bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity.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.

Submitted to the Annals of Applied Statistics

FAST INFERENCE OF INDIVIDUAL ADMIXTURE COEFFICIENTS USING GEOGRAPHIC DATA

By Kevin Caye*, Flora Jay†, Olivier Michel*, and Olivier François*

Université Grenoble-Alpes and Université Paris-Sud†*

Accurately evaluating the distribution of genetic ancestry across geographic space is one of the main questions addressed by evolutionary biologists. This question has been commonly addressed through the application of Bayesian estimation programs allowing their users to estimate individual admixture proportions and allele frequencies among putative ancestral populations. Following the explosion of high-throughput sequencing technologies, several algorithms have been proposed to cope with computational burden generated by the massive data in those studies. In this context, incorporating geographic proximity in ancestry estimation algorithms is an open statistical and computational challenge. In this study, we introduce new algorithms that use geographic information to estimate ancestry proportions and ancestral genotype frequencies from population genetic data. Our algorithms combine matrix factorization methods and spatial statistics to provide estimates of ancestry matrices based on least-squares approximation. We demonstrate the benefit of using spatial algorithms through extensive computer simulations, and we provide an example of application of our new algorithms to a set of spatially referenced samples for the plant species *Arabidopsis thaliana*. Without loss of statistical accuracy, the new algorithms exhibit runtimes that are much shorter than those observed for previously developed spatial methods. Our algorithms are implemented in the `R` package, `tess3r`.

1. Introduction. High-throughput sequencing technologies have enabled studies of genetic ancestry for model and non-model species at an unprecedented pace. In this context, ancestry estimation algorithms are important for demographic analysis, medical genetics including genome-wide association studies, conservation and landscape

Keywords and phrases: Ancestry Estimation Algorithms, Genotypic Data, Geographic Data, Fast Algorithms

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

genetics (Pritchard et al., 2000; Tang et al., 2005; Schraiber et al., 2015; Segelbacher et al., 2010; François et al., 2016). With increasingly large data sets, Bayesian approaches to the inference of population structure, exemplified by the computer program `structure` (Pritchard et al., 2000), have been replaced by approximate algorithms that run several orders faster than the original version (Tang et al., 2005; Alexander et al., 2011; Frichot et al., 2014; Raj et al., 2014). Considering K ancestral populations or genetic clusters, those algorithms estimate ancestry coefficients following two main directions: model-based and model-free approaches. In model-based approaches, a likelihood function is defined for the matrix of ancestry coefficients, and estimation is performed by maximizing the log-likelihood function. For `structure` and related models, model assumptions include linkage equilibrium and Hardy-Weinberg equilibrium in ancestral populations. The first approximation to the original algorithm was based on an expectation-minimization algorithm (Tang et al., 2005), and more recent likelihood algorithms are implemented in the programs `admixture` and `faststructure` (Alexander et al., 2011; Raj et al., 2014). In model-free approaches, ancestry coefficients are estimated by using least-squares methods or factor analysis. Model-free methods make no assumptions about the biological processes that have generated the data. To estimate ancestry matrices, Engelhardt et al. (2010) proposed to use sparse factor analysis, Frichot et al. (2014) used sparse non-negative matrix factorization algorithms, and Popescu et al. (2014) used kernel-principal component analysis. Least-squares methods accurately reproduce the results of likelihood approaches under the model assumptions of those methods. In addition, model-free methods provide approaches that are valid when the assumptions of likelihood approaches are not met (Frichot et al., 2014). Model-free methods are generally faster than model-based methods.

Among model-based approaches to ancestry estimation, an important class of methods have improved the Bayesian model of `structure` by incorporating geographic data through spatially informative prior distributions (Chen et al., 2007; Corander et al., 2008). Under isolation-by-distance patterns (Wright, 1943; Malécot, 1948), spatial algorithms provide more robust estimates of population structure than non-spatial algorithms which can lead to biased estimates of the number of clusters (Durand et al., 2009). Some Bayesian methods are based on Markov chain Monte Carlo algorithms which are computer-intensive (François

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

et al., 2010). Recent efforts to improve the inference of ancestral relationships in a geographical context have mainly focused on the localization of recent ancestors (Baran et al., 2013; Lao et al., 2014; Yang et al., 2014; Bhaskar et al., 2017; Rañola et al., 2014). In these applications, spatial information is used in a predictive framework that assigns ancestors to putative geographic origins. While fast geographic estimation of individual ancestry proportions has been proposed previously (Caye et al., 2016; Bradburd et al., 2016), there is a growing need to develop individual ancestry estimation algorithms that reduce computational cost in a geographically explicit framework.

In this study, we present two new algorithms for the estimation of ancestry matrices based on geographic and genetic data. The new algorithms solve a least squares optimization problem as defined by Caye et al. (2016), based on Alternating Quadratic Programming (AQP) and Alternating Projected Least Squares (APLS). While AQP algorithms have a well-established theoretical background (Bertsekas, 1995), this is not the case of APLS algorithms. Using coalescent simulations, we provide evidence that the estimates computed by APLS algorithms are good approximations to the solutions of AQP algorithms. In addition, we show that the performances of APLS algorithms scale to the dimensions of modern data sets. We discuss the application of our algorithms to data from European ecotypes of *Arabidopsis thaliana*, for which individual genomic a geographic data are available (Horton et al., 2012).

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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. New methods. In this section we present two new algorithms for estimating individual admixture coefficients and ancestral genotype frequencies assuming K ancestral populations. In addition to genotypes, the new algorithms require individual geographic coordinates of sampled individuals.

Q and G-matrices. Consider a genotypic matrix, $\mathbf{Y}$, recording data for n individuals at L polymorphic loci for a p-ploid species (common values for p are $p = 1, 2$). For autosomal SNPs in a diploid organism, the genotype at locus ℓ is an integer number, 0, 1 or 2, corresponding to the number of reference alleles at this locus. In our algorithms, disjunctive forms are used to encode each genotypic value as the indicator of a heterozygote or a homozygote locus /citepFrichot2014. For a diploid organism each genotypic value $0, 1, 2$ is encoded as 100, 010 and 001. For p-ploid organisms, there are $(p + 1)$ possible genotypic values at each locus, and each value corresponds to a unique disjunctive form. While our focus is on SNPs, the algorithms presented in this section extend to multi-allelic loci without loss of generality. Moreover, the method can be easily extended to genotype likelihoods by using the likelihood to encode each genotypic value (Korneliussen et al., 2014).

Our algorithms provide statistical estimates for the matrix $\mathbf{Q} \in \mathbb{R}^{K\times n}$ which contains the admixture coefficients, $\mathbf{Q}_{i,k}$, for each sampled individual, i, and each ancestral population, k. The algorithms also provide estimates for the matrix $\mathbf{G} \in \mathbb{R}^{(p+1)L\times K}$, for which the entries, $\mathbf{G}_{(p+1)\ell+j,k}$, correspond to the frequency of genotype j at locus ℓ in population k. Obviously, the Q and G-matrices must satisfy the following set of probabilistic constraints

$$\mathbf{Q}, \mathbf{G} \geq 0\,, \quad \sum_{k=1}^{K} \mathbf{Q}_{i,k} = 1\,, \quad \sum_{j=0}^{p} \mathbf{G}_{(p+1)\ell+j,k} = 1\,,$$

for all i, k and ℓ. Using disjunctive forms and the law of total probability, estimates of $\mathbf{Q}$ and $\mathbf{G}$ can be obtained by factorizing the genotypic matrix as follows $\mathbf{Y}{=}\mathbf{Q}\,\mathbf{G}^T$ (Frichot et al., 2014). Thus the inference problem can be solved by using constrained nonnegative matrix factorization methods (Lee et al., 1999; Cichocki et al., 2009). In the sequel, we shall use the notations Δ_Q and Δ_G to represent the sets of probabilistic constraints put on the $\mathbf{Q}$ and $\mathbf{G}$ matrices respectively.

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

Geographic weighting. Geography is introduced in the matrix factorization problem by using weights for each pair of sampled individuals. The weights impose regularity constraints on ancestry estimates over geographic space. The definition of geographic weights is based on the spatial coordinates of the sampling sites, (x_i). Samples close to each other are given more weight than samples that are far apart. The computation of the weights starts with building a complete graph from the sampling sites. Then the weight matrix is defined as follows

$$w_{ij} = \exp(-\mathrm{dist}(x_i, x_j)^2/\sigma^2),$$

where $\mathrm{dist}(x_i, x_j)$ denotes the geodesic distance between sites x_i and x_j, and σ is a range parameter.

Next, we introduce the *Laplacian matrix* associated with the geographic weight matrix, $\mathbf{W}$. The Laplacian matrix is defined as $\mathbf{\Lambda} = \mathbf{D} - \mathbf{W}$ where $\mathbf{D}$ is a diagonal matrix with entries $\mathbf{D}_{i,i} = \sum_{j=1}^{n} \mathbf{W}_{i,j}$, for $i = 1, \ldots, n$ (Belkin et al.). Elementary matrix algebra shows that (Cai et al., 2011)

$$\mathrm{Tr}(\mathbf{Q}^T \mathbf{\Lambda} \mathbf{Q}) = \frac{1}{2} \sum_{i,j=1}^{n} w_{ij} \|\mathbf{Q}_{i,\cdot} - \mathbf{Q}_{j,\cdot}\|^2 .$$

In our approach, assuming that geographically close individuals are more likely to share ancestry than individuals at distant sites is thus equivalent to minimizing the quadratic form $\mathcal{C}(\mathbf{Q}) = \mathrm{Tr}(\mathbf{Q}^T \mathbf{\Lambda} \mathbf{Q})$ while estimating the matrix $\mathbf{Q}$.

Least-squares optimization problems. Estimating the matrices $\mathbf{Q}$ and $\mathbf{G}$ from the observed genotypic matrix $\mathbf{Y}$ is performed through solving an optimization problem defined as follows (Caye et al., 2016)

$$\begin{aligned} \min_{Q,G} \quad & \mathrm{LS}(\mathbf{Q}, \mathbf{G}) = \|\mathbf{Y} - \mathbf{Q}\mathbf{G}^T\|_{\mathrm{F}}^2 + \alpha \mathcal{C}(\mathbf{Q}), \\ \text{s.t.} \quad & \mathbf{Q} \in \Delta_Q, \\ & \mathbf{G} \in \Delta_G. \end{aligned} \tag{2.1}$$

The notation $\|\mathbf{M}\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix, $\mathbf{M}$. The regularization parameter α controls the regularity of ancestry estimates over geographic space. Large values of α imply that ancestry coefficients

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

have similar values for nearby individuals, whereas small values ignore spatial autocorrelation in observed allele frequencies.

The Alternating Quadratic Programming (AQP) method. Because the polyedrons Δ_Q and Δ_G are convex sets and the LS function is convex with respect to each variable $\mathbf{Q}$ or $\mathbf{G}$ when the other one is fixed, the problem (2.1) is amenable to the application of block coordinate descent (Bertsekas, 1995). The APQ algorithm starts from initial values for the G and Q-matrices, and alternates two steps. The first step computes the matrix $\mathbf{G}$ while $\mathbf{Q}$ is kept fixed, and the second step permutates the roles of $\mathbf{G}$ and $\mathbf{Q}$. Let us assume that $\mathbf{Q}$ is fixed and write $\mathbf{G}$ in a vectorial form, $g = \mathrm{vec}(\mathbf{G}) \in \mathbb{R}^{K(p+1)L}$. The first step of the algorithm actually solves the quadratic programming subproblem,

$$g^\star = \arg\min_{g \in \Delta_G} (-2 v_Q^T g + g^T \mathbf{D}_Q g)\,, \tag{2.2}$$

where $\mathbf{D}_Q = \mathbf{I}_{(p+1)L} \otimes \mathbf{Q}^T\mathbf{Q}$ and $v_Q = \mathrm{vec}(\mathbf{Q}^T\mathbf{Y})$. Here, $\otimes$ denotes the Kronecker product and $\mathbf{I}_d$ is the identity matrix with d dimensions. The block structure of the matrix $\mathbf{D}_Q$ allows us to decompose the subproblem (2.2) into L independent quadratic programming problems with $K(p+1)$ variables. Now, consider that $\mathbf{G}$ is the value obtained after the first step of the algorithm, and write $\mathbf{Q}$ in a vectorial form, $q = \mathrm{vec}(\mathbf{Q}) \in \mathbb{R}^{nK}$. The second step solves the following quadratic programming subproblem. Find

$$q^\star = \arg\min_{q \in \Delta_Q} (-2 v_G^T q + q^T \mathbf{D}_G q)\,, \tag{2.3}$$

where $\mathbf{D}_G = \mathbf{I}_n \otimes \mathbf{G}^T\mathbf{G} + \alpha \mathbf{\Lambda} \otimes \mathbf{I}_K$ and $v_G = \mathrm{vec}(\mathbf{G}^T\mathbf{Y}^T)$. Unlike subproblem (2.2), subproblem (2.3) can not be decomposed into smaller problems. Thus, the computation of the second step of the AQP algorithm implies to solve a quadratic programming problem with nK variables which can be problematic for large samples (n is the sample size). The AQP algorithm is described in details in Appendix A.1. For AQP, we have the following convergence result.

THEOREM 2.1. *The AQP algorithm converges to a critical point of problem* (2.1).

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

PROOF. The quadratic convex functions defined in subproblems (2.2) and (2.3) have finite lower bounds. The convex sets Δ_Q and Δ_G are compact non-empty sets. Thus the sequence generated by the AQP algorithm is well-defined, and has limit points. According to Corollary 2 of Grippo et al. (2000), we conclude that the AQP algorithm converges to a critical point of problem (2.1). □

Alternating Projected Least-Squares (APLS). In this paragraph, we introduce an APLS estimation algorithm which approximates the solution of problem (2.1), and reduces the complexity of the AQP algorithm. The APLS algorithm starts from initial values of the G and Q-matrices, and alternates two steps. The matrix $\mathbf{G}$ is computed while $\mathbf{Q}$ is kept fixed, and *vice versa*. Assume that the matrix $\mathbf{Q}$ is known. The first step of the APLS algorithm solves the following optimization problem. Find

$$\mathbf{G}^\star = \arg\min \|\mathbf{Y} - \mathbf{Q}\mathbf{G}^T\|_{\mathrm{F}}^2 \,. \tag{2.4}$$

This operation can be done by considering $(p+1)L$ (the number of columns of $\mathbf{Y}$) independent optimization problems running in parallel. The operation is followed by a projection of $\mathbf{G}^\star$ on the polyedron of constraints, Δ_G. For the second step, assume that $\mathbf{G}$ is set to the value obtained after the first step is completed. We compute the eigenvectors, $\mathbf{U}$, of the Laplacian matrix, and we define the diagonal matrix $\mathbf{\Delta}$ formed by the eigenvalues of $\mathbf{\Lambda}$ (The eigenvalues of $\mathbf{\Lambda}$ are non-negative real numbers). According to the spectral theorem, we have

$$\mathbf{\Lambda} = \mathbf{U}^T \mathbf{\Delta} \mathbf{U} \,.$$

After this operation, we project the data matrix $\mathbf{Y}$ on the basis of eigenvectors as follows

$$\mathrm{proj}(\mathbf{Y}) = \mathbf{U}\mathbf{Y} \,,$$

and, for each individual, we solve the following optimization problem

$$q_i^\star = \arg\min \|\mathrm{proj}(\mathbf{Y})_i - \mathbf{G}q\|^2 + \alpha\lambda_i \|q\|^2 \,, \tag{2.5}$$

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

where $\text{proj}(\mathbf{Y})_i$ is the ith row of the projected data matrix, $\text{proj}(\mathbf{Y})$, and λ_i is the ith eigenvalue of $\mathbf{\Lambda}$. The solutions, $q_i^\star$, are then concatenated into a matrix, $\text{conc}(q)$, and $\mathbf{Q}$ is defined as the projection of the matrix $\mathbf{U}^T\text{conc}(q)$ on the polyedron Δ_Q. The complexity of step (2.5) grows linearly with n, the number of individuals. While the theoretical convergence properties of AQP algorithms are lost for APLS algorithms, the APLS algorithms are expected to be good approximations of AQP algorithms. The APLS algorithm is described in details in Appendix A.2.

Choice of hyper-parameters. In ancestry estimation programs, a number of practices have evolved in order to set the model hyper-parameters. Those practices rely on heuristics or empirical rules for determining the prior parameters. For example, the program `structure` implements weakly informative prior distributions for ancestry proportions (Wang, 2017), the program `admixture` has a set of regularization parameters that encourages shrinkage and aggressive parsimony on ancestry estimates (Alexander et al., 2011), and so does the Bayesian version TESS 2.3 (Durand et al., 2009). Choosing the number of ancestral populations is based on cross-validation methods or information theoretic measures. Our model has three hyper-parameters: the number of factors, K, the penalty constant, α, and the range parameter, σ. Determining those constants is notoriously difficult and can be costly in applications. In order to reduce the computational burden, the hyper-parameters α and σ are set as user-defined options. This option allows an advanced user to explore different values with cross-validation or with her own heuristics. Less advanced users could use the default values of the hyper-parameters evaluated in our simulation study.

Range parameter. Testing correlations between genetic and geographic data has a long tradition in population genetics. Popular approaches are based on Mantel tests (Mantel, 1967) and spatial autocorrelation measures (Hardy et al., 1999; Epperson et al., 1996). Prior to the application of our spatial ancestry estimation program, we investigated biologically relevant values for the range parameter by using spatial variograms (Cressie, 1993). The variogram was extended to genotypic

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

data as follows

$$\gamma(h) = \frac{1}{2|N(h)|} \sum_{i,j \in N(h)} \frac{1}{L} \sum_{l=1}^{(p+1)L} |Y_{i,l} - Y_{j,l}|, \tag{2.6}$$

where $N(h)$ is defined as the set of individuals separated by geographic distance h. Visualizing the variogram provides useful information on the level of spatial autocorrelation in the data, and yields empirical estimates of the range parameter. More naive estimates such as an average geodesic distance computed over a fraction of neighboring sites in the sample also performed well in simulations, and they are also proposed to the program users.

Regularization parameter. A default value for the regularization parameter α was set so that the weights for the loss function and for the penalty term $\mathcal{C}(\mathbf{Q})$ are of similar order. We proposed to divide each term by its maximum value. This amounts to consider α equal to $L/\lambda_{\max}$, where $\lambda_{\max}$ is the largest eigenvalue of the Laplacian matrix (The Laplacian matrix has nonnegative eigenvalues).

Number of factors. The number of ancestral populations, K, can be evaluated by using a cross-validation technique based on imputation of masked genotypes (Wold, 1978; Eastment et al., 1982; Alexander et al., 2011; Frichot et al., 2014). The cross-validation procedure partitions the genotypic matrix entries into a learning set and a test set in which 5% of all genotypes are tagged as masked entries. The genotype probabilities for the masked entries are predicted from the factor estimates obtained from unmasked entries. Then, the error between the predicted and truly observed genotype frequencies is computed, and smaller values of that criterion indicate better choices.

Comparison with `tess3`. The algorithm implemented in a previous version of `tess3` also provides another approximation of the solution of problem (2.1). The `tess3` algorithm first computes a Cholesky decomposition of the Laplacian matrix. Then, by a change of variables, the least-squares problem is transformed into a sparse nonnegative matrix factorization problem (Caye et al., 2016). Solving the sparse nonnegative matrix factorization problem relies on the application of existing methods (Kim et al., 2011; Frichot et al., 2014). The methods

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

implemented in `tess3` have an algorithmic complexity that increases linearly with the number of loci and the number of clusters. They lead to estimates that accurately reproduce those of the Monte Carlo algorithms implemented in the Bayesian method `tess` 2.3 (Caye et al., 2016). Like for the AQP method, the `tess3` algorithms have an algorithmic complexity that increases quadratically with the sample size.

Ancestral population differentiation statistics and local adaptation scans. Assuming K ancestral populations, the Q and G-matrices obtained from the AQP and from the APLS algorithms were used to compute single-locus estimates of a population differentiation statistic similar to F_{ST}, as follows

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

where q_k is the average of ancestry coefficients over sampled individuals, $q_k = \sum_{i=1}^{n} Q_{i,k}/n$, for the cluster k, f_k is the ancestral allele frequency in population k at the locus of interest,

$$f_k = \sum_{j=1}^{p} j G_{(p+1)(\ell)+j,k}/p,$$

and $f = \sum_{k=1}^{K} q_k f_k$ (Martins et al., 2016). For a particular locus, the formula for F_{ST}^{Q} corresponds to the proportion of the genetic variation (or variance) in ancestral allele frequency that can be explained by latent population structure

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

where σ_T^2 is the total variance and σ_S^2 is the error variance (Weir, 1996). Following ANOVA theory, the F_{ST}^{Q} statistics were used to perform statistical tests of neutrality at each locus, by comparing the observed values to their expectations from the genome-wide background. The test was based on the squared z-score statistic, $z^2 = (n-K)F_{\mathrm{ST}}^{Q}/(1-F_{\mathrm{ST}}^{Q})$, for which a chi-squared distribution with $K-1$ degrees of freedom was assumed under the null-hypothesis. To avoid an increased number of false positive tests, we adopted an empirical null-hypothesis testing approach that recalibrates the null-hypothesis for the background levels of

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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 expected at selectively neutral SNPs (Efron, 2004). The calibration of the null-hypothesis was achieved by using genomic control to adjust the test statistics (Devlin et al., 1999; François et al., 2016). After recalibration of the null-hypothesis, the control of the false discovery rate was achieved by using the Benjamini-Hochberg algorithm (Benjamini et al.).

`R` *package.* We implemented the AQP and APLS algorithms and improved graphical tools in the `R` package `tess3r`, available from Github and submitted to the Comprehensive R Archive Network (R Core Team, 2016).

3. Simulated and real data sets.

Coalescent simulations. We used the computer program `ms` to perform coalescent simulations of neutral and outlier SNPs under spatial models of admixture (Hudson, 2002). Two ancestral populations were created from the simulation of Wright's two-island models. The simulated data sets contained admixed genotypes for n individuals for which the admixture proportions varied continuously along a longitudinal gradient (Durand et al., 2009; François et al., 2010). In those scenarios, individuals at each extreme of the geographic range were representative of their population of origin, while individuals at the center of the range shared intermediate levels of ancestry in the two ancestral populations (Caye et al., 2016). For those simulations, the Q matrix, $\mathbf{Q}_0$, was entirely described by the location of the sampled individuals.

Neutrally evolving ancestral chromosomal segments were generated by simulating DNA sequences with an effective population size $N_0 = 10^6$ for each ancestral population. The mutation rate per bp and generation was set to $\mu = 0.25 \times 10^{-7}$, the recombination rate per generation was set to $r = 0.25 \times 10^{-8}$, and the parameter m was set to obtain neutral levels of F_{ST} ranging between values of 0.005 and 0.10. The number of base pairs for each DNA sequence was varied between 10k to 300k to obtain numbers of polymorphic loci ranging between 1k and 200k after filtering out SNPs with minor allele frequency lower than 5%. To create SNPs with values in the tail of the empirical distribution of F_{ST}, additional ancestral chromosomal segments were generated by simulating DNA sequences with a migration rate m_s lower than m. The simulations reproduced the reduced levels of diversity and the in-

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

creased levels of differentiation expected under hard selective sweeps occurring at one particular chromosomal segment in ancestral populations (Martins et al., 2016). For each simulation, the sample size was varied in the range $n = 50$-700.

We compared the AQP and APLS algorithm estimates with those obtained with the `tess3` algorithm. Each program was run 5 times on the same simulated data. Using $K = 2$ ancestral populations, we computed the root mean squared error (RMSE) between the estimated and known values of the Q-matrix, and between the estimated and known values of the G-matrix. To evaluate the benefit of spatial algorithms, we compared the statistical errors of APLS algorithms to the errors obtained with the `snmf` method that reproduces the outputs of the `structure` program accurately (Frichot et al., 2014) (Frichot et al., 2015). To quantify the performances of neutrality tests as a function of ancestral and observed levels of F_{ST}, we used the area under the precision-recall curve (AUC) for several values of the selection rate. Subsamples from a real data set were used to perform a runtime analysis of the AQP and APLS algorithms (*A. thaliana* data, see below). Runtimes were evaluated by using a single computer processor unit Intel Xeon 2.0 GHz.

Application to human data. To evaluate the robustness of our approach to a situation where admixture was the consequence of large displacement rather than contact between proximal populations, we studied the case of African-American populations. This is an interesting case for which the incorporation of geographic data could potentially bias estimation of ancestry coefficients. Genotypes with minor allele frequency greater than 5% were obtained from a public release of the 1000 Genomes project phase 3 for African Americans (ASW, 61 individuals), Africans (YRI from Nigeria and LWK from Kenya, 207 individuals), and Europeans (GBR from the United Kingdom and TSI from Italy, 198 individuals) (1000 Genomes Project Consortium, 2015). A total of 6,994,677 SNPs were analyzed with geographic data corresponding to the country of origin of individual samples. We compared the estimates from the APLS algorithm applied with its default parameter settings to the results of the `snmf` program that do not make use of geographic information.

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

Application to European ecotypes of Arabidopsis thaliana. We used the APLS algorithm to survey spatial population genetic structure and to investigate the molecular basis of adaptation by considering 214k SNPs from 1,095 European ecotypes of the plant species *A. thaliana* (Horton et al., 2012). The cross-validation criterion was used to evaluate the number of clusters in the sample, and a statistical analysis was performed to evaluate the range of the variogram from the data. We used `R` functions of the `tess3r` package to display interpolated admixture coefficients on a geographic map of Europe (R Core team 2016). A gene ontology enrichment analysis using the software AMIGO (Carbon et al., 2009) was performed in order to evaluate which molecular functions and biological processes might be involved in local adaptation of *A. thaliana* in Europe.

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

Statistical errors. We used coalescent simulations of neutral polymorphisms under spatial models of admixture to compare the statistical errors of the AQP and APLS estimates with those of the `tess3` algorithm (Caye et al., 2016). The ground truth for the Q-matrix ($\mathbf{Q}_0$) was computed from the mathematical model for admixture proportions used to generate the data. For the G-matrix, the ground truth matrix ($\mathbf{G}_0$) was computed from the empirical genotype frequencies in the two population samples before an admixture event. The root mean squared errors (RMSE) for the $\mathbf{Q}$ and $\mathbf{G}$ estimates decreased as the sample size and the number of loci increased (Figure 1). For all algorithms, the statistical errors were generally small when the number of loci was greater than 10k SNPs. Those results provided evidence that the three algorithms produced equivalent estimates of the matrices $\mathbf{Q}_0$ and $\mathbf{G}_0$. The results also provided a check that the APLS and `tess3` algorithms converged to the same estimates as those obtained after the application of the AQP algorithm, which is guaranteed to converge mathematically.

The benefit of including spatial information in algorithms. Using neutral coalescent simulations of spatial admixture, we compared the statistical estimates obtained from the spatial algorithm APLS and the non-spatial algorithm `snmf` (Frichot et al., 2014). For various levels of ancestral population differentiation, estimates obtained from the spatial algorithm were more accurate than for those obtained using non-spatial approaches (Figure 2). For the larger samples, much finer population structure was detected with the spatial method than with the non-spatial algorithm (Figure 2).

In simulations of outlier loci, we used the area under the precision-recall curve (AUC) for quantifying the performances of tests based on the estimates of ancestry matrices, $\mathbf{Q}$ and $\mathbf{G}$. In addition, we computed AUCs for F_{ST}-based neutrality tests using truly ancestral genotypes. As they represented the maximum reachable values, AUCs based on truly ancestral genotypes were always higher than those obtained for tests based on reconstructed matrices. For all values of the relative selection intensity, AUCs were higher for spatial methods than for non-spatial methods (Figure 4, the relative selection intensity is the ratio of migration rates at neutral and adaptive loci). For high selection intensities, the performances of tests based on estimates of ancestry matrices were

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

close to the optimal values reached by tests based on true ancestral frequencies. These results provided evidence that including spatial information in ancestry estimation algorithms improves the detection of signatures of hard selective sweeps having occurred in unknown ancestral populations.

Sensitivity of estimates to spatial measurements. Next, we used the simulated data sets to evaluate the robustness of APLS estimates to inaccurate measurements of spatial coordinates. To this aim, Gaussian noise was added to truly observed geographic coordinates by considering values of the noise-to-signal ratio ranging from 0 to 3. We computed variograms in all cases, and found that the spatial signal was removed from simulations for noise-to-signal ratios greater than two, while the signal was still observable with a noise-to-signal ratio lower than one. For all simulations, we compared the relative error of APLS Q-matrix estimates to those obtained from an non-spatial method (`snmf`). For small levels of uncertainty in spatial coordinates the errors of APLS estimates were lower than those of `snmf` (Figure 3). For simulations with $n = 500$ individuals and $L = 10^5$ loci, a larger noise-to-signal ratio increased statistical errors in the Q-matrix estimates from the APLS algorithm. For smaller noise-to-signal ratios, RMSEs remained generally lower for the APLS algorithm than for methods without spatial coordinates. For simulations with $n = 50$ individuals and $L = 10^4$ loci, the APLS estimates were more accurate than the non-spatial estimates. This unexpected result could be explained by subtle algorithmic differences in tested programs. To a large extent, estimates from the APLS algorithm were robust to uncertainty in spatial measurements. Standard graphical tests such as a variogram analysis can help deciding whether our spatially explicit algorithm is useful or not.

Runtime and convergence analyses. We subsampled a large SNP data set for *A. thaliana* ecotypes to compare the convergence properties and runtimes of the `tess3`, AQP, and APLS algorithms. In those experiments, we used $K = 6$ ancestral populations, and replicated 5 runs for each simulation. For $n = 100 - 600$ individuals ($L = 50$k SNPs), the APLS algorithm required more iterations (25 iterations) than the AQP algorithm (20 iterations) to converge to its solution (Figure 5). This was less than for `tess3` (30 iterations). For $L = 10 - 200$k SNPs ($n = 150$ individuals), similar results were observed. For 50k SNPs, the runtimes

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

were significantly lower for the APLS algorithm than for the `tess3` and AQP algorithms. For $L = 50$k SNPs and $n = 600$ individuals, it took on average 1.0 min for the APLS and 100 min for the AQP algorithm to compute ancestry estimates. For `tess3`, the runtime was on average 66 min. For $L = 100$k SNPs and $n = 150$ individuals, it took on average 0.6 min (9.0 min) for the APLS (AQP) algorithm to compute ancestry estimates. For `tess3`, the runtime was on average 1.3 min. For those values of n and L, the APLS algorithm implementation ran about 2 to 100 times faster than the other algorithm implementations.

Human data analysis. To evaluate a case of model misspecification, we analyzed data from the 1000 Genomes project for African Americans, Africans from Nigeria and from Kenya, and Europeans from the United Kingdom and from Italy. Using the default values for the hyperparameters, the Laplacian matrix was a block diagonal matrix where each block corresponded to one of the five populations. The spatial variogram exhibited a flat shape. For $K = 2$, the APLS estimates for the African American population were equal to 24.2% for European ancestors and 75.8% for African ancestors. The corresponding `snmf` estimates were equal to 22.4% for European ancestors and 77.6% for African ancestors. For $K = 3$, the APLS estimates for the African American population were equal to 21.4% for European ancestors, 51.8% for West African ancestors and 26.8% for East African ancestors. The corresponding `snmf` estimates were equal to 22.2% for European ancestors, 68.4% for West African ancestors and 9.4% for East African ancestors. Overall, the results obtained with our spatial method for African Americans were similar to those obtained with `snmf`. The main difference between APLS and `snmf` estimates were for African populations. For Africans, `snmf` detected two distinct genetic clusters whereas APLS detected a larger proportion of shared ancestry between Eastern and Western populations.

Application to European ecotypes of Arabidopsis thaliana. We used the APLS algorithm to survey spatial population genetic structure and perform a genome scan for adaptive alleles in European ecotypes of the plant species *A. thaliana*. The cross validation criterion decreased rapidly from $K = 1$ to $K = 3$ clusters, indicating that there were three main ancestral groups in Europe, corresponding to geographic regions in Western Europe, Eastern and Central Europe and Northern Scan-

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

dinavia. For K greater than four, the values of the cross validation criterion decreased in a slower way, indicating that subtle substructure resulting from complex historical isolation-by-distance processes could also be detected (Figure 6). The spatial analysis provided an approximate range of $\sigma = 150$km for the spatial variogram (Figure 6). Figure 7 displays the Q-matrix estimate interpolated on a geographic map of Europe for $K = 6$ ancestral groups. The estimated admixture coefficients provided clear evidence for the clustering of the ecotypes in spatially homogeneous genetic groups.

Targets of selection in A. thaliana genomes. Tests based on the F_{ST}^{Q} statistic were applied to the 241k SNP data set to reveal new targets of natural selection in the *A. thaliana* genome. *A. thaliana* occurs in a broad variety of habitats, and local adaptation to the environment is acknowledged to be important in shaping its genetic diversity through space (Hancock et al., 2011; Fournier-Level et al., 2011). The APLS algorithm was run on the 1,095 European lines of *A. thaliana* with $K = 6$ ancestral populations and $\sigma = 1.5$ for the range parameter. Using the Benjamini-Hochberg algorithm to control the FDR at the level 1%, the program produced a list of 12,701 candidate SNPs, including linked loci and representing 3% of the total number of loci. The top 100 candidates included SNPs in the flowering-related genes SHORT VEGETATIVE PHASE (SVP), COP1-interacting protein 4.1 (CIP4.1) and FRIGIDA (FRI) (p-values $< 10^{-300}$). These genes were detected by previous scans for selection on this dataset (Horton et al., 2012). We performed a gene ontology enrichment analysis using AmiGO in order to evaluate which biological functions might be involved in local adaptation in Europe. We found a significant over-representation of genes involved in cellular processes (fold enrichment of 1.06, p-value equal to 0.0215 after Bonferonni correction).

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

5. Discussion. Including geographic information on sample locations in the inference of ancestral relationships among organisms is a major objective of population genetic studies (Malécot, 1948; Cavalli et al., 1994; Epperson, 2003). Assuming that geographically close individuals are more likely to share ancestry than individuals at distant sites, we introduced two new algorithms for estimating ancestry proportions using geographic information. Based on least-squares problems, the new algorithms combine matrix factorization approaches and spatial statistics to provide accurate estimates of individual ancestry coefficients and ancestral genotype frequencies. The two methods share many similarities, but they differ in the approximations they make in order to decrease algorithmic complexity. More specifically, the AQP algorithm was based on quadratic programming, whereas the APLS algorithm was based on the spectral decomposition of the Laplacian matrix. The algorithmic complexity of APLS algorithm grows linearly with the number of individuals in the sample while the method has the same statistical accuracy as more complex algorithms.

To measure the benefit of using spatial algorithms, we compared the statistical errors observed for spatial algorithms with those observed for non-spatial algorithms. The errors of spatial methods were lower than those observed with non-spatial methods, and spatial algorithms allowed the detection of more subtle population structure. In addition, we implemented neutrality tests based on the spatial estimates of the Q and G-matrices (Martins et al., 2016), and we observed that those tests had higher power to reject neutrality than those based on non-spatial approaches. Thus spatial information helped improving the detection of signatures of selective sweeps having occurred in ancestral populations prior to admixture events. We applied the neutrality tests to perform a genome scan for selection in European ecotypes of the plant species *A. thaliana*. The genome scan confirmed the evidence for selection at flowering-related genes *CIP4.1*, *FRI* and *DOG1* differentiating Fennoscandia from North-West Europe (Horton et al., 2012).

Estimation of ancestry coefficients using fast algorithms that extend non-spatial approaches – such as `structure` – has been intensively discussed during the last years (Wollstein et al., 2015). In these improvements, spatial approaches have received less attention than non-spatial approaches. In this study, we have proposed a conceptual framework for developing fast spatial ancestry estimation methods, and a suite

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

of computer programs implements this framework in the `R` program `tess3r`. Our package provides an integrated pipeline for estimating and visualizing population genetic structure, and for scanning genomes for signature of local adaptation. The algorithmic complexity of our algorithms allows their users to analyze samples including hundreds to thousands of individuals. For example, analyzing more than one thousand *A. thaliana* genotypes, each including more than 210k SNPs, took only a few minutes using a single CPU. In addition, the algorithms have multithreaded versions that run on parallel computers by using multiple CPUs. The multithreaded algorithm, which is available from the `R` program, allows using our programs in large-scale genomic sequencing projects.

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

APPENDIX A: ALGORITHMS

ALGORITHM A.1. AQP algorithm pseudo code. To solve optimization problem (2.1).

Input: the data matrix $\mathbf{Y} \in \{0,1\}^{n\times(p+1)L}$, the Laplacian matrix $\mathbf{\Lambda} \in \mathbb{R}^{n\times n}$, the number of ancestral populations K, the regularization coefficient α, the maximum number of iteration $itMax$

Output: the admixture matrix $\mathbf{Q} \in \mathbb{R}^{n\times K}$, the ancestral genotype frequency matrix $\mathbf{G} \in \mathbb{R}^{K\times(p+1)L}$

Initialize $\mathbf{Q}$ at random;
for $it = 1..itMax$ **do**
  // G optimization step
  for $l = 1..L$ **do**
    $Y^l \leftarrow \mathbf{Y}_{.,(p+1)l..(p+1)l+d}$;
    $\mathbf{D}_Q \leftarrow \mathbf{I}_{p+1} \otimes \mathbf{Q}^T\mathbf{Q}$;
    $v_Q \leftarrow Vec(\mathbf{Q}^T Y^l)$;
    $g^\star \in \arg\min_{g\in\Delta_G} -2v_Q^T g + g^T\mathbf{D}_Q g$;
    $Vec(\mathbf{G}_{(p+1)l..(p+1)l+d,.}) \leftarrow g^\star$;
  end
  // Q optimization step
  $\mathbf{D}_G \leftarrow Id_n \otimes \mathbf{G}^T\mathbf{G} + \alpha\mathbf{\Lambda} \otimes \mathbf{I}_K$;
  $v_G \leftarrow Vec(\mathbf{G}^T\mathbf{Y}^T)$;
  $Vec(\mathbf{Q}^T) \in \arg\min_{q\in\Delta_Q} -2v_G^T q + q^T\mathbf{D}_G q$;
end

ALGORITHM A.2. APLS algorithm pseudo code. To solve the op-

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

timization problem (2.1).

Input: the data matrix $\mathbf{Y} \in \{0,1\}^{n\times(d+1)L}$, the eigenvalues matrix $\mathbf{\Delta}$ and eigenvectors matrices $\mathbf{U}$ such that $\mathbf{\Lambda} = \mathbf{U}^T\mathbf{\Delta}\mathbf{U}$, the number of ancestral populations K, the regularization coefficient α, the maximum number of iteration $itMax$

Output: the admixture matrix $\mathbf{Q} \in \mathbb{R}^{n\times K}$, the ancestral genotype frequency matrix $\mathbf{G} \in \mathbb{R}^{K\times(d+1)L}$

Initialize $\mathbf{Q}$ at random;
$proj(\mathbf{Y}) \leftarrow \mathbf{RY}$;
for $it = 1..itMax$ **do**
  // G optimization step
  for $j = 1..(p+1)L$ **do**
    $g^\star \in \arg\min_{g\in\mathbb{R}^K} ||\mathbf{Y}_{.,j} - \mathbf{Q}g||^2$;
    $\mathbf{G}_{j,.} \leftarrow g^\star$;
  end
  Project $\mathbf{G}$ such that $\mathbf{G} \in \Delta_G$;
  // Q optimization step
  for $i = 1..n$ **do**
    $g_i^\star \in \arg\min_{q\in\mathbb{R}^K} ||proj(\mathbf{Y})_{i,.} - \mathbf{G}^T q||^2 + \alpha\mathbf{\Delta}_{i,i}||q||^2$;
    $proj(\mathbf{Q})_{i,.} \leftarrow g_i^\star$;
  end
  $\mathbf{Q} \leftarrow \mathbf{U}^T proj(\mathbf{Q})$;
  Project $\mathbf{Q}$ such that $\mathbf{Q} \in \Delta_Q$;
end

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

ACKNOWLEDGEMENTS

The authors are grateful to Edoardo M. Airoldi and three anonymous reviewers for their constructive comments. This work has been partially supported by the LabEx PERSYVAL-Lab (ANR-11-LABX-0025-01) funded by the French program Investissement d'Avenir. Olivier François acknowledges support from Grenoble INP and from the Agence Nationale de la Recherche, project AFRICROP ANR-13-BSV7-0017.

REFERENCES

1000 Genomes Project Consortium, T. (2015). A global reference for human genetic variation. *Nature* **526** 68–74.

Alexander, D. H. and Lange, K. (2011). Enhancements to the ADMIXTURE algorithm for individual ancestry estimation. *BMC bioinformatics* **12** 246.

Baran, Y., Quintela, I., Carracedo, Á., Pasaniuc, B. and Halperin, E. (2013). Enhanced localization of genetic samples through linkage-disequilibrium correction. *American Journal of Human Genetics* **92** 882–894.

Belkin, M. and Niyogi, P. Laplacian eigenmaps for eimensionality reduction and data representation. *Neural Computation* **6** 1373–1396.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* **1** 289–300.

Bertsekas, D. P. (1995). *Nonlinear Programming*. Athena Scientific, Nashua, USA.

Bhaskar, A., Javanmard, A., Courtade, T. A. and Tse, D. (2017). Novel probabilistic models of spatial genetic ancestry with applications to stratification correction in genome-wide association studies. *Bioinformatics* **33** 879-885.

Bradburd, G. S., Ralph, P. L. and Coop, G. M. (2016). A spatial framework for understanding population structure and admixture. *PLoS Genetics* **12** e1005703.

Cai, D., He, X., Han, J. and Huang, T. S. (2011). Graph regularized nonnegative matrix factorization for data representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33** 1548–1560.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., AmiGO Hub and Web Presence Working Group (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics* **25** 288–9.

Cavalli, L. L., Menozzi, P. and Piazza, A. (1994). *The History and Geography of Human Genes*. Princeton University Press, Princeton, USA.

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

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

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

CICHOCKI, A., ZDUNEK, R., PHAN, A. H. and AMARI, S. I. (2009). *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-Way Data Analysis and Blind Source Separation*. John Wiley & Sons, Ltd, Chichester, UK.

CORANDER, J., SIRÉN, J. and ARJAS, E. (2008). Bayesian spatial modeling of genetic population structure. *Computational Statistics* **23** 111–129.

CRESSIE, N. A. C. (1993). *Statistics for Spatial Data. Wiley Series in Probability and Statistics*. John Wiley & Sons, Inc., Hoboken, NJ, USA.

DEVLIN, B. and ROEDER, K. (1999). Genomic control for association studies. *Biometrics* **55** 997–1004.

DURAND, E., JAY, F., GAGGIOTTI, O. E. and FRANÇOIS, O. (2009). Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution* **26** 1963–1973.

EASTMENT, H. and KRZANOWSKI, W. (1982). Cross-validatory choice of the number of components from a principal component analysis. *Technometrics* **24** 73–77.

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

ENGELHARDT, B. E. and STEPHENS, M. (2010). Analysis of population structure: A unifying framework and novel methods based on sparse factor analysis. *PLoS Genetics* **6** e1001117.

EPPERSON, B. K. (2003). *Geographical Genetics*. Princeton University Press, Princeton, USA.

EPPERSON, B. K. and LI, T. (1996). Measurement of genetic structure within populations using Moran's spatial autocorrelation statistics. *Proceedings of the National Academy of Sciences* **93** 10528–10532.

FOURNIER-LEVEL, A., KORTE, A., COOPER, M. D., NORDBORG, M., SCHMITT, J. and WILCZEK, A. M. (2011). A map of local adaptation in Arabidopsis thaliana. *Science* **334** 86–89.

FRANÇOIS, O. and DURAND, E. (2010). Spatially explicit Bayesian clustering models in population genetics. *Molecular Ecology Resources* **10** 773–784.

FRANÇOIS, O. and WAITS, L. P. (2016). *Clustering and assignment methods in landscape genetics* 114–128. John Wiley & Sons, Ltd, Chichester, UK.

FRANÇOIS, O., MARTINS, H., CAYE, K. and SCHOVILLE, S. D. (2016). Controlling false discoveries in genome scans for selection. *Molecular Ecology* **25** 454–469.

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

FRICHOT, E., MATHIEU, F., TROUILLON, T., BOUCHARD, G. and FRANÇOIS, O. (2014). Fast and efficient estimation of individual ancestry coefficients. *Genetics* **196** 973–983.

GRIPPO, L. and SCIANDRONE, M. (2000). On the convergence of the block nonlinear Gauss-Seidel method under convex constraints. *Operations Research Letters* **26** 127–136.

HANCOCK, A. M., BRACHI, B., FAURE, N., HORTON, M. W., JARYMOWYCZ, L. B., SPERONE, F. G., TOOMAJIAN, C., ROUX, F. and BERGELSON, J. (2011). Adaptation to climate across the Arabidopsis thaliana genome.

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

Science **334** 83–86.

HARDY, O. J. and VEKEMANS, X. (1999). Isolation by distance in a continuous population: reconciliation between spatial autocorrelation analysis and population genetics models. *Heredity* **83** 145–154.

HORTON, M. W., HANCOCK, A. M., HUANG, Y. S., TOOMAJIAN, C., ATWELL, S., AUTON, A., MULIYATI, N. W., PLATT, A., SPERONE, F. G., VILHJÁLMSSON, B. J., NORDBORG, M., BOREVITZ, J. O. and BERGELSON, J. (2012). Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature Genetics* **44** 212–216.

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

KIM, J. and PARK, H. (2011). Fast Nonnegative Matrix Factorization: an Active-Set-Like Method and Comparisons. *SIAM Journal on Scientific Computing* **33** 3261–3281.

KORNELIUSSEN, T. S., ALBRECHTSEN, A. and NIELSEN, R. (2014). ANGSD: Analysis of next generation sequencing data. *BMC bioinformatics* **15** 356.

LAO, O., LIU, F., WOLLSTEIN, A. and KAYSER, M. (2014). GAGA: A new algorithm for genomic inference of geographic ancestry reveals fine level population substructure in Europeans. *PLoS Computational Biology* **10** e1003480.

LEE, D. D. and SEUNG, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Nature* **401** 788–791.

MALÉCOT, G. (1948). *Les Mathématiques de l'Hérédité*. Masson et Cie, Paris, France.

MANTEL, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer research* **27** 209–220.

MARTINS, H., CAYE, K., LUU, K., BLUM, M. G. B. and FRANÇOIS, O. (2016). Identifying outlier loci in admixed and in continuous populations using ancestral population differentiation statistics. *Molecular Ecology* **25** 5029–5042.

POPESCU, A. A., HARPER, A. L., TRICK, M., BANCROFT, I. and HUBER, K. T. (2014). A novel and fast approach for population structure inference using Kernel-PCA and optimization. *Genetics* **198** 1421–1431.

PRITCHARD, J. K., STEPHENS, M. and DONNELLY, P. (2000). Inference of population structure using multilocus genotype data. *Genetics* **155** 945–959.

RAÑOLA, J. M., NOVEMBRE, J. and LANGE, K. (2014). Fast spatial ancestry via flexible allele frequency surfaces. *Bioinformatics* **30** 2915–2922.

RAJ, A., STEPHENS, M. and PRITCHARD, J. K. (2014). FastSTRUCTURE: Variational inference of population structure in large SNP data sets. *Genetics* **197** 573–589.

SCHRAIBER, J. G. and AKEY, J. M. (2015). Methods and models for unravelling human evolutionary history. *Nature Reviews Genetics* **16** 727–740.

SEGELBACHER, G., CUSHMAN, S. A., EPPERSON, B. K., FORTIN, M. J., FRANCOIS, O., HARDY, O. J., HOLDEREGGER, R., TABERLET, P., WAITS, L. P. and MANEL, S. (2010). Applications of landscape genetics in conservation biology: Concepts and challenges. *Conservation Genetics* **11** 375–385.

TANG, H., PENG, J., WANG, P. and RISCH, N. J. (2005). Estimation of individual

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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: analytical and study design considerations. *Genetic Epidemiology* **28** 289–301.

WANG, J. (2017). The computer program structure for assigning individuals to populations: easy to use but easier to misuse. *Molecular Ecology Resources* in press.

WEIR, B. S. (1996). *Genetic data analysis II: methods for discrete population genetic data* **Vol.2**. Sinauer Associates, Sunderland, MA, USA.

WOLD, S. (1978). Cross-validatory estimation of the number of components in factor and principal components models. *Technometrics* **20** 397–405.

WOLLSTEIN, A. and LAO, O. (2015). Detecting individual ancestry in the human genome. *Investigative genetics* **6** 7.

WRIGHT, S. (1943). Isolation by Distance. *Genetics* **28** 114–138.

YANG, W.-Y., PLATT, A., CHIANG, C. W.-K., ESKIN, E., NOVEMBRE, J. and PASANIUC, B. (2014). Spatial localization of recent ancestors for admixed individuals. *Genes, Genomes, Genetics* **4** 2505–2518.

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

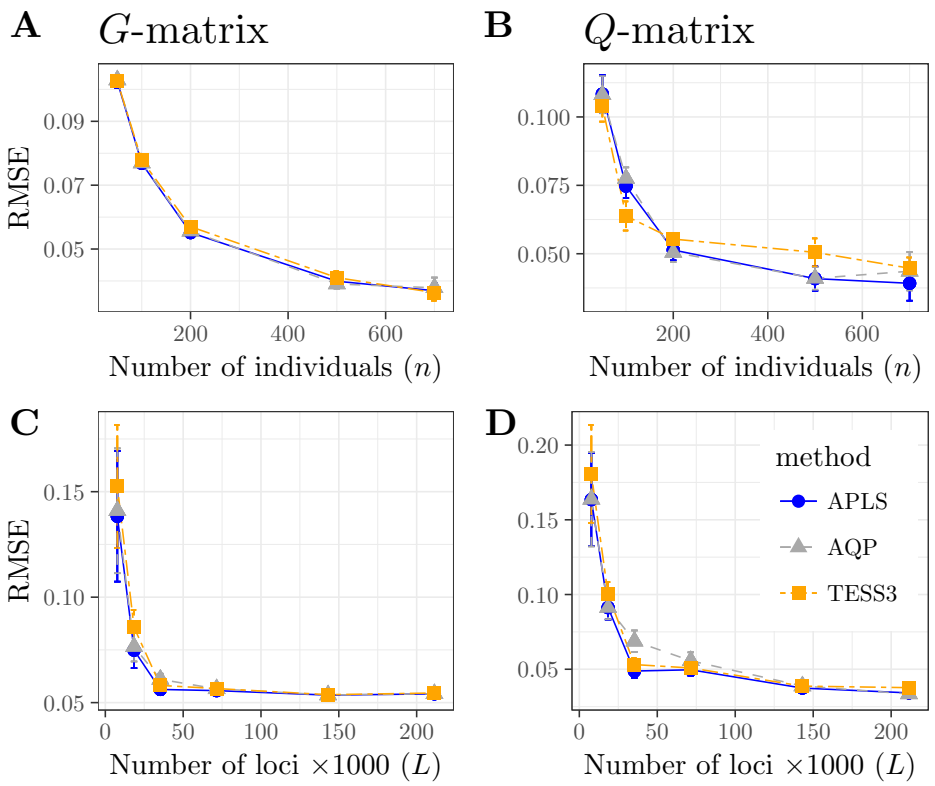

FIG 1. **Root Mean Squared Errors (RMSEs) for the G and Q matrix estimates**. *Simulations of spatially admixed populations. A-B) Statistical errors for APLS, AQP and* `tess3` *estimates as a function of the sample size, n ($L \sim 10^4$). C-D) Statistical errors for APLS, AQP and* `tess3` *estimates as a function of the number of loci, L ($n = 200$).*

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

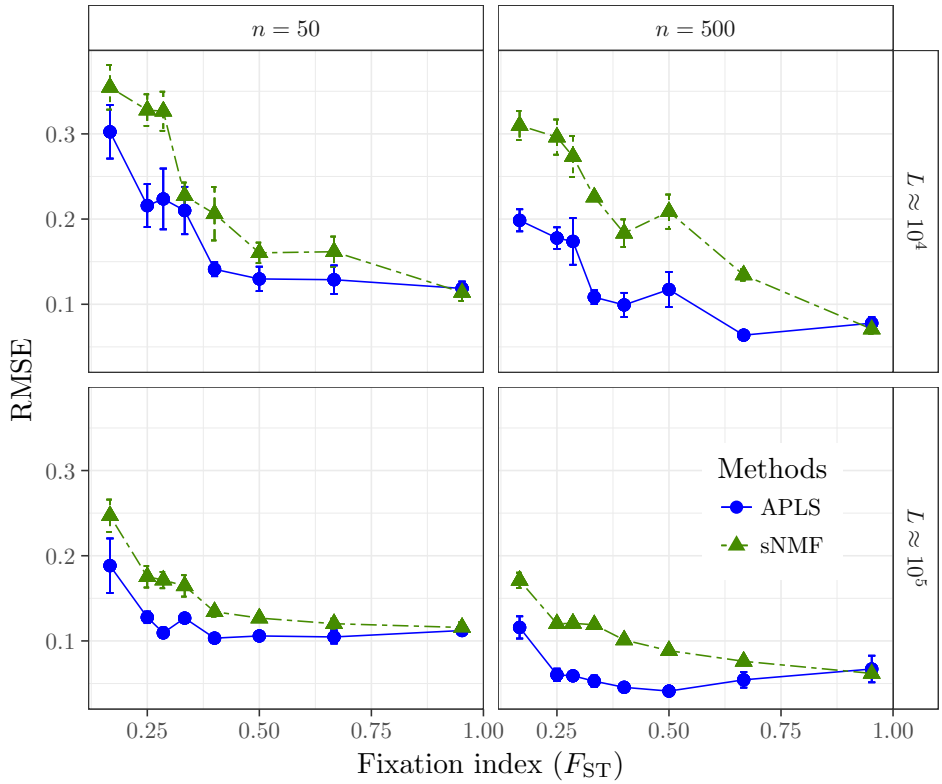

FIG 2. **Root Mean Squared Errors (RMSEs) for the Q estimates.** *Simulations of spatially admixed populations for several values of fixation index (F_{ST}) between ancestral populations. Ancestral populations are simulated with Wright's two-island models and the fixation index is defined as $1/(1+4N_0m)$ where m is the migration rate and N_0 the effective population size. The statistical errors for sNMF and APLS are represented as a function of F_{ST}.*

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

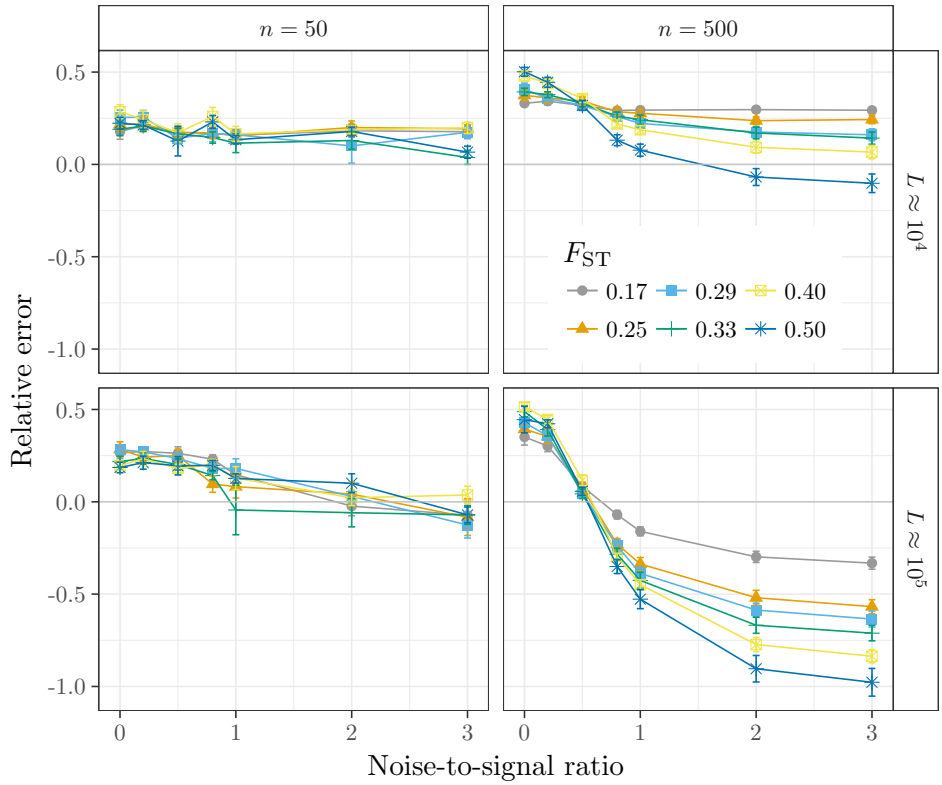

FIG 3. **Impact of uncertainty in geographic coordinates on ancestry estimates.** *Relative statistical error of ancestry estimates obtained from the APLS algorithm for several levels of the noise-to-signal ratio and values of the fixation index. The* `snmf` *algorithm was considered as the reference for the non-spatial method (value 0).*

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

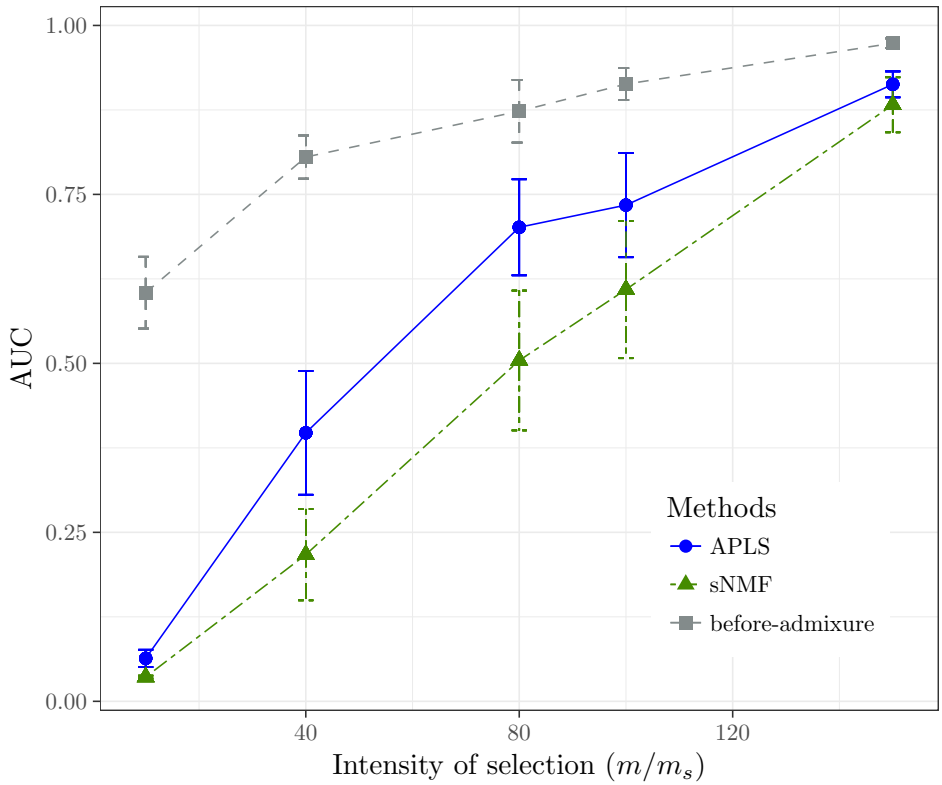

FIG 4. **Area under the precision-recall curve (AUC)**. *Neutrality tests applied to simulations of spatially admixed populations. AUCs for tests based on F_{ST} with the true ancestral populations, spatial ancestry estimates computed with APLS algorithms, non-spatial (*`structure`*-like) ancestry estimates computed with the* `snmf` *algorithm. The relative intensity of selection in ancestral populations, defined as the ratio m/m_s, was varied in the range $1-160$.*

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

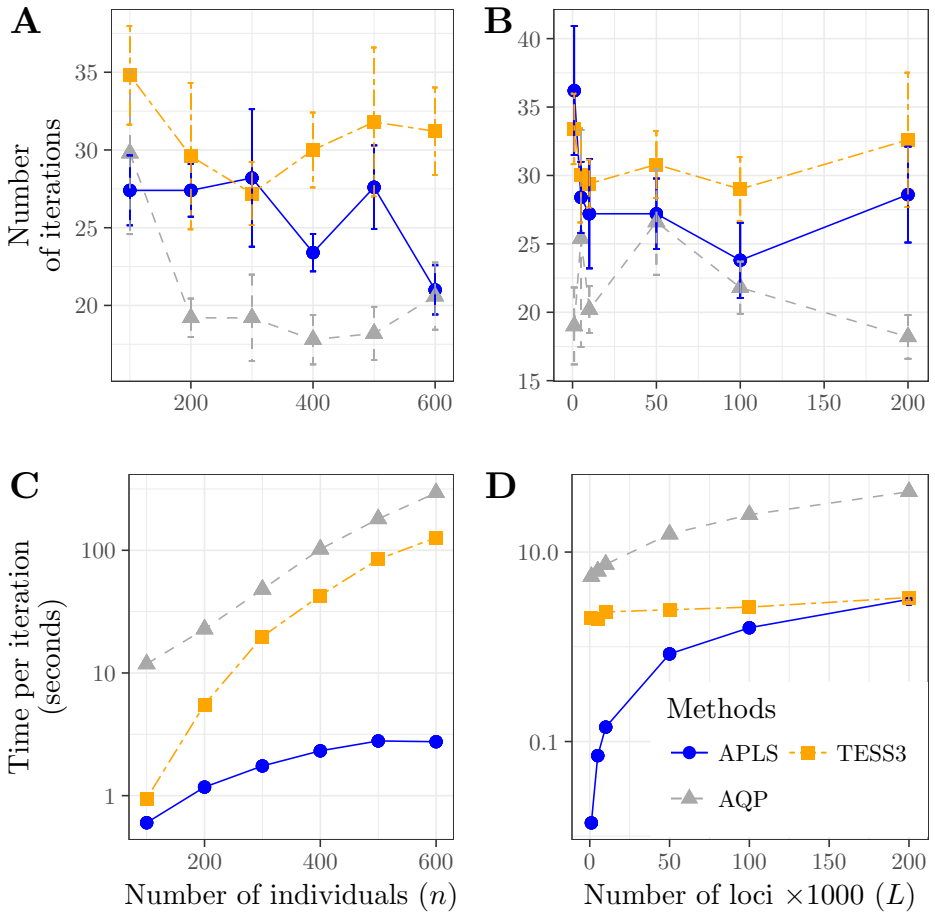

FIG 5. **Number of iterations and runtimes for the AQP, APLS and `tess3` algorithm implementations**. *A-B) Total number of iterations before an algorithm reached a steady solution. C-D) Runtime for a single iteration (seconds). The number of SNPs was kept fixed to $L = 50k$ in A and C. The number of individuals was kept fixed to $n = 150$ in B and D.*

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

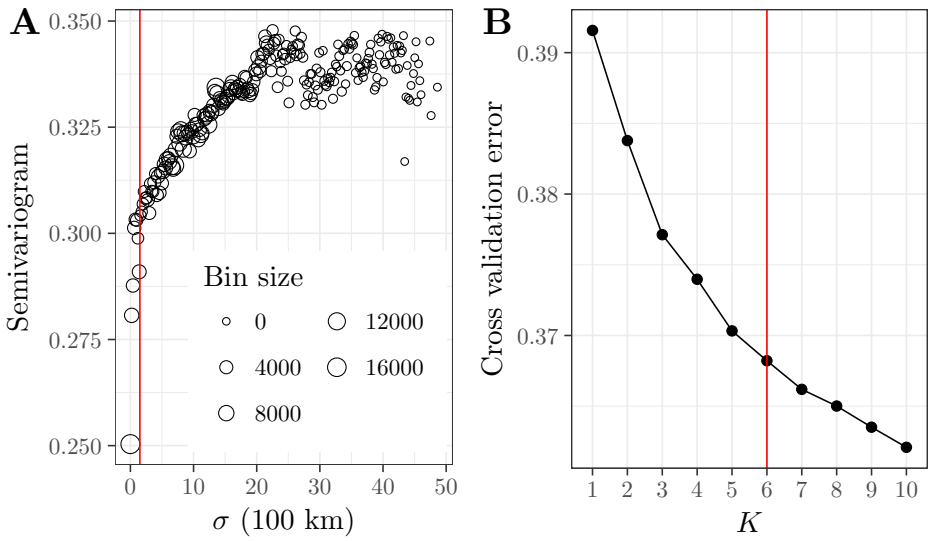

FIG 6. **Choice of σ and K for the APLS algorithm**. *A) Empirical variogram for the A. thaliana data. The red vertical line shows the range value $\sigma = 1.5$. B) Cross validation error as function of the number of ancestral populations, K. The red vertical line shows the number of ancestral populations $K = 6$.*

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

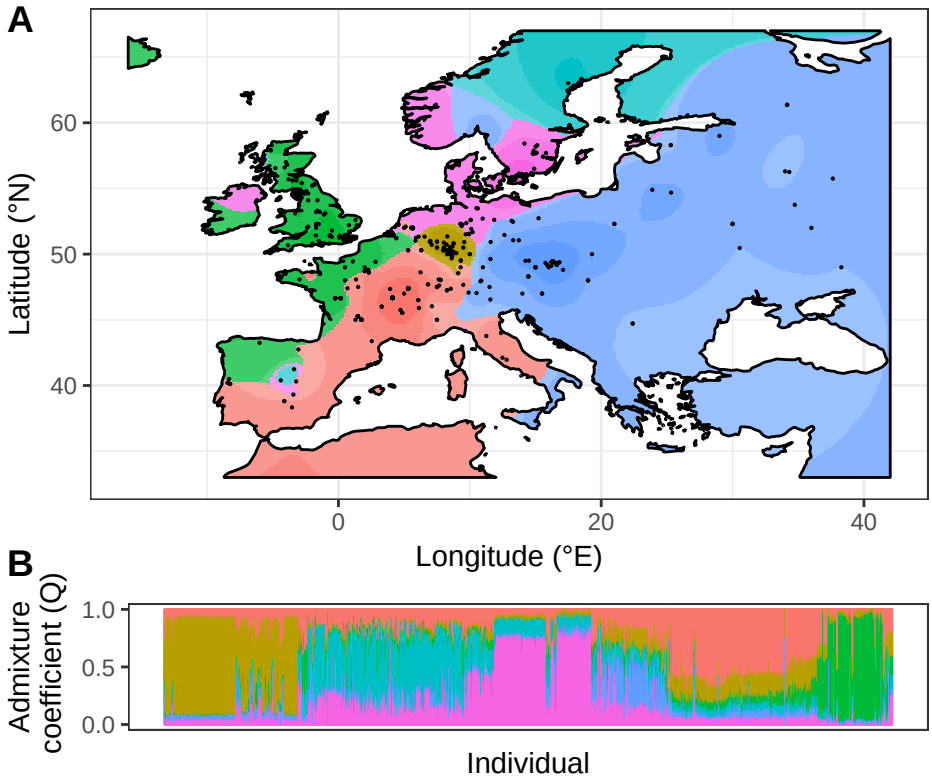

FIG 7. *A. thaliana* **ancestry coeficients**. *Ancestry coefficient estimates computed by the APLS algorithm with* $K = 6$ *ancestral populations and* $\sigma = 1.5$ *for the range parameter. A) Geographic map of ancestry coefficients. B) Barplot of ancestry coefficients.*

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

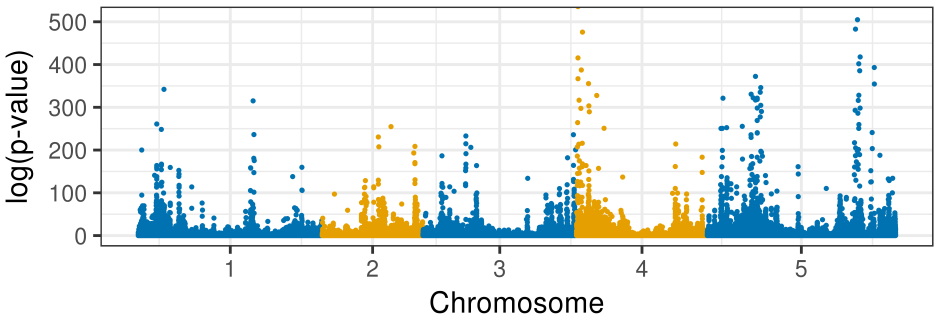

FIG 8. **Local adaptation in European lines of** *A. thaliana* . *Manhattan plot of* $-\log(p\text{-value})$. p-value were computed from population structure estimated by the APLS algorithm with $K = 6$ ancestral populations and $\sigma = 1.5$ for the range parameter.

bioRxiv preprint doi: https://doi.org/10.1101/080291; this version posted September 27, 2017. 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.

Kevin Caye and Olivier François
Université Grenoble-Alpes
Centre National de la Recherche Scientifique
TIMC-IMAG UMR 5525
Grenoble, 38042, France
E-mail: kevin.caye@imag.fr
olivier.francois@imag.fr

Flora Jay
Universit Paris-Sud
Universit Paris-Saclay
Laboratoire de Recherche en Informatique UMR 7206
CNRS UMR 8623
Orsay, 91400, France
E-mail: flora.jay@lri.fr

Olivier Michel
Université Grenoble-Alpes
Centre National de la Recherche Scientifique
GIPSA-lab UMR 5216
Grenoble, 38042, France
E-mail: olivier.michel@gipsa-lab.grenoble-inp.fr