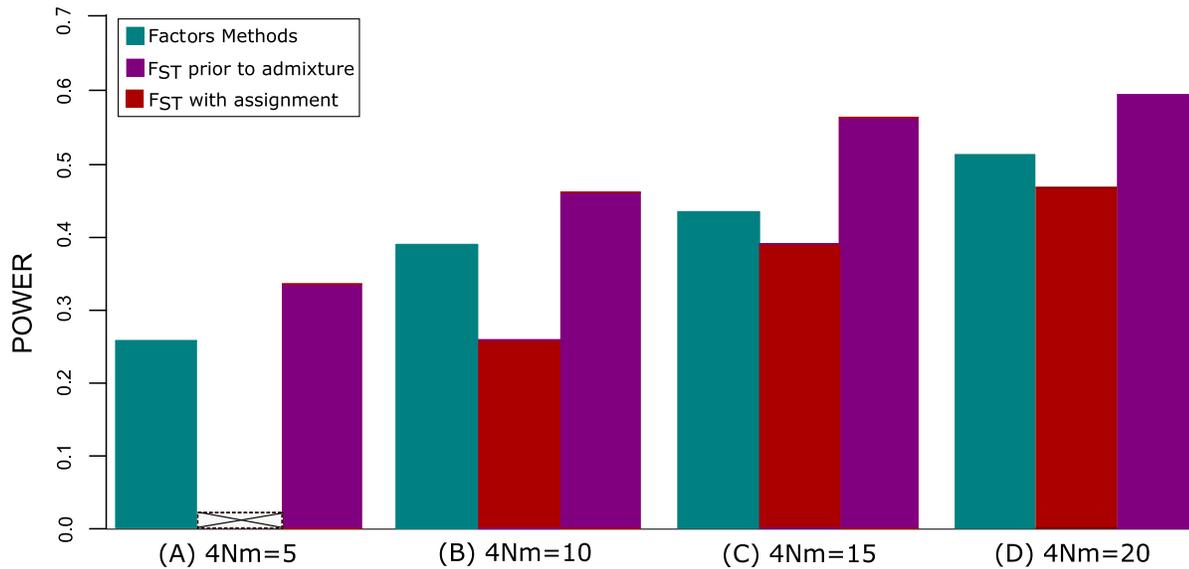


Figure 2. Power values of snmf, tess3, pcadapt (Factor methods) and classical F_{ST} tests with assignment and prior to admixture. All data sets contained 5% of outlier loci. Considering an expected FDR of $\alpha = 0.1$: (A) Power values for the case $4Nm = 5$. The F_{ST} test based on assignment of 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 $4Nm = 20$.

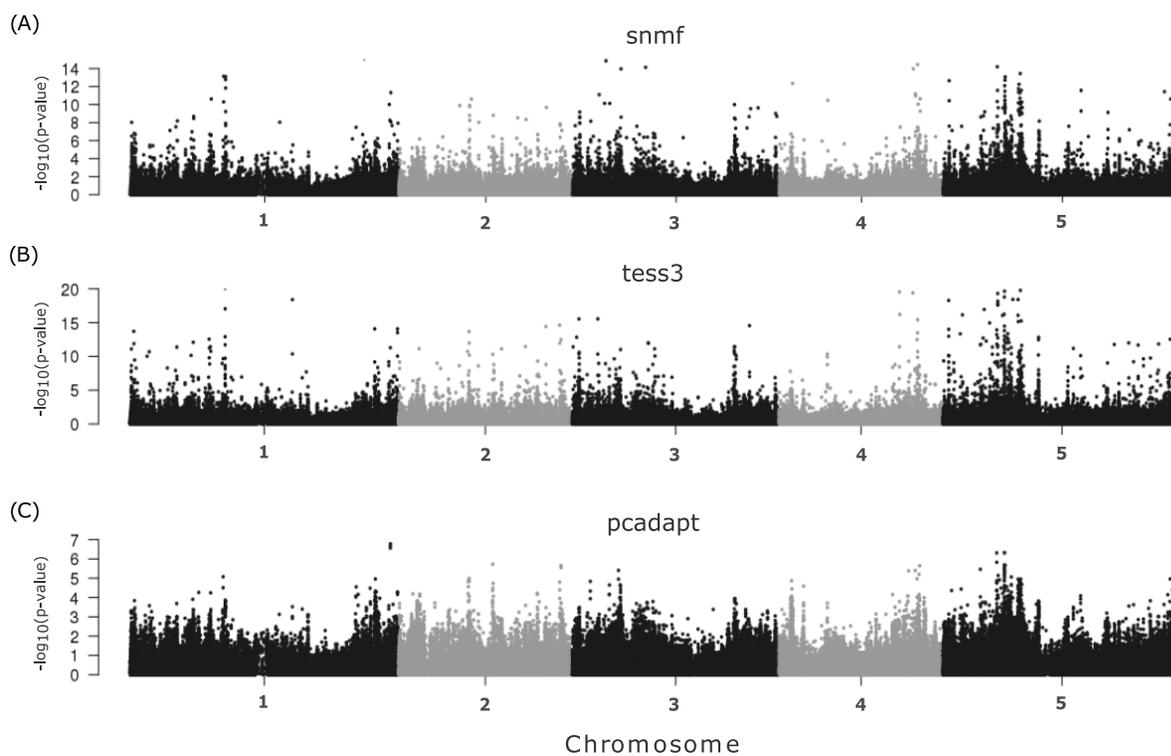


Figure 3. Manhattan plots of minus $\log_{10}(\text{p-values})$ for the *A. thaliana*. Considering the tests using: (A) snmf, (B) tess3 and (C) pcadapt.

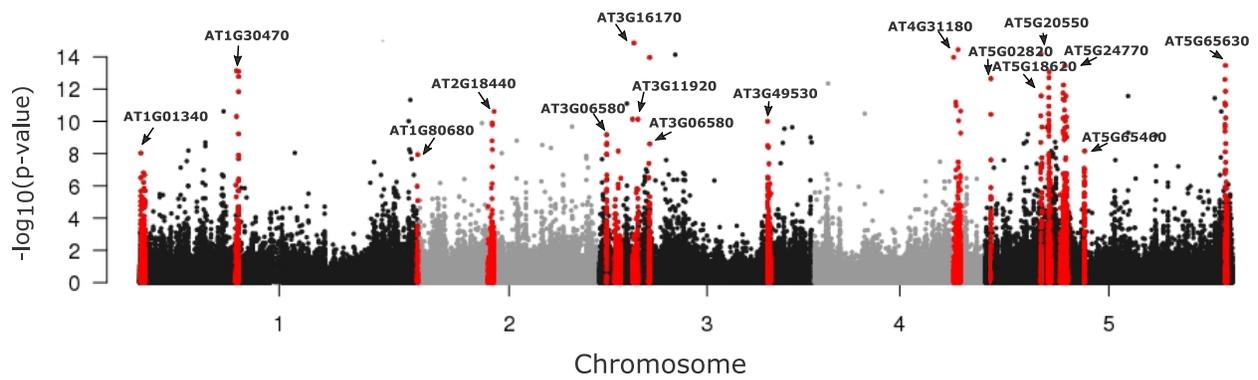


Figure 4. Manhattan plot of minus $\log_{10}(p\text{-values})$ for the *A. thaliana*. Candidate loci detected by genome scans for selection are colored in red (expected FDR level of 1%).

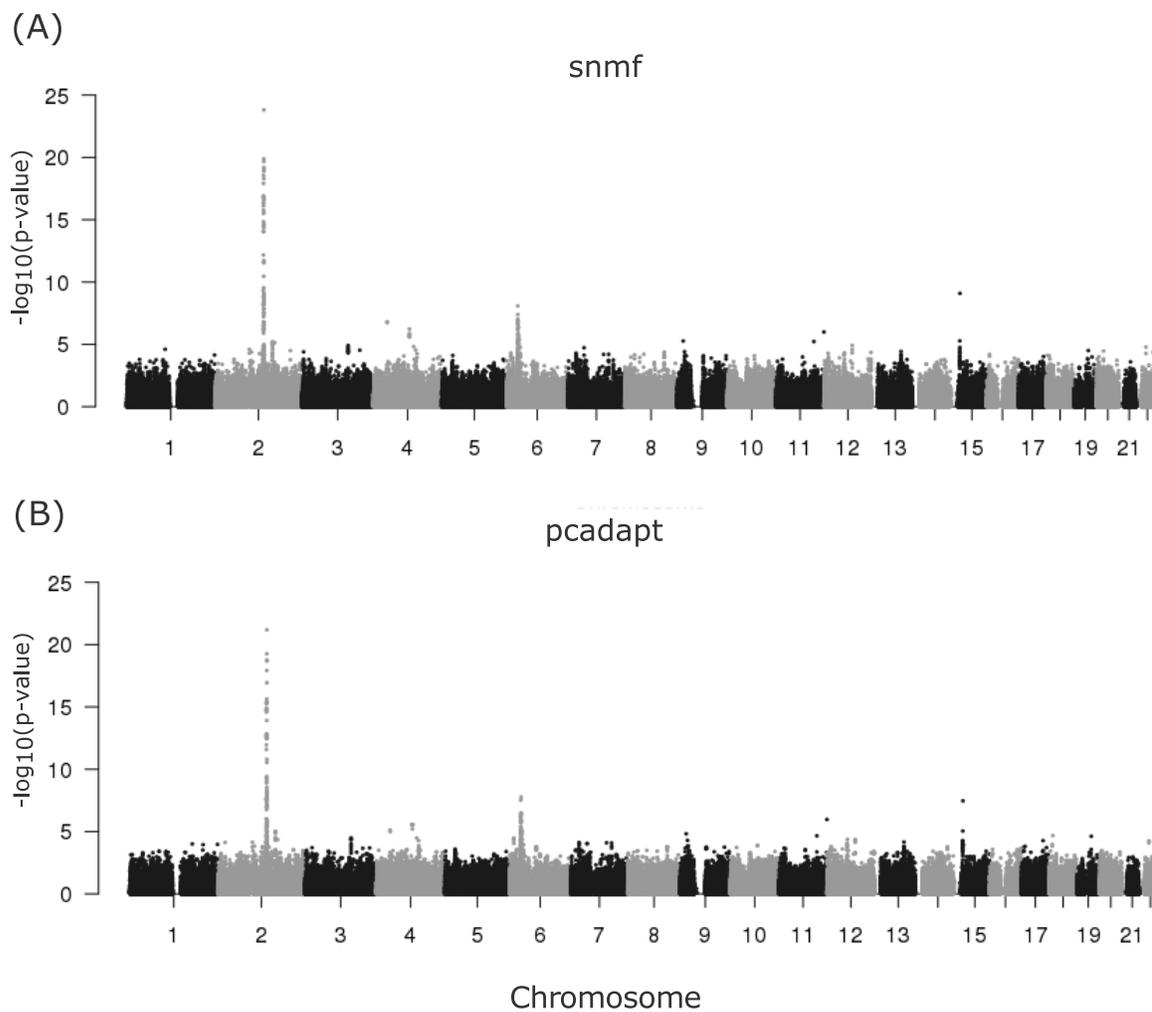


Figure 5. Manhattan plots of minus $\log_{10}(\text{p-values})$ for the 22 chromosomes of the POPRES data set. Considering the tests using: (A) snmf and (B) pcadapt.

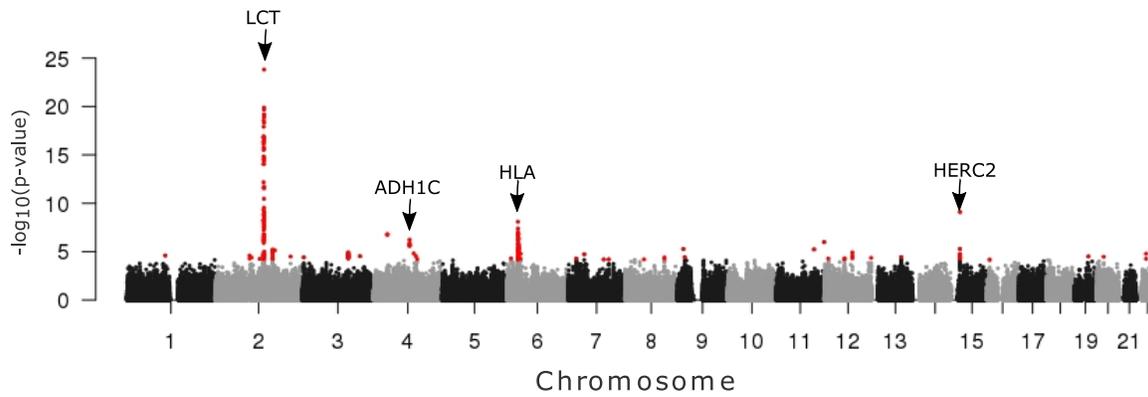


Figure 6. Manhattan plot of minus $\log_{10}(p\text{-values})$ for the POPRES data. Candidate loci detected by genome scans for selection are colored in red (expected FDR level of 10%)

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

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

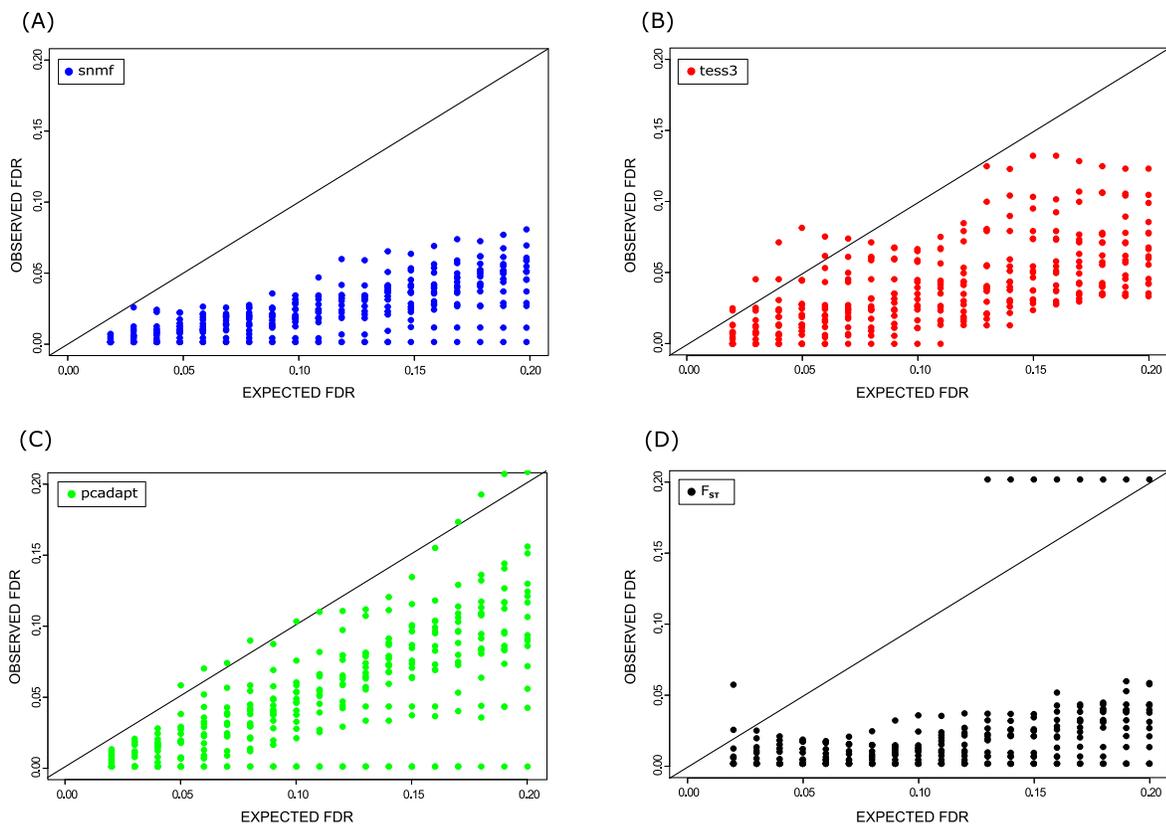


Figure S1. Observed false discovery rates. The tests are based on (A) *snmf*, (B) *tess3*, (C) *pcadapt*, (D) F_{ST} . Sixteen data sets containing 10% of outlier loci were used in each panel.

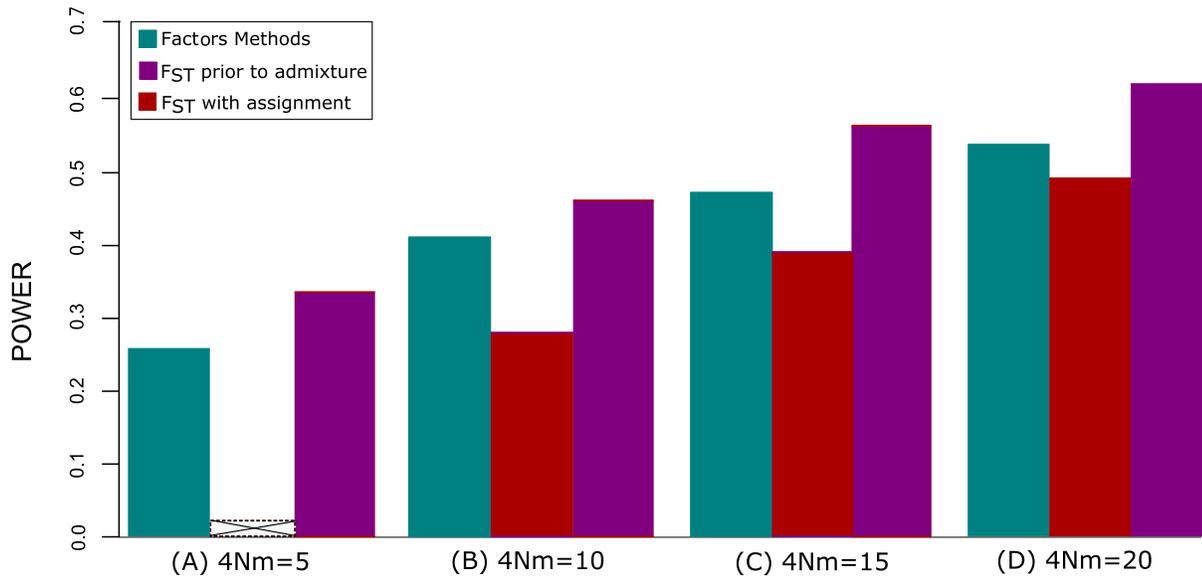


Figure S2. Power values of snmf, tess3, pcadapt (Factor methods) and classical F_{ST} tests with assignment and prior to admixture. All data sets contained 10% of outlier. Considering for a expected FDR of $\alpha = 0.1$: (A) Power values for the case $4Nm = 5$. The F_{ST} test based on assignment of 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 $4Nm = 20$.

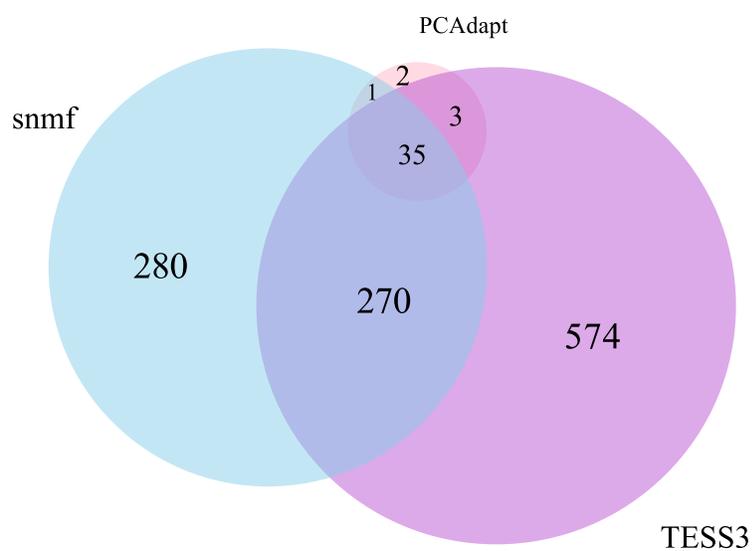


Figure S3. Venn diagrams. Intersection between the lists of loci obtained for each method applied to the *A. thaliana* data set. The *pcadapt* tests turned out to be more conservative.