Genetics: Early Online, published on February 4, 2014 as 10.1534/genetics.113.160572

Fast and efficient estimation of individual ancestry coefficients

Eric Frichot,[1] François Mathieu,[1] Théo Trouillon,[1,2]
Guillaume Bouchard,[2] Olivier François[1]

[1] Université Joseph Fourier Grenoble 1, Centre National de la Recherche Scientifique,

TIMC-IMAG UMR 5525, Grenoble, 38042, France.

[2] Xerox Research Center Europe, Meylan, F38240, France.

Running Title: Super Fast Inference of Population Structure

KeyWords: Inference of Population Structure, Ancestry Coefficients, Non-negative Matrix Factorization Algorithms

Corresponding Author:

Olivier François

Faculty of Medicine

Grenoble

F38706 La Tronche

France

+334 56 52 00 25 (ph.)

+334 56 52 00 55 (fax)

olivier.francois@imag.fr

Abstract

Inference of individual ancestry coefficients, which is important for population genetic and association studies, is commonly performed using computer-intensive likelihood algorithms. With the availability of large population genomic data sets, fast versions of likelihood algorithms have attracted considerable attention. Reducing the computational burden of estimation algorithms remains, however, a major challenge. Here, we present a fast and efficient method for estimating individual ancestry coefficients based on sparse non-negative matrix factorization algorithms. We implemented our method in the computer program `sNMF`, and applied it to human and plant data sets. The performances of `sNMF` were then compared to the likelihood algorithm implemented in the computer program `ADMIXTURE`. Without loss of accuracy, `sNMF` computed estimates of ancestry coefficients with run-times approximately 10 to 30 times shorter than those of `ADMIXTURE`.

1 Introduction

Inference of population structure from multi-locus genotype data is commonly performed using likelihood methods implemented in the computer programs `STRUCTURE`, `FRAPPE` or `ADMIXTURE` (Pritchard et al. 2000a, Tang et al. 2005, Alexander et al. 2009). These programs compute probabilistic quantities called *ancestry coefficients* that represent the proportions of an individual genome that originate from multiple ancestral gene pools. Estimation of ancestry proportions is important with many respects, for example in delineating genetic clusters, drawing inference about the history of a species, screening genomes for signatures of natural selection, and performing statistical corrections in genome-wide association studies (Pritchard et al. 2000b; Marchini et al. 2004; Price et al. 2006; Frichot et al. 2013).

Individual ancestry coefficients can be estimated using either supervised or unsupervised statistical methods. Supervised estimation methods use predefined source populations as ancestral populations. Classical supervised estimation approaches were based on least-squares regression of allele frequencies in hybrid and source populations (Roberts and Hiorns 1965; Cavalli-Sforza and Bodmer 1971). Unsupervised approaches attempt to infer ancestral gene pools from the data using likelihood methods. An undesired feature of likelihood methods is that they can be computer intensive, with typical runs lasting several hours or more. With the use of dense genomic data and increased sample sizes, reducing the time lag necessary to perform estimation is a major challenge of population genetic data analysis.

A fast approach to the estimation of ancestry coefficients is by using principal component analysis (PCA, Patterson et al. 2006). PCA is an exploratory method that describes high-dimensional data using a small number of dimensions, and makes no assumptions about sampled and ancestral populations. Using PCA can lead to results surprisingly close to likelihood methods, and connections between methods have been intensively investigated during the last years (Patterson et al. 2006; Engelhardt and Stephens 2010; Frichot et al. 2012; Lawson et al. 2012; Lawson and Falush 2012). But a drawback of PCA is that interpretation in terms of ancestry is often difficult, as it can be confounded by demographic factors or irregular sampling designs (Novembre and Stephens 2008; Mc Vean 2009; François et al. 2010).

In this study, we introduce computationally fast algorithms that lead to estimates of ancestry coefficients comparable to those obtained with `STRUCTURE` or `ADMIXTURE`. The algorithms were implemented in the computer program `sNMF` based on sparse non-negative matrix factorization (NMF) and least-squares optimization (Lee and Seung 1999; Kim and Park 2007; Kim and Park 2011). Like PCA, NMF algorithms are flexible approaches that are robust to departures from traditional population genetic model assumptions. In addition, NMF algorithms produce estimates of ancestry proportions with run-times that are much shorter than `STRUCTURE` or `ADMIXTURE`. This study assesses the utility of NMF algorithms when analyzing population genetic data sets, and compares the performances of the algorithms inplemented in `sNMF` with those implemented in `ADMIXTURE` on the basis of human and plant data.

2 Materials and Methods

To provide statistical estimates of ancestry proportions using multi-locus genotype data sets, we implemented sparse NMF least-squares optimization algorithms in the computer program `sNMF`.

Modeling ancestry coefficients. We considered allelic data for a sample of n multi-locus genotypes at L loci representing single nucleotide polymorphisms (SNPs). The data were stored into a genotypic matrix (X) where each entry records the number of derived alleles at locus ℓ for individual i. For autosomes in a diploid organism, the number of derived alleles at locus ℓ is then equal to 0, 1 or 2. In our algorithm, we used 3 bits of information to encode each 0, 1 or 2 value as an indicator of a heterozygote or a homozygote locus. In other words, the value 0 was encoded as 100, 1 was encoded as 010 and 2 as 001. The use of a binary coding warrants that the entries sum up to L for each row of the transformed data matrix.

Admixture models generally suppose that the genetic data originate from the admixture of K ancestral populations, where K is unknow a priori. Given K populations, the probability that individual i carries j derived alleles at locus ℓ can be written as follows

$$p_{i\ell}(j) = \sum_{k=1}^{K} q_{ik} g_{k\ell}(j) \quad j = 0, 1, 2, \tag{1}$$

where q_{ik} is the fraction of individual i's genome that originates from the ancestral population k, and $g_{k\ell}(j)$ represents the homozygote ($j = 0, 2$) or the heterozygote ($j = 1$) frequency at locus ℓ in population k. Since it makes no assumption about Hardy-Weinberg equilibrium, the above framework is appropriate to deal with inbreeding and outbreeding in ancestral populations. Using our binary coding, equation (1) writes as

$$P = QG\,, \tag{2}$$

where $P = (p_{i\ell})$ is a $n \times 3L$ matrix, $Q = (q_{ik})$ is a $n \times K$ matrix, and $G = (g_{k\ell}(j))$ is a $K \times 3L$ matrix. The Q matrix records ancestry proportions for each individual in the sample. Though the focus of the above framework is on estimating ancestry estimates for each sampled individual, it can be easily modified to provide ancestry estimates based on allele frequencies in population samples.

Least-squares estimates of ancestry proportions. We approached the inference of ancestry coefficients by using least squares (LS) optimization algorithms (Engelhardt and Stephens 2010). Estimates of the Q and G matrices were obtained after minimizing the following least-squares criterion

$$\mathrm{LS}(Q,G) = \|X - QG\|_{\mathrm{F}}^2\,, \tag{3}$$

where $\|M\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix M (Berry et al. 2007). Without constraints on Q and G, the solutions of the LS problem are given by the singular value decomposition of the matrix X, and the resulting matrices Q and G contain the scores and loadings of a PCA. To obtain ancestry coefficients, the matrices Q and G must have non-negative entries such that

$$\sum_{k=1}^{K} q_{ik} = 1\,, \quad \sum_{j=0}^{2} g_{k\ell}(j) = 1. \tag{4}$$

With the constraints of equation (4), estimating ancestry coefficients and genotypic frequencies is equivalent to performing a non-negative matrix factorization (NMF) of the data matrix, X. NMF was previously applied to gene expression data (Kim and Park 2007), and algorithms for NMF were surveyed and compared in (Kim and Park 2011). In `sNMF`, estimates of Q and G were computed using the Alternating Non-negativity-constrained Least Squares (ANLS) algorithm with the active set (AS) method (Berry et al. 2007; Kim and Park 2011). We modified the ANLS-AS algorithm as follows.

Our algorithm begins with the initialization of the Q entries with non-negative values. Then it iterates the following cycles until convergence. The first step of an algorithm cycle consists of computing a non-negative matrix G that minimizes the quantity

$$\mathrm{LS}_1(G) = \left\|X - QG\right\|_F^2\,, G \geq 0. \tag{5}$$

The G matrix was obtained by setting all negative entries to zero, after solving classical linear regression equations. The obtained solution was then normalized so that its entries satisfy equation (4).

Given G, the second step of the cycle consists of computing a non-negative matrix Q that minimizes the quantity

$$\mathrm{LS}_2(Q) = \left\| \begin{pmatrix} G^T \\ \sqrt{\alpha}\; e_{1\times K} \end{pmatrix} Q - \begin{pmatrix} X^T \\ 0_{1\times n} \end{pmatrix} \right\|_F^2 \tag{6}$$

where $e_{1\times K}$ is a row vector having all entries equal to 1, $0_{1\times n}$ is a vector of length n with all entries equal to 0, and α is a non-negative *regularization parameter*. This minimization problem was solved using the block principal pivoting method proposed by Kim and Park (2011). The obtained solution, Q, was then normalized so that the row entries sum up to 1. Iterations were stopped based on a stationarity criterion derived from the Karush-Kuhn-Tucker conditions (Kim and Park 2011), and when the relative difference between two successive values of the criterion was less than a tolerance threshold of $\epsilon = 10^{-4}$.

For α greater than zero, the algorithm amounts to performing *sparse* NMF for the data matrix X. We tested values α greater than zero because they can reduce the variance of Q and G estimates for the smaller data sets, force irrelevant estimates to zero and improve the numerical behavior the ANLS minimization algorithm. In addition, the programming structures used in `sNMF` optimized the time spent in memory access. Several algorithmic methods were also used to accelerate computation of matrix products. While we evaluated sNMF run-times using a single computer processor unit (2.4 GHz 64bits Intel Xeon), a multi-threaded version of the `sNMF` program was also developed for multi-processor systems.

Data sets. Ancestry inference and run-time analyses were performed on 6 world-wide samples of genomic DNA from 52 populations of the Human Genome Diversity Project – Centre d'Etude du Polymorphisme Humain (HGDP-CEPH). Five panels were extracted from the Harvard HGDP-CEPH database. These panels were given IDs HGDP00778, HGDP00542, HGDP00927, HGDP00998 and HGDP01224, and contained precisely ascertained genotypes of $n = 934$ individuals. The genotypes were specifically designed for population genetic analyses (Patterson et al. 2012). Each marker was ascertained in individuals of Han,Papuan, Yoruba, Karitiana and Mongolian ancestry, and the data matrices included 78,253, 48,531, 124,115, 2,635 and 10,664 SNPs respectively (Patterson et al. 2012, Table 1). A sample of 1,043 individuals from the HGDP-CEPH Human Genome Diversity Cell Line Panel was also analyzed. The genotypes were generated on Illumina 650K arrays (Li et al. 2008), and the SNP data were filtered to remove low quality SNPs included in the original files. In addition, we used data from the 1000 Genomes Project. The 1000 Genomes Project data contains the genomes of 1,092 individuals from 14 populations, constructed

using a combination of low-coverage whole-genome and exome sequencing (phase 1 data, The 1000 Genomes Project Consortium 2012). The data matrix included 2.2M polymorphic sites across the human genome (Table 1).

To examine the robustness of `sNMF` to departures from classical population genetic hypotheses, additional analyses were performed on a sample of $n = 168$ European accessions of the plant species *Arabidopsis thaliana*. *A. thaliana* is a widely distributed self-fertilizing plant known to harbour considerable genetic variation and complex patterns of population structure and relatedness (Atwell et al. 2010). We analyzed 216,130 SNPs spread across the genome of *A. thaliana* (Atwell et al. 2010, Table 1).

Comparisons with `ADMIXTURE`. The computer program `ADMIXTURE` (version 1.22) estimates ancestry coefficients based on the likelihood model implemented in `STRUCTURE`. In `ADMIXTURE`, the assumption of Hardy-Weinberg equilibrium in ancestral populations translates into a binomial model for allele counts at each locus. Considering unrelated individuals, the logarithm of the likelihood can thus be computed as follows

$$\mathcal{L}(Q,F) = \sum_i \sum_\ell \left(x_{i\ell} \log(\sum_k q_{ik} f_{k\ell}) + (1 - x_{i\ell}) \log(\sum_k q_{ik}(1 - f_{k\ell})) \right)$$

up to an additive constant that does not influence estimation algorithms. In this formula, $Q = (q_{ik})$ represents the matrix of ancestry coefficients for all individuals, and $F = (f_{k\ell})$ represents a matrix of allele frequencies for all loci. The F matrix can be converted to a G matrix comparable to the one computed by `sNMF` using the binomial model: $g_{k\ell}(0) = (1 - f_{k\ell})^2$, $g_{k\ell}(1) = 2f_{k\ell}(1 - f_{k\ell})$, and $g_{k\ell}(2) = f_{k\ell}^2$. `ADMIXTURE` provides numerical estimates of Q and F that maximize the quantity $\mathcal{L}(Q,F)$. The local optimization algorithm relies on a block relaxation scheme using sequential quadratic programming for block updates, coupled with a quasi-Newton acceleration of convergence.

A difficulty with optimization algorithms used by `ADMIXTURE` and `sNMF` is that the solutions produced can be dependent on the initial values used for Q, F or G. To enable comparisons with estimates obtained with `ADMIXTURE`, the clusters output by runs of each programs were permuted using `CLUMPP` (Jakobsson et al. 2007). Differences in ancestry estimates obtained with `ADMIXTURE`

(Q^{ADM}) and with `sNMF` (Q^{sNMF}) were assessed by two measures. The first measure was defined as the root mean squared error (RMSE) between the matrices Q^{ADM} and Q^{sNMF} obtained from each program

$$\mathrm{RMSE} = \left(\frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} (q_{ik}^{\mathrm{ADM}} - q_{ik}^{\mathrm{sNMF}})^2 \right)^{1/2} .$$

Though G matrices could be mainly considered as nuisance parameters for our estimation problem, a similar RMSE criterion was defined for comparing them. The second measure was defined as the squared Pearson correlation coefficient (R^2) between the matrices Q^{ADM} and Q^{sNMF}. When simulations with known Q-matrices were analysed, one of the two matrices was replaced by the true Q-matrix used to generate the simulated data.

Runs of `ADMIXTURE` and `sNMF` were performed for values of the number of clusters set to $K = 2 - 10$, 15, and 20 for human data sets and set to $K = 2 - 7$ for *A. thaliana*. For `sNMF`, the values of the regularization parameter (α) ranged between 0 and 10,000 using a $\log_{10}$ scale (5 values). Each run was replicated five times for a total of 1,410 experiments. Missing data imputation was initially performed after resampling missing genotypes from empirical frequencies at each locus. The missing values were updated using predictive probabilities after 20 sweeps of algorithm (see below).

Cross-entropy criterion. We employed a cross-validation technique based on imputation of masked genotypes to evaluate the prediction error of ancestry estimation algorithms (Wold 1978, Eastment and Krzanowski 1982). The procedure partitioned the genotypic matrix entries into a training set and a test set. To build the test set, 5% of all genotypes were randomly selected, and tagged as missing values. The occurence probabilities for the masked entries were computed using the program outputs obtained from training sets according to the following formula

$$p_{i\ell}^{\mathrm{pred}}(j) = \sum_{k=1}^{K} q_{ik} g_{k\ell}(j) \,, \quad j = 0, 1, 2 \,. \tag{7}$$

`ADMIXTURE` predicts each masked value by $E[x_{i\ell}|Q^{\mathrm{ADM}}, F^{\mathrm{ADM}}] = 2 \sum_k q_{ik}^{\mathrm{ADM}} f_{k\ell}^{\mathrm{ADM}}$ and the prediction error is estimated by averaging the squares of the deviance residuals for the binomial model (Alexander et al. 2011). Extending the approach employed by `ADMIXTURE` to our non-parametric approach, the predicted values were compared to the masked values, $x_{i\ell}$, by averaging the quantity

defined as $-\log p_{i\ell}^{\text{pred}}(x_{i\ell})$ over all SNPs in the test set. In statistical terms, our criterion provides an estimate of the following quantity

$$H(p^{\text{sample}}, p^{\text{pred}}) = -\sum_{j=0}^{2} p^{\text{sample}}(j) \log p_{i\ell}^{\text{pred}}(j)\,, \quad j = 0, 1, 2. \tag{8}$$

This quantity corresponds to the sum of the Kullback-Leiber divergence between the sampled (p^{sample}) and predicted (p^{pred}) allelic distributions and the Shannon entropy of the sample distribution. It also corresponds to the cross-entropy between p^{sample} and p^{pred}. The number of ancestral gene pools (K) and the regularization parameter (α) were chosen to minimize the cross-entropy criterion. In general smaller values of the criterion indicate better algorithm outputs and estimates. The standard error of the cross-entropy criterion is of order $1/\sqrt{n_L}$ where n_L is the number of masked genotypes. For data sets including 1,000 individuals genotyped at more than $20,000$ SNPs, the third digit of the cross-entropy criterion can be significant.

Simulated data analysis. We adopted a simulation approach to compare RMSEs between the Q-matrix computed by `ADMIXTURE` or by `sNMF` and a known matrix used to generate the simulated data. In addition we assessed whether or not the correct value of K could be identified by `sNMF` using the cross-entropy criterion.

In a first series of simulations, we used the 1,000 Genomes Project data set to generate artificial data showing various levels of admixture. As ancestral populations, we chose the CHB, GBR, and YRI samples (The 1,000 Genomes Project 2012). We considered 50,000 SNPs in linkage equilibrium exhibiting no missing genotypes. The allele frequencies observed in our three ancestral populations were used as the true values for the F matrix. The genotypic matrix was constructed according to the binomial model used by `ADMIXTURE`. For 1000 individuals in each simulated data set, a Q matrix was simulated from a Dirichlet probability distribution, and several parameters were explored. Our experiments reproduced the parameters used for evaluating the accuracy of `ADMIXTURE` in a previous study (Alexander et al. 2009). Runs of `sNMF` were performed for values of the number of clusters set to $K = 2 - 5$ ($\alpha = 0$), and the choice of K was made on the basis on the cross-entropy criterion. For $K = 3$, the values of the regularization parameter (α) was varied between 0 and 100.

Additional data sets were created to mimic the population structure of European populations of *Arabidopsis thaliana* using 10,000 SNPs (168 individuals). To defined ancestral frequencies, we used the western European populations grouping samples from the United Kingdom, Belgium and France (23 individuals), central European populations grouping samples from the Czech Republic (24 individuals), and northern European populations grouping samples from Finland and Northern Sweden (13 individuals). `ADMIXTURE` and `sNMF` grouped those samples within 3 well-separated clusters exhibiting low levels of admixture with other plant populations. The empirical frequencies computed from the 3 populations were considered as the true frequencies for a generative model with $K = 3$ ancestral populations. From empirical frequencies, we computed genotypic frequencies, $f_{k\ell}$, using 4 distinct values of population inbreeding coefficient, $F_{\text{IS}} = 25{-}100\%$, that corresponded to moderate and strong levels of inbreeding. For 168 individuals, 10,000 genotypes were simulated using the sampling equation $p(x_{i\ell} = j) = \sum_k q_{ik} g_{k\ell}(j)$, where q_{ik} corresponds to the Q-matrix computed from the full empirical data set (216K SNPs). In addition, simulated data sets were generated with or without missing data (0 or 20 %). Fifty replicates were created for each value of the inbreeding coefficient and for each value of the ratio of missing data.

3 Results

We used the program `sNMF` to implement non-negative matrix factorization algorithms and to compute least-squares estimates of ancestry coefficients for world-wide human population samples and for European populations of the plant species *Arabidopsis thaliana*. As in the likelihood model implemented in the computer programs `STRUCTURE` and `ADMIXTURE`, `sNMF` supposes that the genetic data originate from the admixture of K parental populations, where K is unknown, and it returns estimates of ancestry proportions for each multi-locus genotype in the sample (Pritchard et al. 2000a, Alexander et al. 2009). To estimate ancestry coefficients, `sNMF` solves a constrained least squares minimization problem using an alternating algorithm based on a block principal pivoting method (Kim and Park 2011, see Materials and Methods).

Comparison of ancestry estimates for HGPD data sets. First we evaluated the ability of `ADMIXTURE` estimates to be accurately reproduced by `sNMF` for 5 Harvard HGDP panels and for the HGDP-CEPH data set (Li et al. 2008; Patterson et al. 2012). For each run of `ADMIXTURE`, we computed a maximum squared correlation coefficient (R^2) and a minimum root mean squared error (RMSE) over runs of `sNMF` performed with the same number of clusters (K). For K ranging from 5 to 10, squared correlation coefficients remained greater than 0.96 across all runs (480 runs, Figure 1). Average values of the RMSE remained lower than 5.5% across all runs (Table S1). These results provided evidence that `sNMF` estimates closely reproduce those obtained with `ADMIXTURE` across the 6 HGDP data sets.

Run-time analysis. Next we performed run-time analyses for `ADMIXTURE` and for `sNMF` using the 1000 Genomes Project phase 1 data in addition to the previous HGDP data sets. The run-times were averaged over distinct random seed values for each value of K. Run-times increased with the number of SNPs in the data set and with the number of clusters in each algorithm (Figure 2, Table 2, Figure S1). For data set HGDP01224 (10.6K SNPs), it took on average 0.8 minute (1.7 minutes) to `sNMF` to compute ancestry estimates for $K = 10$ ($K = 20$) clusters. For `ADMIXTURE`, the run-time was on average equal to 11 minutes (48 minutes) for $K = 10$ ($K = 20$) clusters. For panel HGDP00778 (78K SNPs), it took on average 7.2 minutes (15 minutes) to `sNMF` to compute

ancestry estimates for $K = 10$ ($K = 20$) clusters. For `ADMIXTURE`, the average run-time was 1.5 hour (6.2 hours) for $K = 10$ ($K = 20$) clusters. For the CEPH-HGDP data sets (660K SNPs), it took on average 55 minutes (2.1 hours) to `sNMF` to compute ancestry estimates for $K = 10$ ($K = 20$) clusters. For `ADMIXTURE`, the average run-time was 12 hours (38 hours) for $K = 10$ ($K = 20$) clusters. Run-times increased in a quadratic fashion with K for `ADMIXTURE` whereas they increased linearly for `sNMF` (Figure 2). For the values of K used in our analyses, `sNMF` ran 5 to 30 times faster than `ADMIXTURE` when these programs were applied to HGPD data sets. Regarding the 1000 Genomes Project phase 1 data set, the average run-times of `sNMF` were approximately equal to 2.8 hours (4.6 hours) for number of clusters $K = 5$ ($K = 10$). The `ADMIXTURE` runs led to similar estimates of Q, but a single run on the phase 1 data set took more than 19 hours for $K = 5$ (59 hours for $K = 10$).

Prediction of masked genotypes. To decide which program options could provide the best estimates, we employed a cross-validation technique based on the imputation of masked genotypes (Wold 1978; Alexander et al. 2011). The cross-validation method partitions the genotypic matrix entries into a training set and a test set, that are used for estimation and validation sequentially. To build test sets, 5% of the genotypic matrix entries were tagged as missing values. The masked entries were then predicted using estimates obtained from training sets. Predictions were assessed using a cross-entropy criterion that measured the capability of an algorithm to correctly impute masked genotypes (see Material and Methods). Lower values of the cross-entropy criterion generally indicate better predictive capabilities of an algorithm.

Using the cross-entropy criterion, we performed an extensive analysis of `sNMF` program outputs to assess which values of the number of clusters (K) and regularization parameter (α) could provide the best prediction of masked genotypes (Figure 3, **Figure S2**). For HGDP data sets with moderate size (panels HGDP00998 and HGDP01224), values of K around 7–8 provided the best predictive results. For larger human data sets, cross-entropy values did not stabilize for $K \leq 10$, indicating that more than ten clusters were necessary to describe population structure. Choices of regularization parameter values greater than 1,000 were generally discarded by the cross-entropy criterion. For panels of moderate size, the best ancestry estimates were obtained

for values around $\alpha = 100$. The influence of the regularization parameter was substantial for the smallest data sets, but for the largest ones a wide range of values led to comparable imputation results (Figure S2). Regardless of the value of the regularization parameter, $K = 5$ clusters led to the best results for The 1000 Genomes Project data set (Figure 3). This last result is in accordance with the criteria used for choosing populations included in the 1000 Genomes Project.

Ancestry estimates. To compare ancestry estimates obtained from particular runs of `sNMF` and `ADMIXTURE`, we displayed the Q-matrices computed by each program for the Harvard HGPD panel HGDP00778 (78K SNPs), the HGDP-CEPH data (660K SNPs), The 1000 Genomes Project phase 1 data (2.2M SNPs) and European populations of *A. thaliana* (216K SNPs). Using $K = 7$ ancestral populations for the Harvard HGPD panel HGDP00778, the cross-entropy criterion was equal to 0.747 for the `ADMIXTURE` run, and it was equal to 0.762 for the `sNMF` run. The criterion favored `ADMIXTURE` in this case, but the two runs led to very close estimates of the Q-matrix ($R^2 = 0.99$, **Figure 4**). When the programs were applied to the HGDP-CEPH data ($K = 7$), the cross-entropy criterion was equal to 0.691 for the `ADMIXTURE` run and 0.704 for `sNMF` (**Figure 4**). This particular `ADMIXTURE` run identified clusters that separated the African hunter-gatherer populations from the other populations, whereas `sNMF` identified a unique cluster in Africa. In the `sNMF` run, Middle East populations were separated from European populations (Figure 4). The differences between `ADMIXTURE` and `sNMF` results disappeared when additional runs were performed with distinct random seeds. Using $K = 5$ for the 1000 Genomes Project phase 1 data, `sNMF` identified clusters that correspond to the main geographic regions of the world, similarly to `ADMIXTURE` (Figure 5, cross-entropy = 0.5010). Substantial levels of European ancestry in African-Americans, Mexican-Americans, Puerto-Ricans and Columbians were inferred by `sNMF` and by the other program. An interesting case was with the application of ancestry estimation programs to European populations of *A. thaliana*, a selfing plant characterized by high levels of inbreeding (Atwell et al. 2010). Using $K = 3$, the cross-entropy criterion for `ADMIXTURE` was equal to 0.641 on average, while the average value for `sNMF` was equal to 0.483. The value of the criterion suggests that `sNMF` estimates were more accurate than those obtained from `ADMIXTURE`. The graphical output of the Q matrix displayed clinal variation of ancestry

coefficients occurring along an East-West gradient separating two clusters, and Northern Swedish accessions were grouped into a separate cluster. These results supported previous estimates based on sequence data (Figure S3, François et al. 2008).

Simulated data analysis. To further ascertain the accuracy of `sNMF` estimates and to compare those estimates with `ADMIXTURE`, we employed computer simulations based on the 1,000 Genomes Project and *Arabidopsis thaliana* data sets. We also assessed the ablility of the cross-entropy criterion to correctly identify the value of K when it is known.

In a first series of simulations, we used the 1,000 Genomes Project data to generate genotypes showing various levels of admixture. As our ancestral populations, we chose the CHB, GBR, and YRI samples (The 1,000 Genomes Project 2012), and true Q matrices were created using several parameterizations of the Dirichlet distribution (Table 3). Genotypic matrices were simulated according to the binomial model used by `ADMIXTURE`. In this context, `ADMIXTURE` estimates are thus expected to be more accurate than `sNMF` estimates. For the range of parameters explored in the simulations, root mean squared errors comparing the estimated and true value of the Q-matrix remained lower than 2% for both programs (Table 3). For moderate levels of admixture, differences in statistical errors were lower than 1% regardless of the value of the regularization parameter, α, used in `sNMF`. This result indicated that `sNMF` estimates are generally accurate, and that relatively small values of α do not influence `sNMF` outputs for data sets of size comparable to those used in simulations. Root mean squared errors comparing the estimated and true value of the G-matrix were slightly lower for `ADMIXTURE` than for `sNMF` (Table 3). This could be explained as we simulated from a binomial model (unrelated individuals), and as the number of degrees of freedom in `sNMF` is twice the number of degrees of freedom in `ADMIXTURE`. Regarding the choice of the number of clusters, the cross-entropy criterion was minimal for $K = 3$ for every simulated data sets (Table S2).

To evaluate the relative impact of linkage disequilibrium (LD) on `sNMF` and `ADMIXTURE` ancestry estimates, we considered subsets of SNPs sampled from the 1000 Genomes Project data set. We compared ancestry estimates computed by each program for data sets containing blocks of more than 30 SNPs spaced less than 20 kb apart and for data sets containing SNPs separated by more

than 20 kb (20,000 SNPs, 20 replicates). Using linked blocks of SNPs, the average value of the RMSE over all runs was equal to 0.0859 (0.0297 for unlinked SNPs) for `sNMF` whereas it was equal to 0.0976 for `ADMIXTURE` (0.257 for unlinked SNPs).). Our results show that LD had an impact on the accuracy of ancestry estimates regardless of the program used, and that the magnitude of the effect was similar for `ADMIXTURE` and `sNMF` (Figure S4).

We used another series of simulated data to evaluate the sensitivity of `ADMIXTURE` and `sNMF` estimates to the presence of related individuals and inbreeding in the sample. Based on empirical data, we used simulation models that mimicked the population structure of European populations of *Arabidopsis thaliana*. First we verified that the true value of the number of ancestral populations was correctly recovered by the `sNMF` program using the cross-entropy criterion ($K = 3$). Next we evaluated statistical errors for `ADMIXTURE` and `sNMF` estimates of the Q-matrix. RMSEs remained lower than 4% for both programs. These results showed that the two programs produced accurate estimates of the Q-matrix in the presence of inbreeding and missing data **(Figure 6)**.

`ADMIXTURE` estimates were robust to the inclusion of moderate levels of inbreeding in the sample. When the values of the inbreeding coefficient were equal to $0.25-0.5$, `ADMIXTURE` ancestry estimates were more accurate than `sNMF` estimates. When the value of the inbreeding coefficient were greater than 0.5 and for fully inbred lines, `sNMF` produced better estimates than `ADMIXTURE` **(Figure 6)**. The cross-entropy criterion was smaller for `sNMF` than for `ADMIXTURE`, showing that `sNMF` produced better prediction of masked genotypes than `ADMIXTURE` (Figure S5). This result can be explained by a more accurate estimation of genotypic frequencies for `sNMF` than for `ADMIXTURE` in the presence of strong levels of inbreeding.

4 Discussion

We applied the computer program `sNMF` to the estimation of individual ancestry coefficients using large population genetic data sets for humans and for *A. thaliana*, and compared the program performances to those of `ADMIXTURE`. For 6 HGDP data sets, ancestry estimates obtained with `sNMF` and `ADMIXTURE` strongly agreed with each other. In addition, the `sNMF` program was able to analyze the 1000 Genomes project phase 1 data set within a few hours using a standard computer processing unit. Without significant loss of accuracy, `sNMF` computed estimates of admixture proportions within run-times that were about 10 to 30 times faster than those of `ADMIXTURE`.

The approach used by `sNMF` is based on theoretical connections between likelihood approaches, PCA and NMF methods (Ding et al. 2008; Engelhardt and Stephens 2010; Lawson et al. 2012; Parry and Wang 2013). Several methods can be applied to computing NMF estimates, including the multiplicative update algorithm, the projected-gradient method and the alternating least-squares algorithm (Brunet et al. 2004; Berry et al. 2007; Kim and Park 2011). For population genetic data, we found that alternating least-squares algorithms coupled with the active set method provided the best trade-off between speed and accuracy, and improved significantly over other NMF implementations (Kim and Park 2011).

To decide which algorithm yielded the best estimates, we introduced a predictive criterion based on the computation of cross-entropy and the imputation of masked genotypes. For HGDP data sets, the cross-entropy criterion discarded large values of the `sNMF` regularization parameter (greater than 1,000). For the large data sets, a wide range of values of the regularization parameter reached similar predictive values. For data sets having less than 10,000 SNPs, we found that parsimony (i.e., large values of α) could improve estimation of ancestry coefficients. We observed that a likelihood approach could benefit the analysis of modest-sized data sets or data containing a large number of missing genotypes. For larger data sets and missing less than 20% genotypes, `sNMF` ancestry estimates were statistically close to those obtained with `ADMIXTURE`, and both programs were equally efficient at predicting masked genotypes. Statistical theory actually predicts that errors in evaluating the cross-entropy criterion are of order $O(1/\sqrt{n_L})$ where n_L is the number of masked genotypes. For Harvard HGDP panels, differences between the `ADMIXTURE` and `sNMF` results

could be considered hardly significant and estimates were statistically similar. The example of the Harvard HGDP panels showed that the cross-entropy criterion could also be used to discriminate among program runs regardless of the program used.

The assumptions underlying `STRUCTURE` and `ADMIXTURE` rely on simplified population genetic hypotheses. More specifically, the assumptions include absence of genetic drift, Hardy-Weinberg and linkage equilibrium in ancestral populations. The coding used by `sNMF` enabled the estimation of homozygote and heterozygote frequencies, and avoided Hardy-Weinberg equilibrium assumptions. Although `ADMIXTURE` analyses were robust to small departures from Hardy-Weinberg equilibrium in human data, `sNMF` was more appropriate to deal with inbreed lineages. For European populations of *A. thaliana*, the values of the cross-entropy criterion indicated better predictive results for `sNMF` than for `ADMIXTURE`. The difference between `sNMF` and `ADMIXTURE` predictions could be explained as the binomial model of `ADMIXTURE` is not suited to the high levels of inbreeding observed in *A. thaliana* populations (Atwell et al. 2010). As seen from equation (1), an implicit assumption underlying NMF predictions is that genotypic frequencies can be formed according to instantaneous mixtures of ancestral frequencies without genetic drift. Interpretations of admixture using estimates obtained using likelihood and least-squares methods can be confounded by the existence of phylogenetic relationships among population samples (see Patterson et al. 2012 for an alternative approach), or by complex demographic scenarios such as spatial range expansion (François et al. 2010).

Comparing the relative computational performances of `ADMIXTURE` and `sNMF` was a difficult task because run-times are dependent on several factors. Those factors include the size and other characteristics of each data set, the tolerance threshold used when stopping program iterations, the use of multi-processor algorithms, and the initial values of the Q and G matrices. For example, run-times could be shortened by using initial values obtained after running the program on reduced data sets.

We explain the relative speed of the NMF algorithm by looking at algorithmic complexity for each program. The ANLS algorithm iterates cycles that solve linear regression equations for Q and G. The complexity of a single cycle of `sNMF` is of order $O(KLn)$, where K is the number of

clusters, n the number of individuals and L is the number of loci. The complexity of a single cycle of `ADMIXTURE` is of order $O(K^2Ln)$ (Alexander et al. 2009). Since the default tolerance threshold in this program implies that the program generally runs a small number of cycles (e.g., less than 40 cycles for the 78K SNPs Harvard HGDP panel), we observed that least-squares algorithms ran significantly faster than likelihood algorithms when analyzing large population genomic data sets with large values of K.

5 Acknowledgments

The authors are grateful to Nick Patterson, Eric Stone, and an anonymous reviewer for their useful comments on a previous version of this manuscript. We thank the 1,000 Genomes Project for authorizing us to use the phase 1 data. This work was supported by a grant from la Région Rhône-Alpes to Eric Frichot and Olivier François. Olivier François acknowledges support from Grenoble INP.

6 Web Resources

The `sNMF` code can be downloaded from the following URL:

http://membres-timc.imag.fr/Olivier.Francois/snmf.html.

References

Alexander, D.H., Novembre, J., Lange, K. (2009). Fast model-based estimation of ancestry in unrelated individuals. Genome. Res. *19*, 1655–1664.

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

Atwell, S., Huang, Y.S., Vilhjálmsson, B.J., Willems, G., Horton, M., Li, Y., Meng, D., Platt, A., Tarone, A.M., Hu, T.T., et al. (2010). Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature *465*, 627–631.

Berry, M.W., Browne, M., Langville, A.N., Pauca, V.P., Plemmons, R.J. (2007). Algorithms and applications for approximate nonnegative matrix factorization. Comput. Stat. Data. An. *52*, 155–173.

Brunet, J.P., Tamayo, P., Golub, T.R., Mesirov, J.P. (2004). Metagenes and molecular pattern discovery using matrix factorization. P. Natl. A. Sci. *101*, 4164–4169.

Cavalli-Sforza, L.L., Bodmer, W.F. (1971). The genetics of human populations. (New York: Dover Publications).

Ding, C., Li, T., Peng, W. (2008). On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. Comput. Stat. Data. An. *52*, 3913–3927.

Eastment, H.T., Krzanowski, W.J. (1982). Cross-validatory choice of the number of components from a principal component analysis. Technometrics *24*, 73–77.

Engelhardt, B.E., Stephens, M. (2010). Analysis of population structure: a unifying framework and novel methods based on sparse factor analysis. PLoS. Genet. *6*, 12.

François, O., Blum, M.G.B., Jakobsson, M., Rosenberg, N.A. (2008). Demographic history of European populations of *Arabidopsis thaliana*. PLoS. Genet. *4*, e1000075.

François, O., Currat, M., Ray, N., Han, E., Excoffier, L., Novembre, J. (2010). Principal component analysis under population genetic models of range expansion and admixture. Mol.

Biol. Evol. *27*, 1257–1268.

Frichot, E., Schoville, S.D., Bouchard, G., François, O. (2012). Correcting principal component maps for effects of spatial autocorrelation in population genetic data. Front. Genet. *3*, 254.

Frichot, E., Schoville, S.D., Bouchard, G., François, O. (2013). Testing for associations between loci and environmental gradients using latent factor mixed models. Mol. Biol. Evol. *30*, 1687–1699.

Kim, H., Park, H. (2007). Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics *23*, 1495–1502.

Kim, J., Park, H. (2011). Fast Nonnegative Matrix Factorization: An active-set-like method and comparisons. SIAM. J. Sci. Comput. *33*, 3261–3281.

Jakobsson, M., Rosenberg, N.A. (2007). CLUMPP: a cluster matching and permutation program for dealing with label switching and multimodality in analysis of population structure. Bioinformatics *23*, 1801–1806.

Lawson, D.J., Falush, D. (2012). Population identification using genetic data. Annu. Rev. Genom. Hum. G. *13*, 337–361.

Lawson, D.J., Hellenthal, G., Myers, S., and Falush, D. (2012). Inference of population structure using dense haplotype data. PLoS. Genet. *8*, e1002453.

Lee, D.D., Seung, H.S. (1999). Learning the parts of objects by non-negative matrix factorization. Nature *401(6755)*, 788–791.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., Myers, R.M. (2008). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

Marchini, J., Cardon, L.R., Phillips, M.S., Donnelly, P. (2004). The effects of human population structure on large genetic association studies. Nat. Genet. *36*, 512–517.

McVean, G. (2009). A genealogical interpretation of principal components analysis. PLoS. Genet. *5*, 10.

Novembre, J., Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. Nat. Genet. *40*, 646–649.

Parry, R.M., Wang, M.D. (2013). A fast least-squares algorithm for population inference. BMC bioinformatics *14*, 28.

Patterson, N., Price, A., Reich, D. (2006). Population structure and eigenanalysis. PLoS. Genet. *2*, e190.

Patterson, N.J., Moorjani, P., Luo, Y., Mallick, S., Rohland, N., Zhan, Y., Genschoreck, T., Webster, T., Reich, D. (2012). Ancient admixture in human history. Genetics *192*, 1065–1093.

Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A., Reich, D. (2006). Principal components analysis corrects for stratification in genome-wide association studies. Nat. genet. *38*, 904–909.

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

Pritchard, J.K., Stephens, M., Rosenberg, N.A., Donnelly, P. (2000). Association mapping in structured populations. Am. J. Hum. Genet. *67*, 170.

Roberts, D.F., Hiorns, R.W. (1965). Methods of analysis of the genetic composition of a hybrid population. Hum. Biol. *37*, 38–43.

Tang, H., Peng, J., Wang, P., Risch, N. (2005). Estimation of individual admixture: Analytical and study design considerations. Genet. Epidemiol. *28*, 289–301.

The 1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

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

7 Figure titles and legends

Figure 1. Correlation between `sNMF` and `ADMIXTURE` estimates. Squared correlation (coefficient of determination, R^2) between the ancestry coefficients estimated by each program. For each number of clusters (K), the result corresponds to the maximum correlation over 5 runs, averaged over values of the regularization parameter lower than 1,000 and over 6 HGDP data sets. The shaded area corresponds to a 95% confidence interval displayed for each value of the regularization parameter, α.

Figure 2. Run-times for `sNMF` and `ADMIXTURE` runs. Averaged time elapsed before the stopping criterion of the `sNMF` (blue) and `ADMIXTURE` (orange) programs is met. Time is expressed in unit of hours. A) Run-time analysis for Harvard HGDP panel 01224 (10.6K SNPs). B) Run-time analysis for Harvard HGDP panel 00778 (78K SNPs). C) Run-time analysis for the HGDP-CEPH data (660K SNPs).

Figure 3. Values of the cross-entropy criterion for `sNMF` runs (Human data sets). Minimal values of the cross-entropy criterion over 5 runs of `sNMF` for A-E) 5 Harvard HGDP panels, F) HGDP-CEPH data, and G) the 1000 Genomes Project data. The number of clusters ranged from 2 to 10.

Figure 4. Graphical representation of ancestry estimates obtained for HGDP data sets ($K = 7$). A) HGDP00778 panel (78K SNPs). Estimated ancestry coefficients using `ADMIXTURE` (top, cross-entropy = 0.747) and `sNMF` (bottom, cross-entropy = 0.762 and $\alpha = 100$). B) HGDP-CEPH data set (660K SNPs). Estimated ancestry coefficients using `ADMIXTURE` (top, cross-entropy = 0.691) and `sNMF` (bottom, cross-entropy = 0.704 and $\alpha = 100$).

Figure 5. Graphical representation of ancestry estimates obtained for the 1000 Genomes Project data set. A) Estimated ancestry coefficients using `sNMF` with $K = 5$ and $\alpha = 10,000$ (cross-entropy = 0.5010). B) Estimated ancestry coefficients using `sNMF` with $K = 6$ and $\alpha = 10,000$ (cross-entropy = 0.5011) (FIN: Finnish, GBR: British, IBS: Spanish, CEU: CEPH Utah residents, TSI: Tuscan, CHS: Southern Han Chinese, CHB: Han Chinese, JPT: Japanese, YRI: Yoruba, LWK: Luhya, ASW: African-American, PUR: Puerto Rican, CLM: Colombian, MXL: Mexican-American).

Figure 6. Accuracy of `ADMIXTURE` and `sNMF` in the presence of related individuals. RMSEs between estimated Q-matrices and a known matrix used to generate simulated data. Simulations mimicked the population structure of European populations of *Arabidopsis thaliana*. A-B) Moderate levels of inbreeding, $F_{\mathrm{IS}} = 25 - 50\%$, C-D) Strong levels of inbreeding, $F_{\mathrm{IS}} = 75 - 100\%$.

Tables

Table 1: **Data sets used in this study.**

Data set	Sample size	Number of SNPs	Reference
HGDP00778	934	78K	(Patterson et al. 2012)
HGDP00542	934	48.5K	–
HGDP00927	934	124K	–
HGDP00998	934	2.6K	–
HGDP01224	934	10.6K	–
HGDP-CEPH	1,043	660K	(Li et al. 2008)
1000 Genomes	1,092	2.2M	(The 1000 Genomes Project Consortium 2012)
A. thaliana	168	216K	(Atwell et al. 2010)

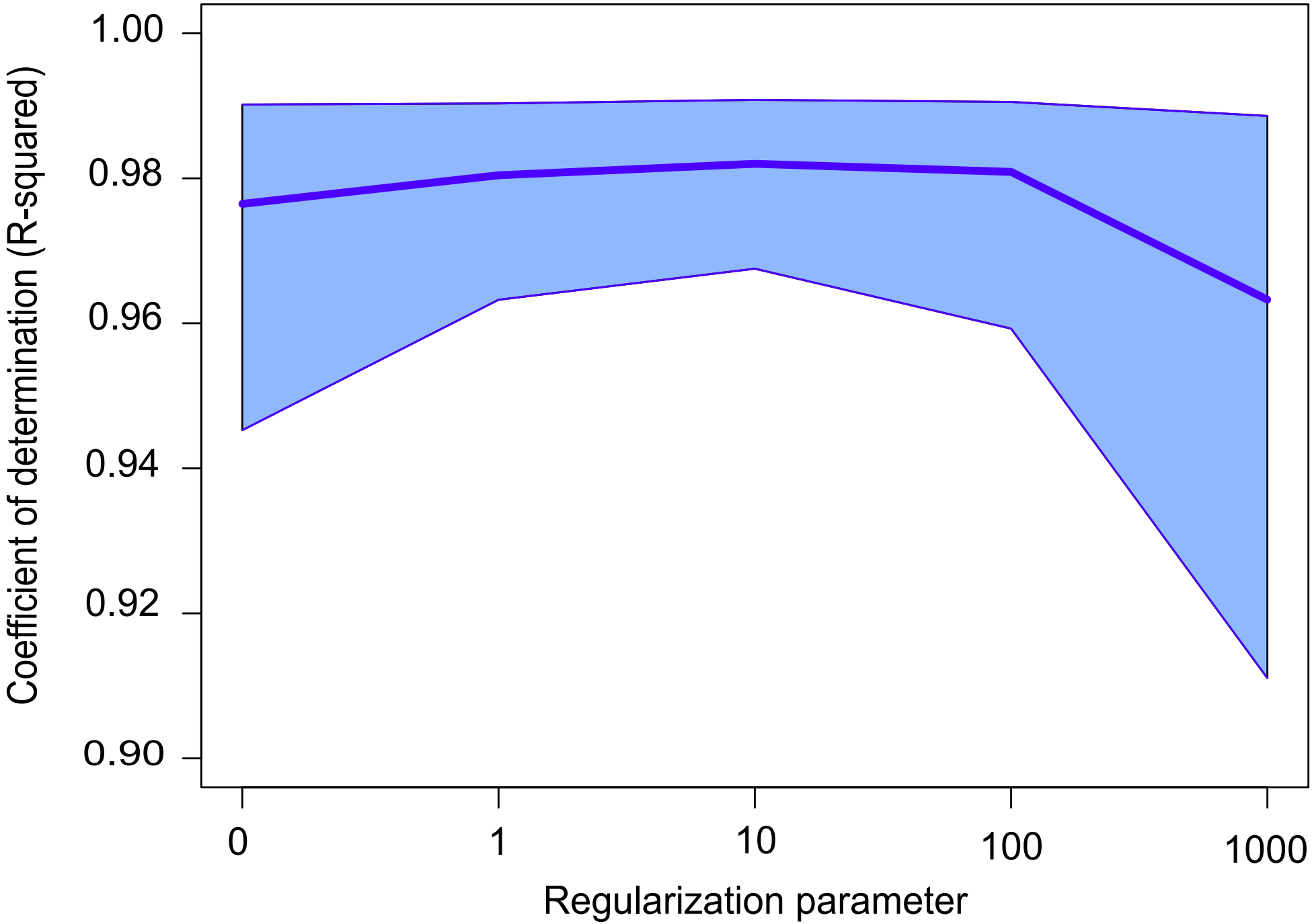
Table 2: **Run-time summary for sNMF and ADMIXTURE.** Average values and their 95% confidence intervals.

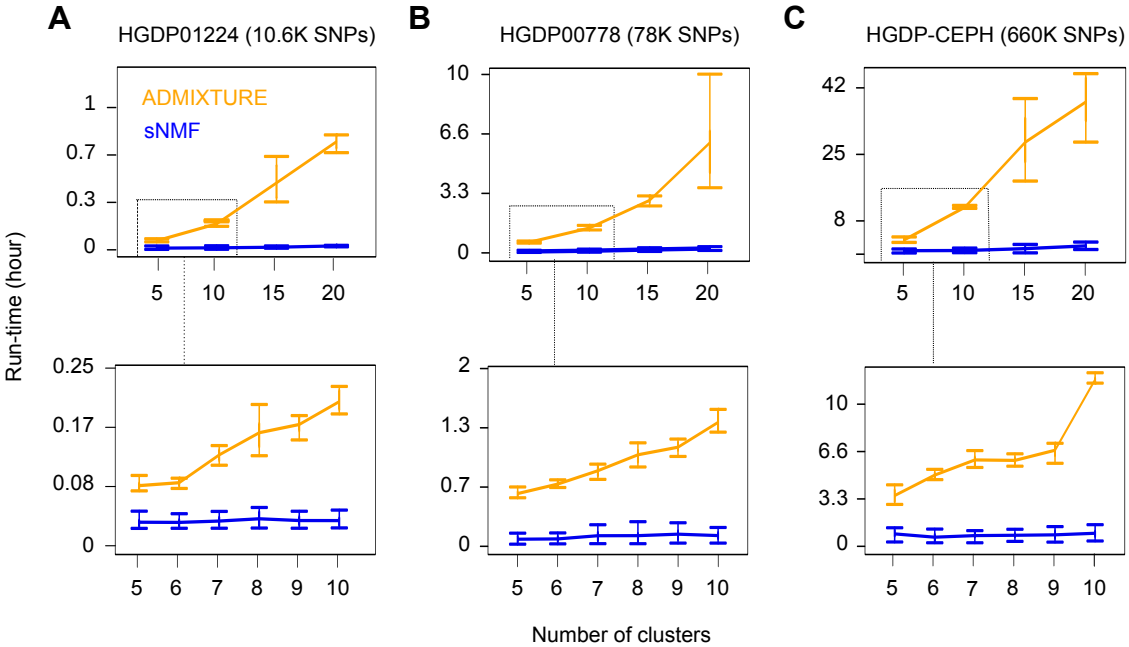
Data set	time	$K = 5$		$K = 10$		$K = 20$	
(Number of SNPs)	unit	sNMF	ADMIXTURE	sNMF	ADMIXTURE	sNMF	ADMIXTURE
HGDP01224	(minute)	**0.68**	4.4	**0.8**	11	**1.7**	48
(10.6K)		[0.1,1.6]	[3.4,4.9]	[0.18,1.7]	[9.9,12]	[1.3,1.8]	[41,55]
HGDP00778	(hour)	**0.087**	0.61	**0.12**	1.5	**0.25**	6.2
(78K)		[0.03,0.15]	[0.55,0.66]	[0.044,0.12]	[1.3,1.5]	[0.14,0.34]	[3.8,9.4]
HGDP-CEPH	(hour)	**0.9**	3.7	**0.92**	12	**2.1**	38
(660K)		[0.33,1.3]	[3,4.3]	[0.38,1.5]	[11,12]	[1.3,3.0]	[29,45]
1000 Genomes Project	(hour)	**2.8**	(19)	**4.6**	(59)	–	–
(2.2M)		[1.1,4.7]	–	[1.5,8.3]	–	–	–

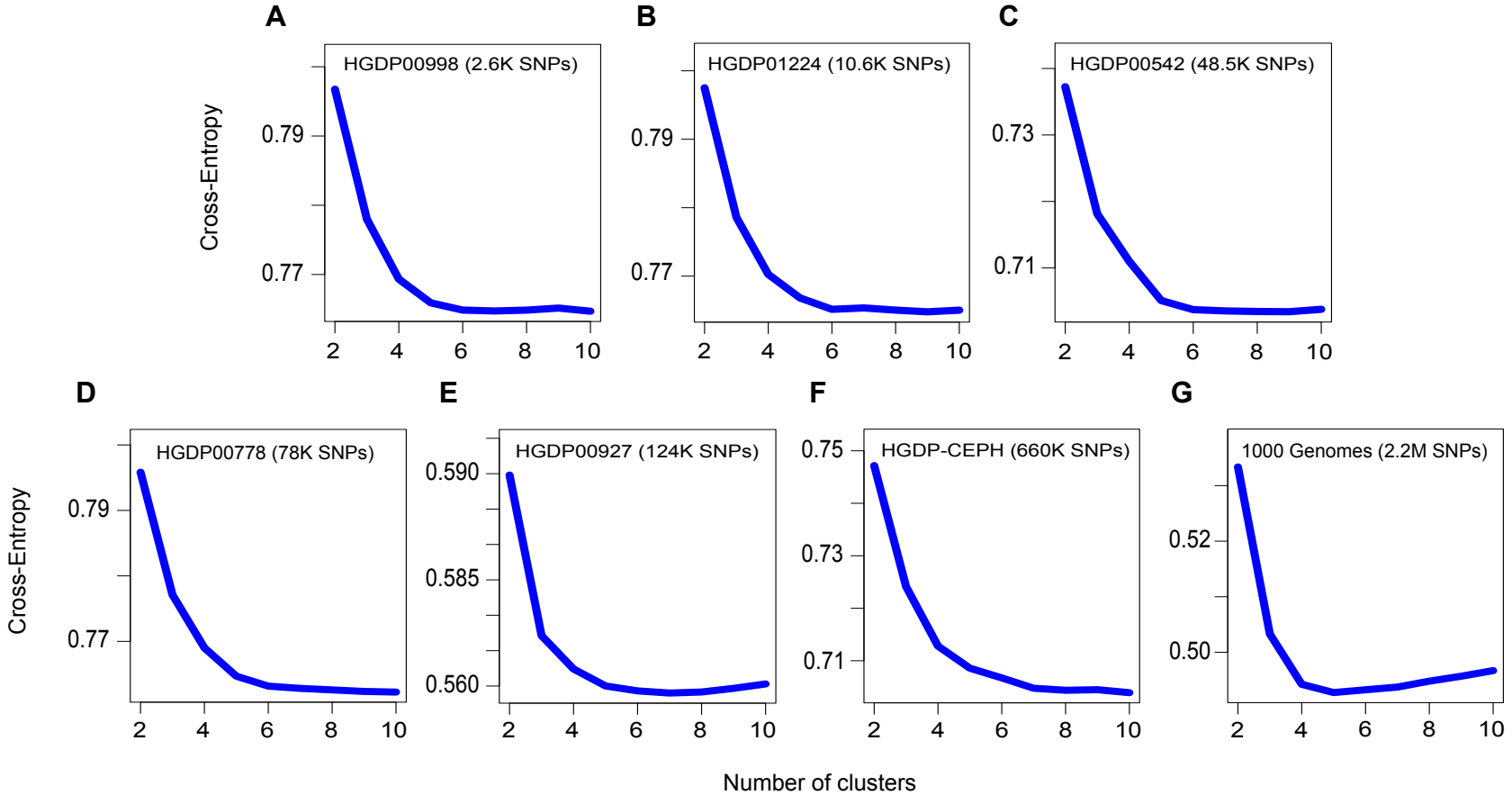
Table 3: **Statistical errors for ADMIXTURE and sNMF on simulated data sets.**

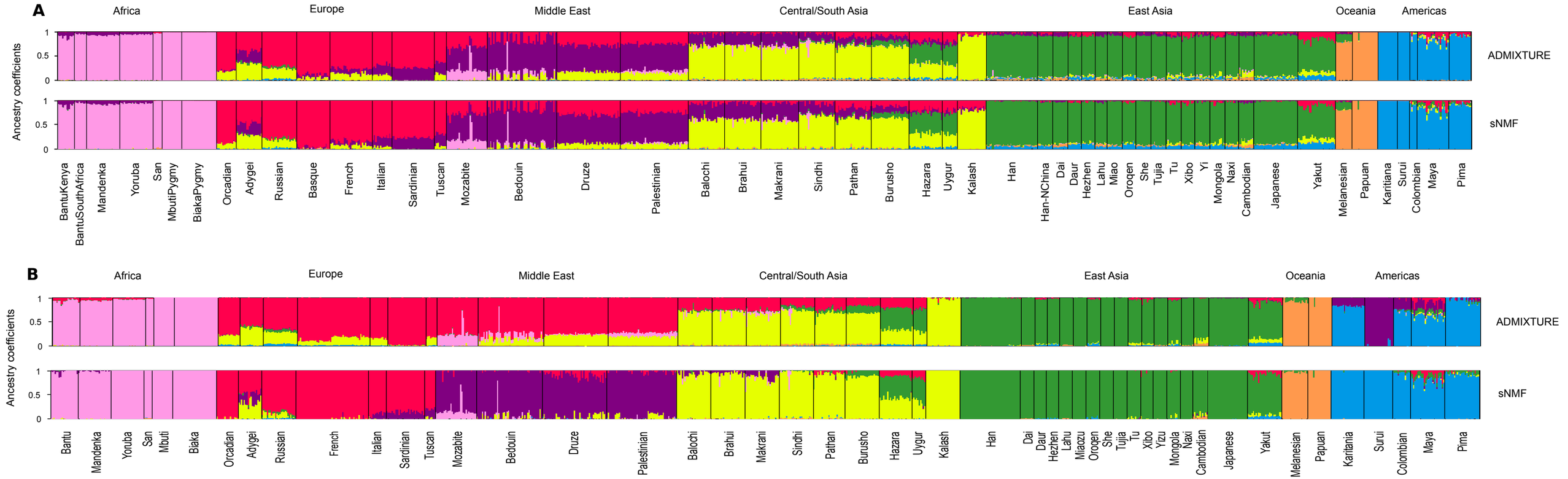
	Dir(1, 1, 1)	Dir(.5, .5, .5)	Dir(.1, .1, .1)	Dir(.2, .2, .05)	Dir(.2, .2, .5)	Dir(.05, .05, .01)
Q-matrix						
ADMIXTURE	0.023	0.012	0.004	0.006	0.010	0.003
sNMF $\alpha = 0$	0.020	0.011	0.007	0.007	0.013	0.009
sNMF $\alpha = 100$	0.024	0.014	0.006	0.006	0.014	0.006
G-matrix						
ADMIXTURE	0.029	0.022	0.016	0.022	0.022	0.022
sNMF $\alpha = 0$	0.034	0.027	0.021	0.028	0.028	0.028
sNMF $\alpha = 100$	0.034	0.027	0.021	0.028	0.028	0.028

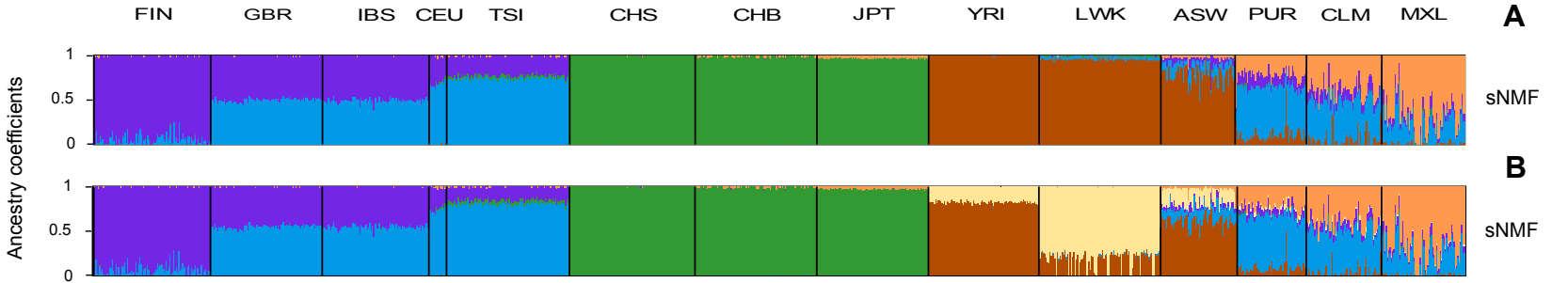
Dir: Dirichlet distribution used to simulate "true" admixture coefficients using 3 ancestral populations.

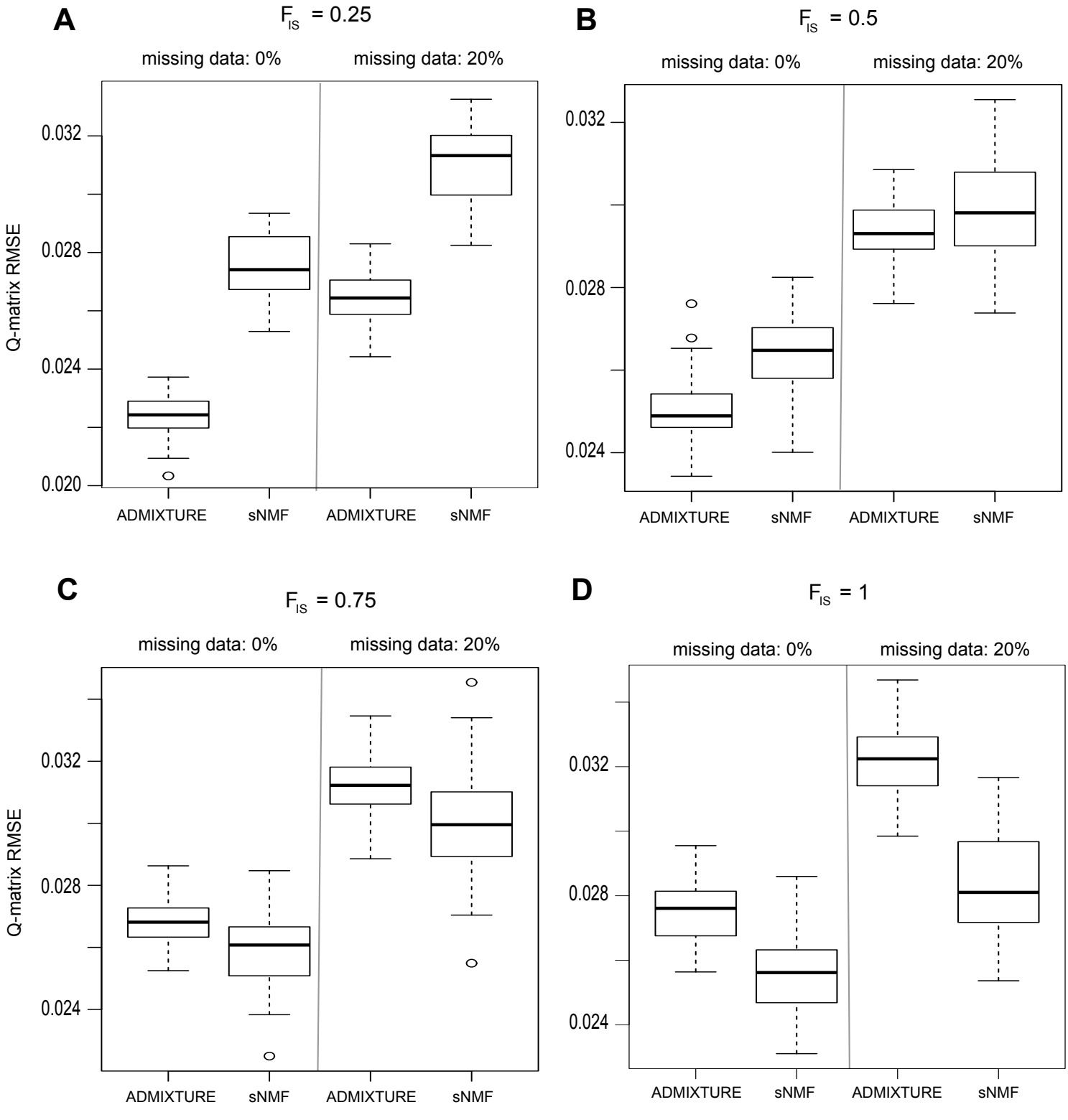

1.00
0.98
0.96
0.94
0.92
0.90
Coefficient of determination (R-squared)
0
1
10
100
1000
Regularization parameter


A
HGDP01224 (10.6K SNPs)
B
HGDP00778 (78K SNPs)
C
HGDP-CEPH (660K SNPs)
ADMIXTURE
sNMF
Run-time (hour)
Number of clusters


A
HGDP00998 (2.6K SNPs)
B
HGDP01224 (10.6K SNPs)
C
HGDP00542 (48.5K SNPs)
D
HGDP00778 (78K SNPs)
E
HGDP00927 (124K SNPs)
F
HGDP-CEPH (660K SNPs)
G
1000 Genomes (2.2M SNPs)
Cross-Entropy
Number of clusters


A
Africa
Europe
Middle East
Central/South Asia
East Asia
Oceania
Americas
Ancestry coefficients
1
0.5
0
ADMIXTURE
sNMF
BantuKenya
BantuSouthAfrica
Mandenka
Yoruba
San
MbutiPygmy
BiakaPygmy
Orcadian
Adygei
Russian
Basque
French
Italian
Sardinian
Tuscan
Mozabite
Bedouin
Druze
Palestinian
Balochi
Brahui
Makrani
Sindhi
Pathan
Burusho
Hazara
Uygur
Kalash
Han
Han-NChina
Dai
Daur
Hezhen
Lahu
Miao
Oroqen
She
Tujia
Tu
Xibo
Yi
Mongola
Naxi
Cambodian
Japanese
Yakut
Melanesian
Papuan
Karitiana
Surui
Colombian
Maya
Pima
B
Africa
Europe
Middle East
Central/South Asia
East Asia
Oceania
Americas
Ancestry coefficients
1
0.5
0
ADMIXTURE
sNMF
Bantu
Mandenka
Yoruba
San
Mbuti
Biaka
Orcadian
Adygei
Russian
French
Italian
Sardinian
Tuscan
Mozabite
Bedouin
Druze
Palestinian
Balochi
Brahui
Makrani
Sindhi
Pathan
Burusho
Hazara
Uygur
Kalash
Han
Dai
Daur
Hezhen
Lahu
Miaozu
Oroqen
She
Tujia
Tu
Xibo
Yizu
Mongola
Naxi
Cambodian
Japanese
Yakut
Melanesian
Papuan
Karitania
Surui
Colombian
Maya
Pima


FIN
GBR
IBS
CEU
TSI
CHS
CHB
JPT
YRI
LWK
ASW
PUR
CLM
MXL
A
B
sNMF
sNMF
Ancestry coefficients
1
0.5
0
1
0.5
0


A
F_{IS} = 0.25
missing data: 0%
missing data: 20%
Q-matrix RMSE
0.032
0.028
0.024
0.020
ADMIXTURE
sNMF
ADMIXTURE
sNMF
B
F_{IS} = 0.5
missing data: 0%
missing data: 20%
0.032
0.028
0.024
ADMIXTURE
sNMF
ADMIXTURE
sNMF
C
F_{IS} = 0.75
missing data: 0%
missing data: 20%
Q-matrix RMSE
0.032
0.028
0.024
ADMIXTURE
sNMF
ADMIXTURE
sNMF
D
F_{IS} = 1
missing data: 0%
missing data: 20%
0.032
0.028
0.024
ADMIXTURE
sNMF
ADMIXTURE
sNMF