bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

August 18, 2016

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

Running Title: Identifying outlier loci in continuous populations

Keywords: Genome Scans for Selection, Admixed Populations, Inference of Population Structure, Population Differentiation tests.

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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, 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 & Krakauer, 1973; Beaumont & Nichols, 1996; Akey *et al.*, 2002; Weir *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 & 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).

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 (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 & 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).

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 *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; Duforet-Frebourg *et al.*, 2016; Chen *et al.*, 2016; Galinsky *et al.*, 2016; Luu *et al.*, 2016).

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, 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.

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

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

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, 1951)

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

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}

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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.

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 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 *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 & 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}.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 & Rohlf, 2012). 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 *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 & Roeder (1999)).

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

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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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) for F_{ST} concludes the proof of equation (1).

3 Simulation experiments and data sets

Simple simulation models. In a first series of simulations, we created replicate data sets close to the underlying assumptions of population differentiation tests (Lewontin & Krakauer, 1973; Beaumont & 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 & Hewitt, 1985),

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

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).

Complex simulation models. To evaluate the power of tests in realistic landscape simulations, we used six publicly available data sets previously described by Lotterhos & 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 $\times$ 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 & Whitlock (2015)). All selected loci adapted to this landscape.

Computer programs We performed genome scans for selection using three factor methods: `snmf` (Frichot *et al.*, 2014), `tess3` (Caye *et al.*, 2016), `pcadapt` (Luu *et al.*, 2016; 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). 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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

to formula (1). We compared the results of `snmf` and `tess3` with the results of the program `pcadapt` (Luu *et al.*, 2016). 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.

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 & 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).

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 & 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`.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

4 Results

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 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 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 *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 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 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 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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

similar to the power obtained when we applied outlier tests to the data before admixture (Figure 2B). The results for data sets with 10% of selected loci were similar to those obtained with 5% of selected loci (Figure S2).

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 & 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 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 3B). 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.

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 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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 et al. 2016). For the three methods, the Manhattan plots exhibited peaks at the same chromosome positions (Figure 5). 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 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 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 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 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 (Mühlenbock *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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

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 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 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).

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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, 2011; 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.*, 2013). 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;

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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; Chen *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 & 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 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 & 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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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).

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).

Aknowlegments.

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 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, 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).

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

References

Adams J, Ward RH (1973). Admixture studies and the detection of selection. *Science* 180, 1137–1143.

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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

438 Alexander DH, Lange K (2011.) Enhancements to the ADMIXTURE algorithm for individual
439 ancestry estimation. *BMC Bioinformatics* 12, 246.

440 Alexander DH, Novembre J, Lange K (2009). Fast model-based estimation of ancestry in unrelated
441 individuals. *Genome Research* 19, 1655–1664.

442 Ascencio-Ibáñez JT, Sozzani R, Lee TJ, *et al.* (2008). Global analysis of *Arabidopsis* gene expres-
443 sion uncovers a complex array of changes impacting pathogen response and cell cycle during
444 geminivirus infection. *Plant Physiology* 148, 436–454.

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

447 Barton NH, Hewitt GM (1985). Analysis of hybrid zones. *Annual review of Ecology and Systematics*
448 16, 113–148.

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

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

453 Bersaglieri T, Sabeti PC, Patterson N, *et al.* (2004). Genetic signatures of strong recent positive
454 selection at the lactase gene. *The American Journal of Human Genetics* 74, 1111–1120.

455 Bierne N, Roze D, Welch JJ (2013). Pervasive selection or is it...? Why are FST outliers sometimes
456 so frequent? *Molecular Ecology* 22(8), 2061–2064.

457 Blumberg BS, Hesser JE (1971). Loci differentially affected by selection in two american black
458 populations. *Proceedings of the National Academy of Sciences* 68, 2554–2558.

459 Bryc K, Auton A, Nelson MR, *et al.* (2010). Genome-wide patterns of population structure and
460 admixture in West Africans and African Americans. *Proceedings of the National Academy of*
461 *Sciences* 107, 786–791.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

462 Catinot J, Huang JB, Huang PY, *et al.* (2015). ETHYLENE RESPONSE FACTOR 96 positively
463 regulates *Arabidopsis* resistance to necrotrophic pathogens by direct binding to GCC elements
464 of jasmonate-and ethylene-responsive defence genes. *Plant, Cell & Environment* 38, 2721–2734.

465 Cavalli-Sforza LL, Menozzi P, Piazza A. *The History and Geography of Human Genes*. Princeton
466 University Press, Princeton, USA, 1994.

467 Cavalli-Sforza LL (1966). Population structure and human evolution. *Proceedings of the Royal*
468 *Society of London B: Biological Sciences* 164, 362–379.

469 Caye K, Deist TM, Martins H, Michel O, François O (2016). TESS3: Fast inference of spatial
470 population structure and genome scans for selection. *Molecular Ecology Resources* 16, 540–548.

471 Chawade A, Bräutigam M, Lindlöf A, Olsson O, Olsson B (2007). Putative cold acclimation
472 pathways in *Arabidopsis thaliana* identified by a combined analysis of mRNA co-expression
473 patterns, promoter motifs and transcription factors. *BMC Genomics* 8, 1.

474 Chen C, Durand E, Forbes F, François O (2007). Bayesian clustering algorithms ascertaining
475 spatial population structure: A new computer program and a comparison study. *Molecular*
476 *Ecology Notes* 7, 747–756.

477 Chen GB, Lee SH, Zhu ZX, Benyamin B, Robinson MR (2016). EigenGWAS: finding loci under
478 selection through genome-wide association studies of eigenvectors in structured populations.
479 *Heredity* 117(1), 51–61.

480 Chen H, Kim HU, Weng H (2011). Malonyl-coA synthetase, encoded by ACYL ACTIVATING
481 ENZYME13, is essential for growth and development of *Arabidopsis*. *The Plant Cell* 23(6),
482 2247–2262.

483 Chevalier C, Nafati M, Mathieu-Rivet E, *et al.* (2011). Elucidating the functional role of endoredu-
484 plication in tomato fruit development. *Annals of Botany* 107(7), 1159–1169.

485 Darwin C. *On The Origin of Species by Means of Natural Selection.* John Murray, London, UK,
486 1859.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

487 DeBlasio SL, Luesse DL, Hangarter RP (2005). A plant-specific protein essential for blue-light-
488 lnduced chloroplast movements. *Plant Physiology* 139, 101–114.

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

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

493 Durand E, Jay F, Gaggiotti OE, François O (2009). Spatial inference of admixture proportions
494 and secondary contact zones. *Molecular Biology and Evolution* 26(9), 1963–1973.

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

498 Falush D, Van Dorp L, Lawson D (2016). A tutorial on how (not) to over-interpret STRUC-
499 TURE/ADMIXTURE bar plots. *bioRxiv* 066431, doi: http://dx.doi.org/10.1101/066431

500 François O, Blum MGB, Jakobsson M, Rosenberg NA (2008). Demographic history of Euro-
501 pean populations of *Arabidopsis thaliana*. *PLoS Genetics* 4(5), e1000075. doi: 10.1371/jour-
502 nal.pgen.1000075.

503 François O, Durand E (2010). Spatially explicit Bayesian clustering models in population genetics.
504 *Molecular Ecology Resources* 10, 773–784.

505 François O, Martins H, Caye K, Schoville SD (2016). Controlling false discoveries in genome scans
506 for selection. *Molecular Ecology* 25, 454–469.

507 Frichot E, François O (2015). LEA: an R package for Landscape and Ecological Association studies.
508 *Methods in Ecology and Evolution* 6(8), 925–929.

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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

511 Galinsky KJ, Bhatia G, Loh PR, *et al.* (2016). Fast principal component analysis reveals convergent
512 evolution of ADH1B in Europe and East Asia *The American Journal of Human Genetics* 98(3),
513 456–472.

514 Guo KM, Babourina O, Christopher DA, Borsics T, Rengel Z (2008). The cyclic nucleotide-gated
515 channel, AtCNGC10, influences salt tolerance in *Arabidopsis*. *Physiologia Plantarum* 134, 499–
516 507.

517 Han Y, Gu S, Oota H, *et al.* (2007) Evidence of positive selection on a class I ADH locus. *The*
518 *American Journal of Human Genetics* 80(3), 441–456.

519 Hancock AM, Brachi B, Faure N, *et al.* (2011) Adaptation to climate across the *Arabidopsis*
520 *thaliana* genome. *Science* 334, 83–86.

521 Hao W, Song M, Storey JD (2016). Probabilistic models of genetic variation in structured popu-
522 lations applied to global human studies. *Bioinformatics* 32(5), 713–721.

523 He XJ, Mu RL, Cao WH, Zhang ZG, Zhang JS, Chen SY (2005). AtNAC2, a transcription factor
524 downstream of ethylene and auxin signaling pathways, is involved in salt stress response and
525 lateral root development. *The Plant Journal* 44, 903–916.

526 Hewitt G (2000). The genetic legacy of the Quaternary ice ages. *Nature* 405(6789), 907–913.

527 Holsinger KE, Weir BS (2009). Genetics in geographically structured populations: Defining, esti-
528 mating and interpreting F_{ST}. *Nature Reviews Genetics* 10, 639–650.

529 Horton MW, Hancock AM, Huang YS, *et al.* (2012). Genome-wide patterns of genetic variation
530 in worldwide *Arabidopsis thaliana* accessions from the RegMap panel. *Nature Genetics* 44,
531 212–216.

532 Hudson RR (2002). Generating samples under a Wright–Fisher neutral model of genetic variation.
533 *Bioinformatics* 18(2), 337–338.

534 Kalinowski ST (2011). The computer program STRUCTURE does not reliably identify the main
535 genetic clusters within species: simulations and implications for human population structure.
536 *Heredity* 106(4), 625–632.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

Kim YO, Kang H (2005). Cold-inducible zinc finger-containing glycine-rich RNA-binding protein contributes to the enhancement of freezing tolerance in *Arabidopsis thaliana*. *The Plant Journal* 42, 890–900.

Lao O, Lu TT, Nothnagel M, *et al.* (2008). Correlation between genetic and geographic structure in Europe. *Current Biology* 18(16), 1241–1248.

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

Long JC (1991). The genetic structure of admixed populations. *Genetics* 127(2), 417–428.

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(5), 1031–1046.

Lu Y, Zhu J, Liu P (2005). A two-step strategy for detecting differential gene expression of cDNA microarray data. *Current Genetics* 47(2), 121–131.

Luu K, Bazin E, Blum MGB (2016). pcadapt: An R package for performing genome scans for selection based on principal component analysis. *bioRxiv* 056135, doi: http://dx.doi.org/10.1101/056135.

Manel S, Schwartz MK, Luikart G, Taberlet P (2003). Landscape genetics: Combining landscape ecology and population genetics. *Trends in Ecology & Evolution* 18, 189–197.

Mathieson I, Lazaridis I, Rohland N, *et al.* (2015). Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* 528, 499–503.

Mitchell-Olds T, Schmitt J (2006). Genetic mechanisms and evolutionary significance of natural variation in *Arabidopsis*. *Nature* 441, 947–952.

Mühlenbock P, Karpinska B, Karpinski S (2007). Oxidative stress and redox signalling in plants. *eLS*. doi: 10.1002/9780470015902.a0020135.

Nelson MR, Bryc K, King KS, *et al.* (2008). The population reference sample, POPRES: A resource for population, disease, and pharmacological genetics research. *The American Journal of Human Genetics* 83(3), 347–358.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

563 Novembre J, Johnson T, Bryc K, *et al.* (2008). Genes mirror geography within Europe. *Nature*
564 456, 98–101.

565 Patterson N, Price AL, Reich D (2006). Population structure and eigenanalysis. *PLoS Genetics*
566 2(12), e190.

567 Pritchard JK, Stephens M, Donnelly P (2000). Inference of population structure using multilocus
568 genotype data. *Genetics* 155(2), 945–959.

569 Puechmaille SJ (2016). The program STRUCTURE does not reliably recover the correct popula-
570 tion structure when sampling is uneven: subsampling and new estimators alleviate the problem.
571 *Molecular Ecology Resources* 16(3), 608–627.

572 Radin I, Mansilla N, Rödel G, Steinebrunner I (2015). The *Arabidopsis* COX11 homolog
573 is essential for cytochrome *c* oxidase activity. *Frontiers in Plant Science* 6, 1091.
574 doi:10.3389/fpls.2015.01091.

575 Raj A, Stephens M, Pritchard JK (2014). fastSTRUCTURE: Variational inference of population
576 structure in large SNP data sets. *Genetics* 197, 573–589.

577 Rajjou L, Belghazi M, Huguet R, *et al.* (2006). Proteomic investigation of the effect of salicylic
578 acid on *Arabidopsis* seed germination and establishment of early defense mechanisms. *Plant*
579 *Physiology* 141(3), 910–923.

580 Roth C, Wiermer M (2012). Nucleoporins Nup160 and Seh1 are required for disease resistance in
581 *Arabidopsis*. *Plant Signaling & Behavior* 7(10), 1212–1214.

582 Schoville SD, Bonin A, François O, *et al.* (2012). Adaptive genetic variation on the landscape:
583 Methods and cases. *Annual Review of Ecology, Evolution, and Systematics* 43, 23–43.

584 Sokal R, Rohlf F. *Biometry: The Principles and Practice of Statistics in Biological Research* (4th
585 edn). W.H. Freeman & Company, New York, NY, 2012.

586 Storey JD, Tibshirani R (2003). Statistical significance for genomewide studies. *Proceedings of the*
587 *National Academy of Sciences* 100(16), 9440–9445.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

588 Sun CW, Callis J (1997). Independent modulation of *Arabidopsis thaliana* polyubiquitin mRNAs
589 in different organs and in response to environmental changes. *The Plant Journal* 11, 1017–1027.

590 Tang H, Choudhry S, Mei R, *et al.* (2007). Recent genetic selection in the ancestral admixture of
591 Puerto Ricans. *The American Journal of Human Genetics* 81(3), 626–633.

592 Tang H, Peng J, Wang P, Risch NJ (2005). Estimation of individual admixture: Analytical and
593 study design considerations. *Genetic Epidemiology* 28, 289–301.

594 Visser M, Kayser M, Palstra RJ (2012). HERC2 rs12913832 modulates human pigmentation by
595 attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter.
596 *Genome Research* 22(3), 446–455.

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

599 Wang Y, Zhang WZ, Song LF, *et al.* (2008). Transcriptome analyses show changes in gene expres-
600 sion to accompany pollen germination and tube growth in *Arabidopsis*. *Plant Physiology* 148,
601 1201–1211.

602 Waples RS, Gaggiotti O (2006). What is a population? An empirical evaluation of some genetic
603 methods for identifying the number of gene pools and their degree of connectivity. *Molecular*
604 *Ecology* 15, 1419–1439.

605 Weigel D, Mott R (2009). The 1001 Genomes Project for *Arabidopsis thaliana*. *Genome Biology*
606 10(5), 1–5.

607 Weir BS, Cardon LR, Anderson AD, Nielsen DM, Hill WG (2005). Measures of human population
608 structure show heterogeneity among genomic regions. *Genome Research* 15(11), 1468–1476.

609 Weir BS, Cockerham CC (1984). Estimating F-statistics for the analysis of population structure.
610 *Evolution* 38(6), 1358–1370.

611 Weir BS. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sinauer Assoc.
612 Inc., Sunderland, MA, USA, 1996.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

613 Wollstein A, Lao O (2015). Detecting individual ancestry in the human genome. *Investigative*
614 *Genetics* 6, 1–12.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

6 Figures and Tables

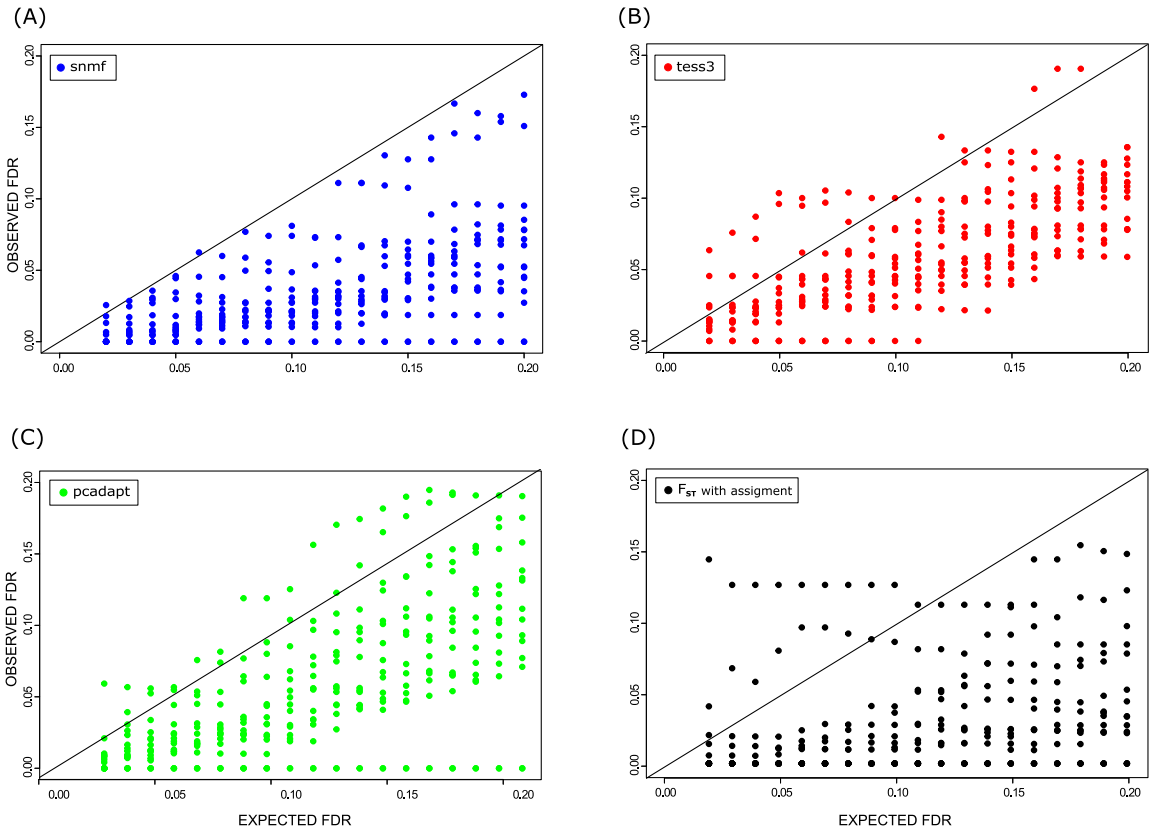

Figure 1. FDR for simulations of admixed populations. Simulation of ancestral populations based on 2-island models with various levels of population differentiation and selection. Sixteen data sets contained 5% of truly selected loci. Observed false discovery rates for an expected level of FDR equal 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.'s (2016) `pcadapt` statistic, (D) Standard F_{ST} test based on assignment of individuals to their most likely genetic cluster.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

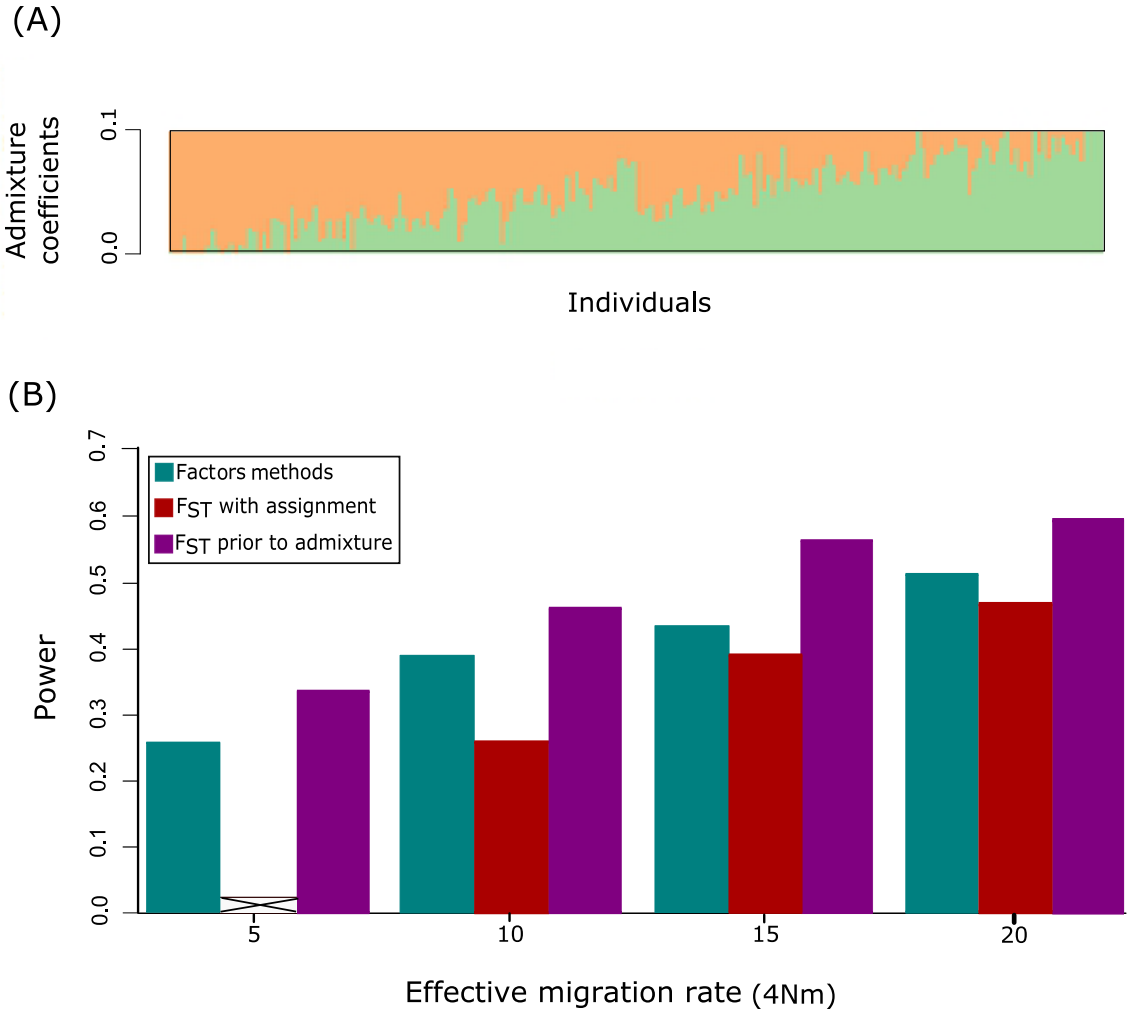

Figure 2. Power in simulations of admixed populations. Simulations of ancestral populations based on 2-island models with various levels of selection and background of levels of population differentiation ($4Nm$). Sixteen data sets contained 5% of truly selected loci. (A) Individual ancestry coefficients 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.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

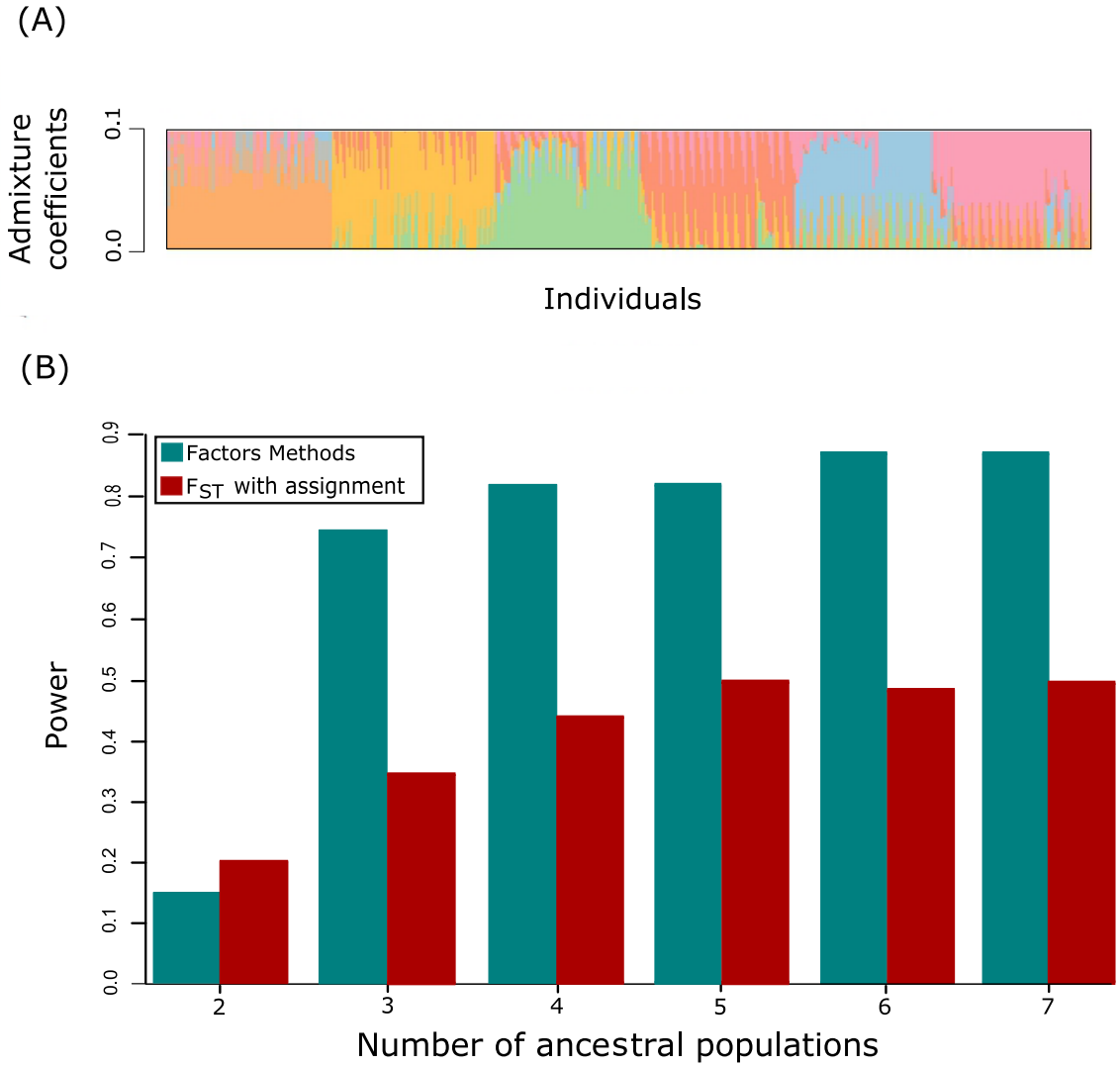

Figure 3. Power in simulations of range expansions. (A) Individual ancestry coefficients estimated using `snmf` with $K = 6$ ancestral populations. (B) Power estimates for tests based on factor methods and for F_{ST} tests in which individuals were assigned to their most likely cluster. Power values were computed by considering an expected FDR value equal to 0.1. Factor methods included `snmf` and `pcadapt`.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

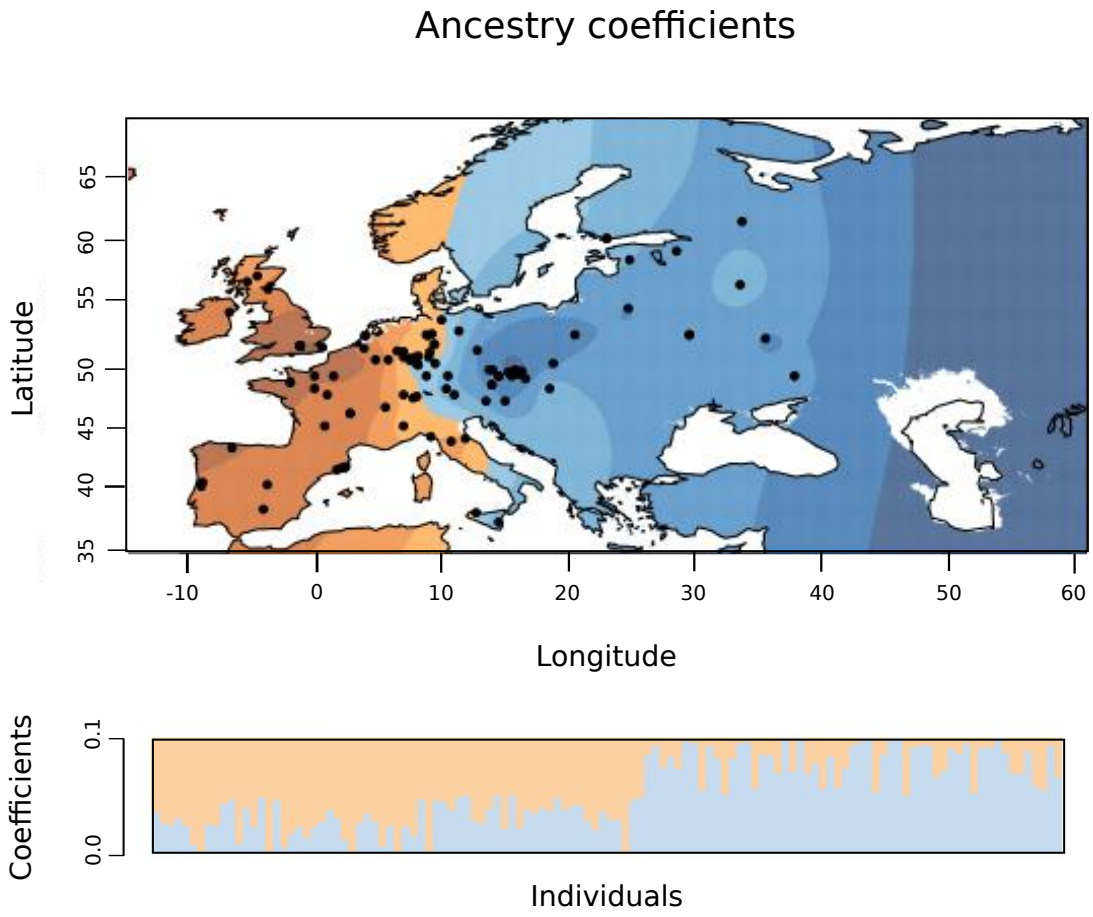

Figure 4. Ancestry coefficients for *Arabidopsis thaliana***.** Coefficients estimated using `snmf` with $K = 2$ ancestral populations interpolated on a geographic map of Europe.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

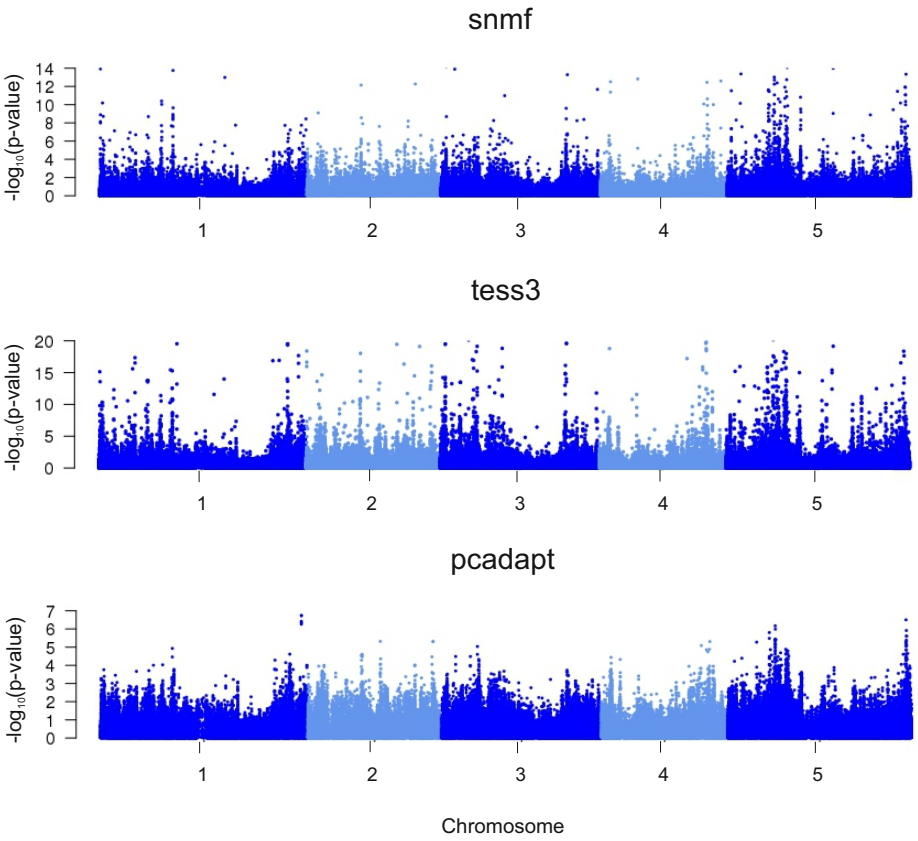

Figure 5. Manhattan plots of minus log10(p-values) for *A. thaliana*. Tests using (A) `snmf`, (B) `tess3` and (C) `pcadapt`. The tests based on `pcadapt` were more conservative than the tests based on the other methods.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

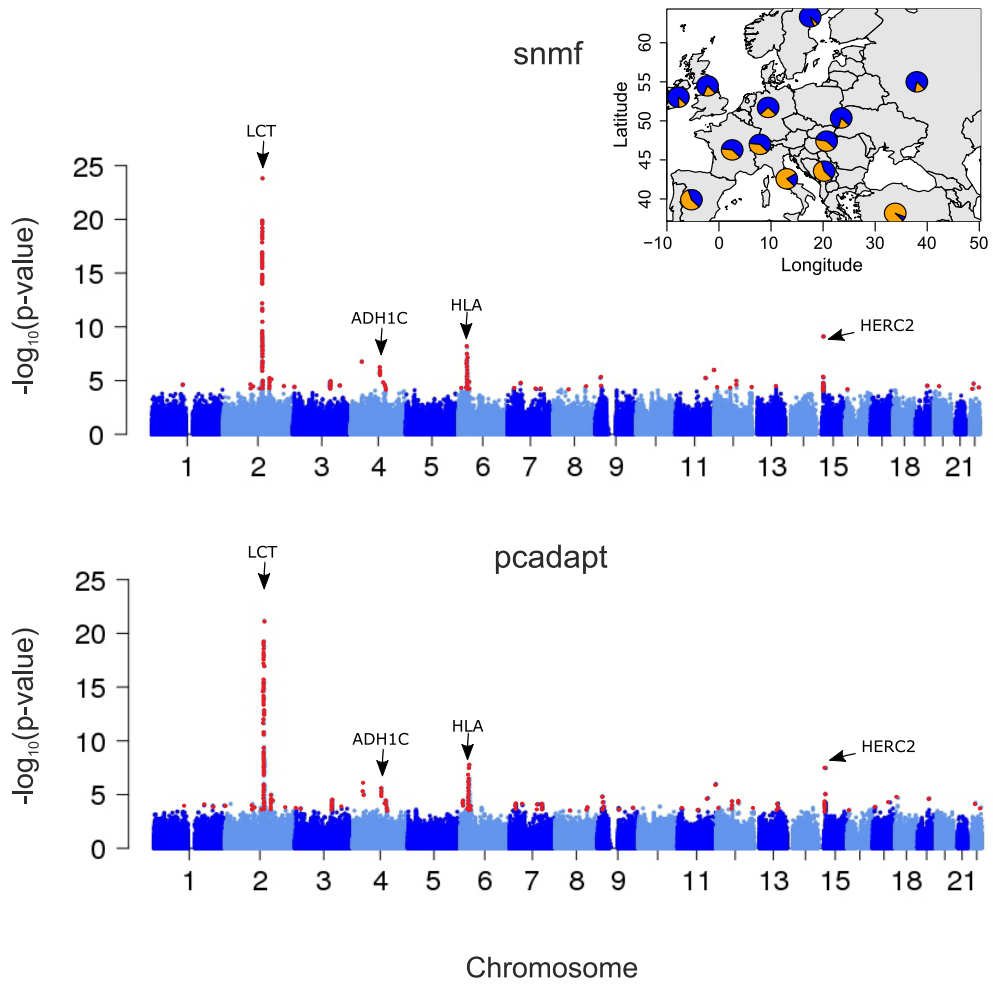

Figure 6. Manhattan plots of minus log10(p-values) for Europeans (POPRES data set). Tests using (A) `snmf` and (B) `pcadapt`. Candidate loci detected by genome scans for selection are colored in red for an expected FDR level of 10%. The inserted figure displays population structure estimated with `snmf` with $K = 2$ populations.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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 & 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 *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 & Kang (2005)
	6202633	AT5G18630	Lipid metabolic process	Wang *et al.* (2008)
	6947843	AT5G20540	Unknown	
	6952417	AT5G20550	Oxidation-reduction process	
	6956660	AT5G20570 (ATRBX1)	Protein ubiquitination	Ascencio-Ibáñez *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 & 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 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.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

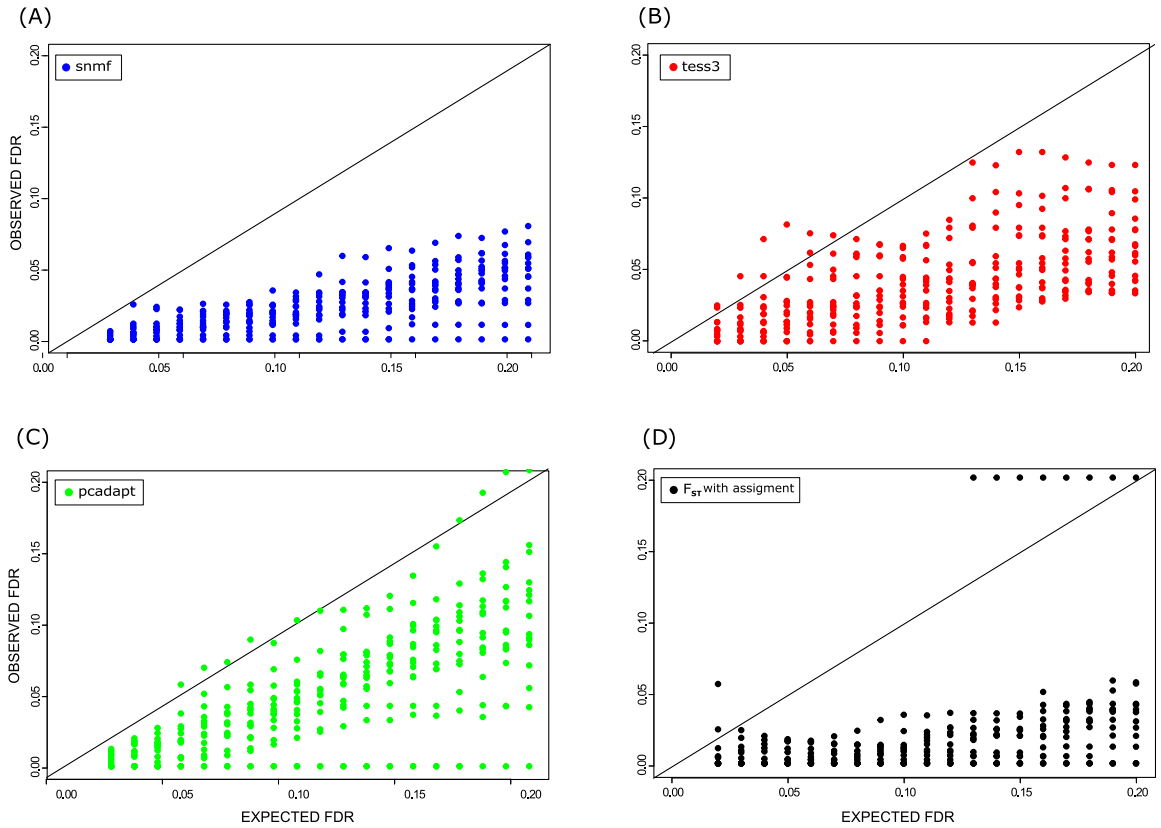

Figure S1. FDR for simulations of admixed populations (10% of outliers). Simulation of ancestral populations based on 2-island models with various levels of population differentiation and selection. Sixteen data sets contained 10% of truly selected loci. Observed false discovery rates for an expected 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.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

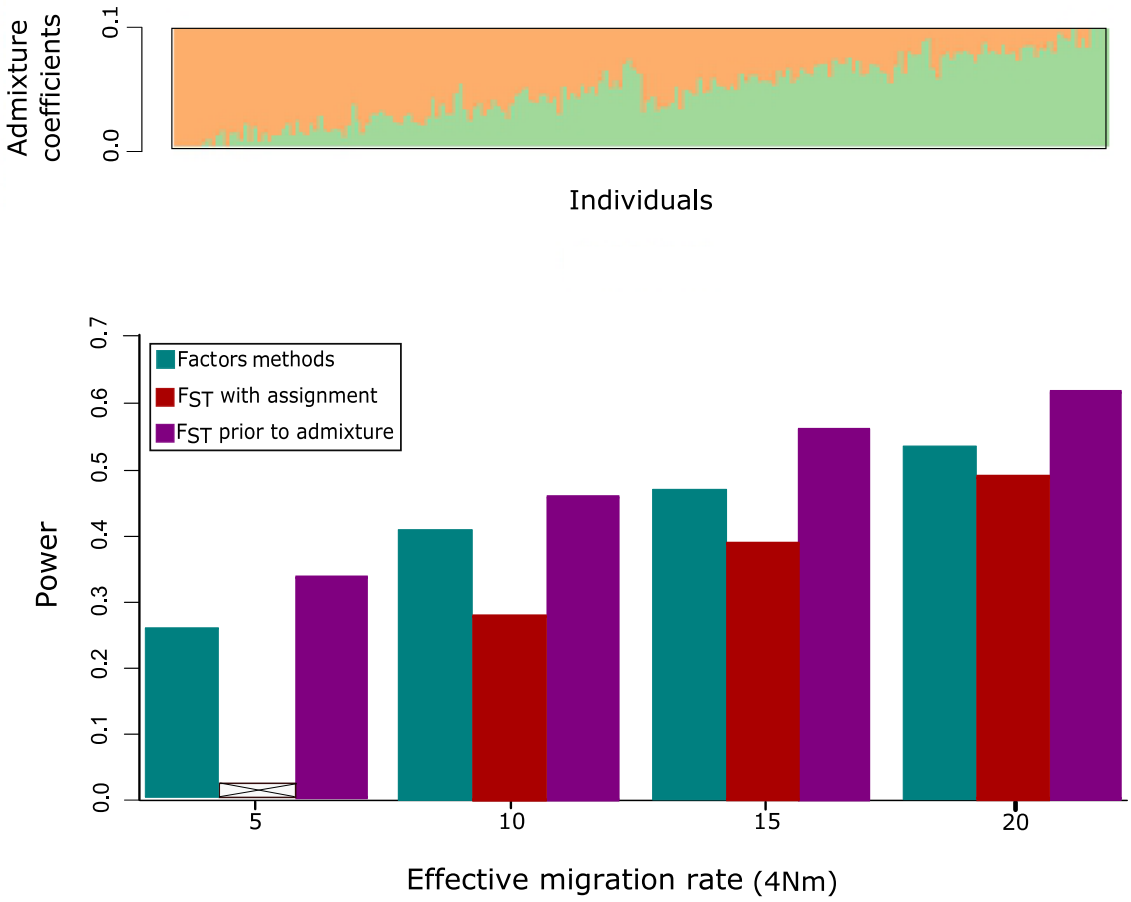

Figure S2. Power in simulations of admixed populations (10% of outliers). Simulations of ancestral populations based on 2-island models with various levels of selection and background of levels of population differentiation ($4Nm$). Sixteen data sets contained 10% of truly selected loci. (A) Individual 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.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

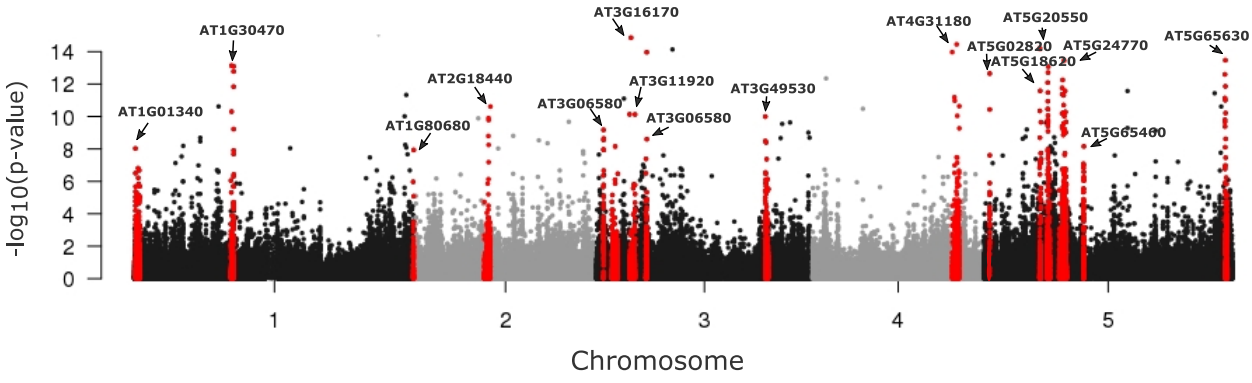

Figure S3. Manhattan plot of minus log10(p-values) for *A. thaliana*. The candidate regions are colored in red. Those regions correspond to an expected FDR level of 1% for `snmf` and `tess3` having more than 5 SNPs in each region.

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

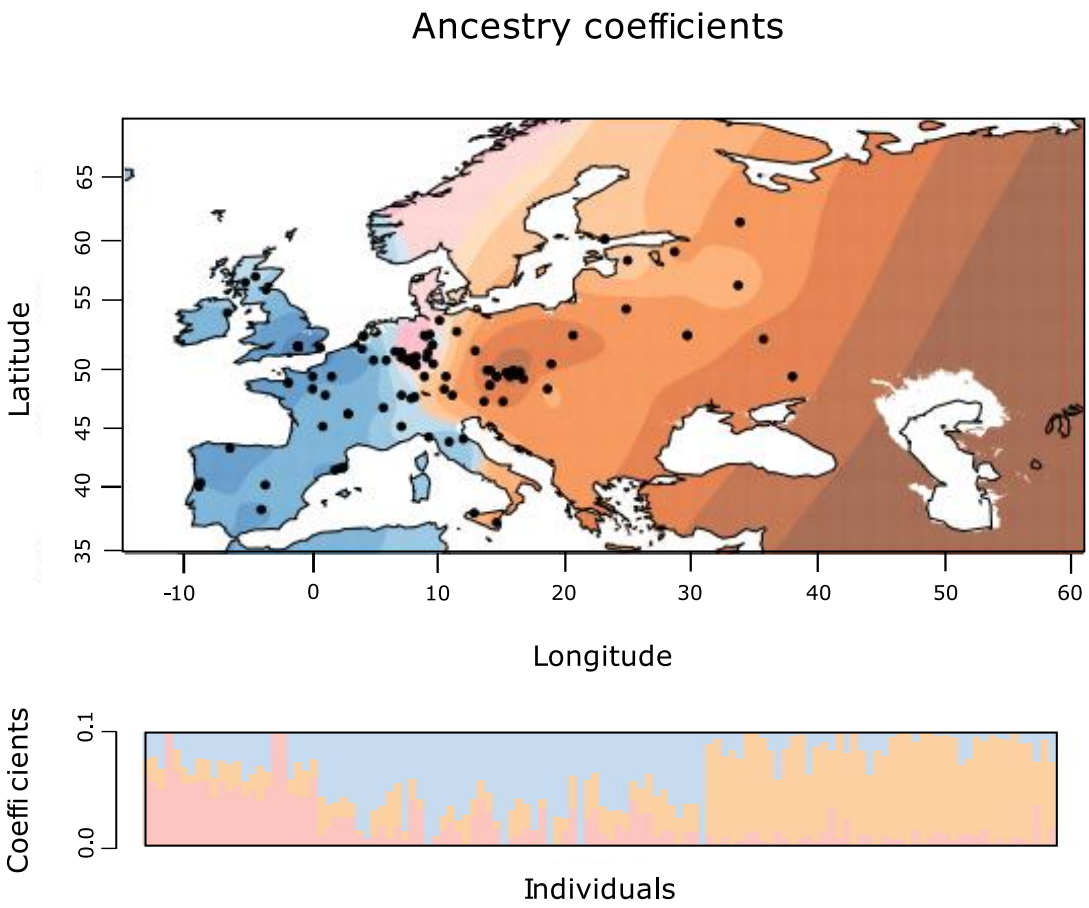

Figure S4. Geographic map of ancestry coefficients for *Arabidopsis thaliana* **using** `snmf` **with** $K = 3$ **ancestral populations.**

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

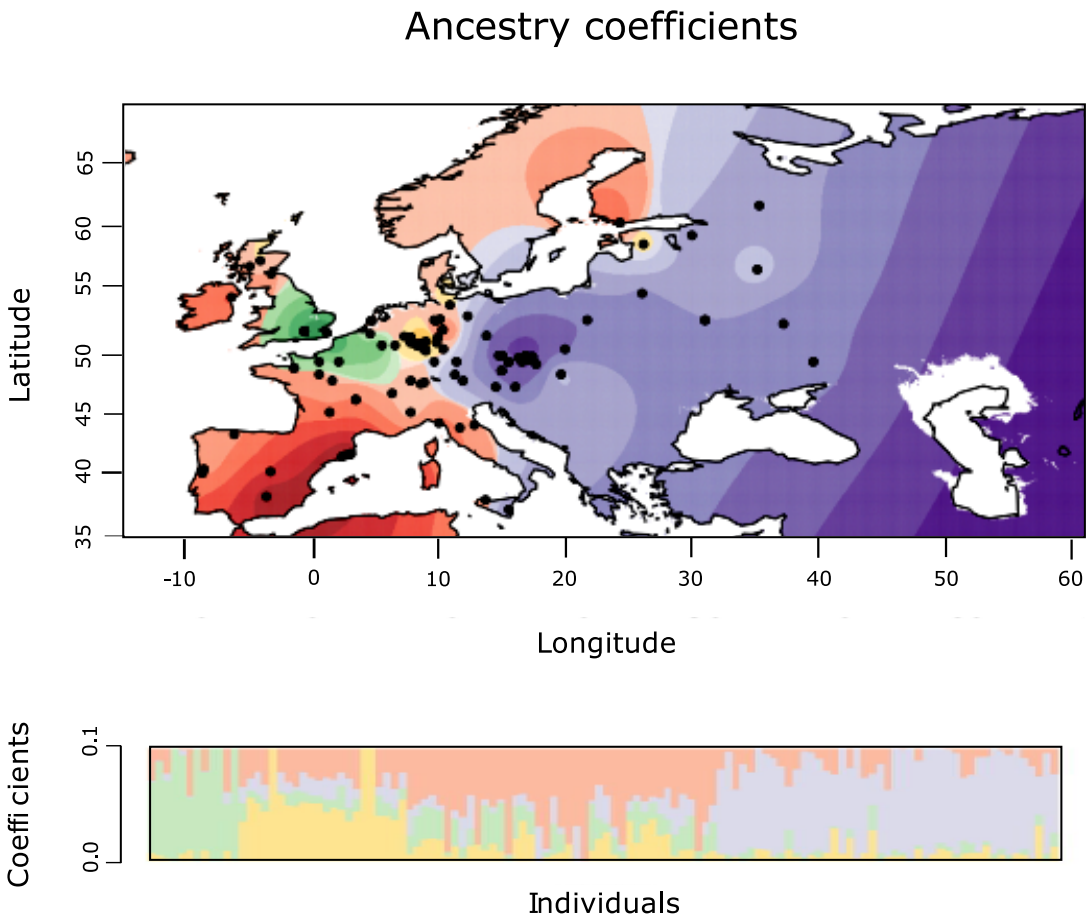

Figure S5. Geographic map of ancestry coefficients for *Arabidopsis thaliana* **using** `snmf` **with** $K = 4$ **ancestral populations.**

bioRxiv preprint doi: https://doi.org/10.1101/054585; this version posted August 18, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

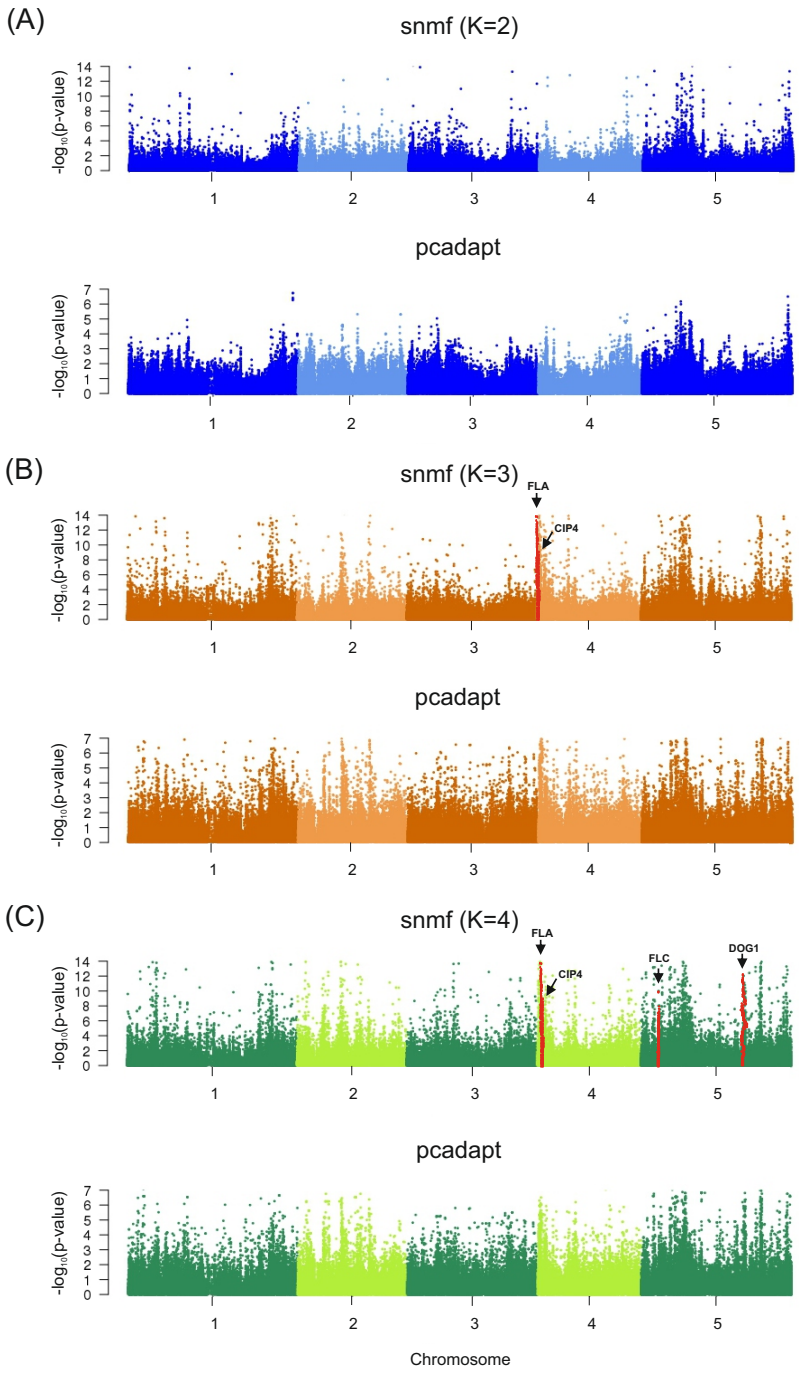

Figure S6. Manhattan plots of minus log10(p-values) for *A. thaliana*. Tests using (A) `snmf` with $K = 2$ ancestral populations and `pcadapt` with 1 principal component, (B) `snmf` with $K = 3$ ancestral populations and `pcadapt` with 2 principal components, (C) `snmf` with $K = 4$ ancestral populations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