1 Materials and Methods

Objectives and input data. Similar to previous versions of the software, the new version `tess 3` runs statistical analyses for geographical samples of multi-locus genotypes. The principal output of `tess 3` consists of estimates of ancestry coefficients for each individual in the sample. The program also computes locus-specific estimates of ancestral genotypic frequencies, and return estimates of population-based differentiation statistics that can be used to perform genome scans for adaptive alleles.

The genotypic matrix entries record allelic data for each individual (i) and each locus (ℓ). `tess 3` requires data that consists of n multi-locus genotypes and their geographic coordinates. The `tess 3` program is particularly suited to the analysis of large genomic data sets, for which the number of loci (L) range between thousands to millions of genetic polymorphisms. For example, this is the case with data representing single nucleotide polymorphisms (SNPs). Here the genotypic matrix records the number of derived or mutant alleles at each locus. Considering autosomes in a diploid organism, each genotype at locus ℓ corresponds to the number of derived alleles at this locus, and it is coded as an integer number 0, 1 or 2.

Geographically constrained least-squares estimates of ancestry proportions. Ancestry estimation models suppose that the genetic data originate from the admixture of ancestral populations, where the number of ancestral population, K, is unknown. Following standard notations, Q_{ik} denotes the fraction of individual i's genome that originates from the ancestral population k, and $G_{k\ell}(j)$ represents the frequency of genotype j at locus ℓ in population k. The Q-matrix, $Q = (Q_{ik})$, denotes the $n \times K$ matrix of individual ancestry coefficients, whereas the G-matrix, $G = (G_{k\ell}(j))$, denotes, the $K \times ML$ matrix of ancestral genotype frequencies, where M is the number of genotypes ($M = 3$ for diploid organism SNPs).

The program `tess 3` computes a nearest neighbor graph for the sampling sites, and provides least-square estimates of ancestry proportions by using graph regularized matrix factorization algorithms (Frichot et al. 2014, Cai et al. 2010). Estimates of the Q and G matrices are obtained after solving the following constrained least-squares problem:

$$\mathrm{LS}(Q, G) = \|X - QG\|_{\mathrm{F}}^2 + \alpha \sum_{s_i \sim s_j} w_{ij} \|Q_{i\cdot} - Q_{j\cdot}\|^2 , \qquad (1)$$

where $\|M\|_{\mathrm{F}}$ denotes the Frobenius norm of a matrix M (Berry et al. 2007), $\|V\|$ is the Euclidean norm of a vector V, α is a non-negative *regularization parameter*. In equation (1), the summation at the right-hand side of the second term runs over all pairs of sides, $s_i \sim s_j$, that share a common edge in the graph, and the w_{ij}'s are weights that decrease with geographic distance between sampling sites, (s_i, s_j).

$$w_{ij} = e^{-d(s_i,s_j)^2/\bar{d}^2} , \tag{2}$$

where $\bar{d}$ is the average pairwise distance between geographic samples (Durand et al. 2009).

For α greater than zero, `tess 3`performs a *graph regularized* non-negative matrix factorization algorithm for the data matrix X (Cai et al. 2011, Lee and Seung 1999). Testing values α greater than zero can reduce the variance of Q and G estimates, and forces individuals that are geographically close to each other to share ancestry from the same source populations (Durand et al. 2009). `tess 3` have a default parameter for α to have same order of magnitude for the data term and the regularization term, but it can be determined by using a cross-validation criterion (Alexander et al.). α can also be determine by looking at the resulting Q and choosing it when one obtain suitable spatially continuous Q.

Estimates of the Q and G matrices are computed using the Alternating Non-negativity-constrained Least Squares (ANLS) algorithm with the active set (AS) method following the method used in the program `sNMF` (Frichot et al. 2014, Kim and Park 2011). The ANLS-AS algorithm begins with the initialization of the Q matrix and computes a non-negative matrix G that minimizes the quantity

$$\|X - QG\|_F^2$$

The obtained solution is normalized so that its entries satisfy equation (2). Given G, we compute the Q matrix that minimizes the quantity

$$\left\| \begin{pmatrix} Vec(X^T) \\ 0 \end{pmatrix} - \begin{pmatrix} Id \otimes G^T \\ \sqrt{\alpha}\ \Gamma \otimes Id \end{pmatrix} Vec(Q^T) \right\|_F^2 \tag{3}$$

where Vec denotes the vectorization of the matrix X formed by stacking the columns of X into a single column vector, Γ is the Cholesky decomposition of L the graph laplacian [1]. With this loss function formulation we can apply same algorithm that for G computation. Iterations are stopped when

algorithm are in a local minimum. Thus we verify that two successive values of the loss function is less than a tolerance threshold of $\epsilon = 10^{-7}$.

Outlier Detection Tests. In addition to the inference of population structure, the program `tess 3` can perform genome scan for adaptive alleles based on ancestral frequency differentiation statistics. More specifically, `tess 3` uses the ancestral genotype frequency matrix to derive allele frequencies in K ancestral populations. Then the program algorithm combines ancestral frequencies with Q-matrix estimates to assess locus-specific population differentiation indices

$$F_{\mathrm{ST}} = 1 - \frac{\sigma_{\mathrm{S}}^2}{\sigma_{\mathrm{T}}^2},$$

where σ_{S}^2 is computed as $\sigma_{\mathrm{S}}^2 = \sum_k q_k f_k (1 - f_k)$, σ_{T}^2 is computed as $\sigma_{\mathrm{T}}^2 = (\sum_k q_k f_k)(1 - \sum_k q_k f_k)$, and q_k is the mean value of Q_{ik} over all individuals in the sample.

Using standard theory (Weir), we transformed the population differentation indices into t-scores, and p-values were corrected for confounding due to neutral population structure using a graphically determined correction factor (Devlin and Roeder 1999). Multiple testing issues were addressed by applying the Benjamini-Hochberg algorithm to the re-calibrated P-values with an expected level false discovery rate equal to $q = 0.05$ (Benjamini and Hochberg 1998).

Simulated Data Sets. ICI Simulated data sets contained 200 admixed genotypes with levels of ancestry that varied accross geographic space. We used the computer program `ms` to perform coalescent simulations of neutral and outlier SNP loci under models of structured populations (2-island models, Hudson, 2002). The principle underlying the use of neutral demographic processes for simulating local selection is that loci with selection coefficient s, have an effectively reduced migration rate, m_{s}, approximated as $m_{\mathrm{s}} = m/s$ (Bazin et al. 2010).

Considering two source populations under a migration-drift equilibrium model, we set the neutral migration rate to the value $mN = 20$. The number of loci was varied in the range $L = 10^3 - 10^5$, and the proportion of outlier loci was 10% or 5%. Outlier loci were generated using the value $m_{\mathrm{s}}N = 2$, which corresponded to the value $s = 10$ for the selection coefficient.

The sample size was set to $n = 200$, 100 genotypes were sampled from each source population, and admixed genotypes were then created according to a longitutinal gradient of ancestry. Individuals at the each extreme of the longitunal range were representative of ancestral populations, while individuals at the center of the range shared intermediate levels of ancestry in the two source populations. This simulation setting mimicks samples from an admixed population after secondary contact (Durand et al. 2009).

Simulated data were analyzed using `tess 3` and `tess 2.3` (Durand et al. 2009). To enable comparisons, the Q-matrix outputs for each program were permuted using `CLUMPP` (Jakobsson et al. 2007). Statistical errors were measured as the root mean squared error (RMSE) between the matrices Q^3 (Q^2) obtained from `tess 3` (`tess 2`) and the true matrix of coefficients (Q^0) used to generate the data.

$$\text{RMSE} = \left(\frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} (q_{ik}^{i} - q_{ik}^{0}) \right)^{1/2} .$$

A similar RMSE criterion was defined for comparing the estimates of G^3 and G^2 obtained from `tess 3` and `tess 2`.

Number of ancestral populations. Runs of `tess` were performed for values of the number of ancestral populations ranging from $K = 1$ to 6, For `tess 3`, 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. The number of ancestral populations, K, was chosen after the evaluation of the cross-entropy criterion for each K. The evaluation procedure partitions the genotypic matrix entries into a training set and a test set, for which the known genotypes masked to the program. Each value of K corresponds to a distinct model of the data, and the cross-entropy criterion evaluates the capability of each model to correctly predict the masked values (Frichot et al. 2014). Smaller values of the criterion indicate better algorithm outputs and estimates.

Arabidopsis thaliana **data.** In addition to our simulated data set, we considered genomic data from the model plant *Arabidopsis thaliana* genotyped for 205,406 SNPs (Atwell et al. 2010). We focused our example on the study of 49 accessions from Scandinavia, and considered ecological gradients linked to temperature by extracting 11 variables from the WorldClim database at

each of the 49 sampling sites (annual mean temperature, mean diurnal range, temperature seasonality, etc). We summarized the 11 variables as a unique ecological factor by computing the first principal component of the temperature variables.

2 Results

Comparison of ancestry estimates for simulated data. First of all, we evaluate the ability of spatial NMF to reproduce result of Tess on simulated data. We created admixed population of two source population under a migration-drift model with different value of Nm as described in section simulated Data Sets. We got a data set with 200 individuals and 2 000 SNPs and ran algorithms with K = 2 ancestral clusters. All other parameters was set to default values. For each run we computed the statistical error for each method (see figure 1). We note that RMSE increase with the migration parameter (from around 5% to around 12%). The results provide evidence that spatial NMF and Tess have close outcomes. Moreover, we can see that, for this particular simulation, spatial NMF Q estimation is a bit better than Tess one. This comes from the simulation which was created using spatially continuous Q-matrix. Therefore the regularization part of (1) enforce better estimation of Q-matrix by spatial NMF algorithm. It is also important to note that Tess and spatial NMF program was used with default parameters (except for the number of ancestral cluster K). Thus it may be possible to have better result if we would have tuned parameters. This task is harder for Tess than for spatial NMF which only have two adjustable parameters and only the number of cluster is mandatory to run the algorithm.

Run-time analysis Next we performed run time analyses for Tess and spatial NMF, using same kind of simulated data set that in the previous part. For Tess the total number of sweeps of MCMC is an important parameter for runtime analyses. We noticed that for this particular simulation scenario 1000 sweeps was enough to have steady estimations. The runtimes were averaged over distinct random seed values of K and L. We see that for both algorithms the runtime increase with the number of SNPs and the number of clusters. For 10 000 SNPs both algorithms take less than 0.15 hours and spatial NMF is only around 3 times faster than Tess for 10 000 SNPs. But for a dataset with 100 000 SNPs spatial NMF is around 30 times faster than

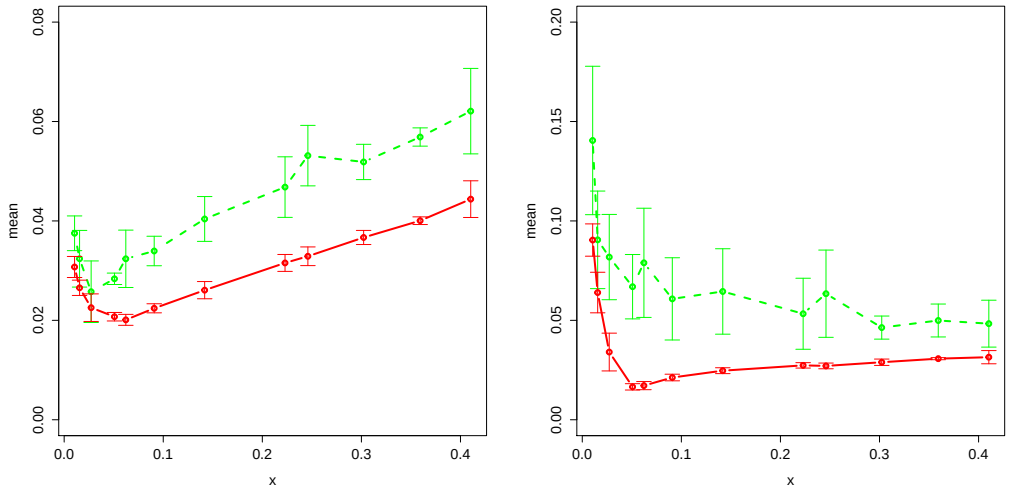

Figure 1: Statistical error of spatial NMF (green) and Tess (red) on simulated data sets with increasing migration rate between ancestral populations. Each data set had 200 individuals and 2 000 SNPs. Tess and spatial NMF was run with K = 2 ancestral populations and default parameters.

Tess, for example with 5 clusters Tess took in average 1.2 hours whereas for spatial NMF the average runtime was 0.03 hours.

False discovery rate of local adaptation with Fst statistic In this serie of experiment we study the usefulness of Fst () to detect local adaptation. We simulated neutral loci and selected loci as described in section simulated Data Sets. We performed an outlier detection test (describe in section) and compute observed false discovery rate and observed proportion of outlier detected (the recall). Figure ... show us we there is a batch of outlier loci which is easy to detect: with an expected Fdr of 0.05 we detect about 60% for the green simulation and 10% for the red simulation. However it is not as easy to detect other outliers, and so the recall rise slowly with the expected Fdr. Indeed with those simulations there are outlier p-values which have same distribution that null p-values. Thus there are hard to detect when we control the false discovery rate. Nevertheless, those simple experiments show us this Fst statistic is useful to detect outlier loci in an admixed population.

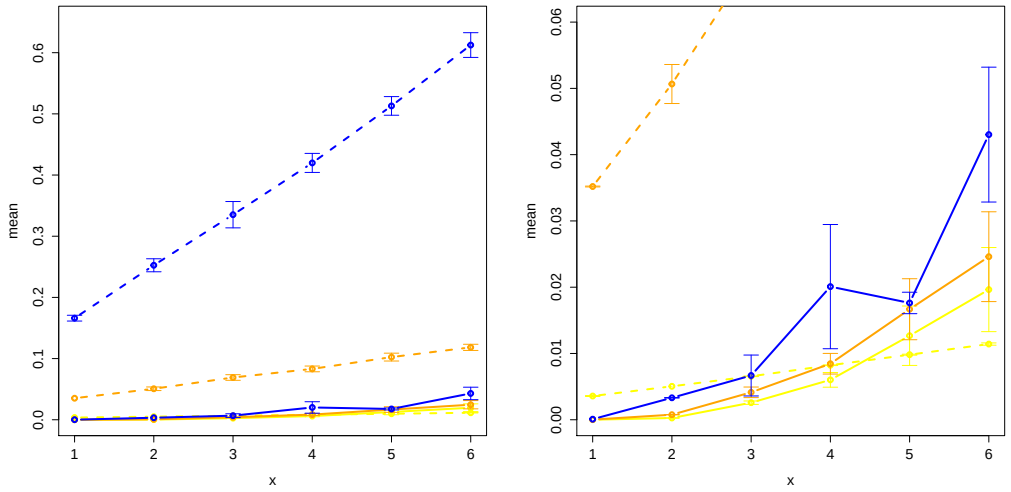

Figure 2: Runtimes for spatial NMF and Tess on Intel Xeon 2.40 GHz CPU. Time is expressed in unit of hours. We ran both methods with 200 individuals for L equal 1000, 10 000 and 100 000 SNPs (from the palest to the darkest green for spatial NMF and from the palest to the darkest red for Tess) and K from 1 to 10.

Arabidopsis Thaliana To finish we run spatial NMF on Arabidopsis Thaliana (ref) in order to compare with other studies on this dataset. The figure ... show that with 3 ancestral clusters we have the usual ancestral coefficients (ref). We also ran Tess on this dataset to compare results. The RMSE between Tess and spatial NMF coefficient was equal to 9.8%. This is a good error in comparison with order we got on simulated data (see fig1). Moreover Tess runtime was 2 hours whereas spatial NMF runtime was five minutes. Then we computed Fst (eq) with this result to perform local adaptation detection.

References

[1] Fan RK Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.

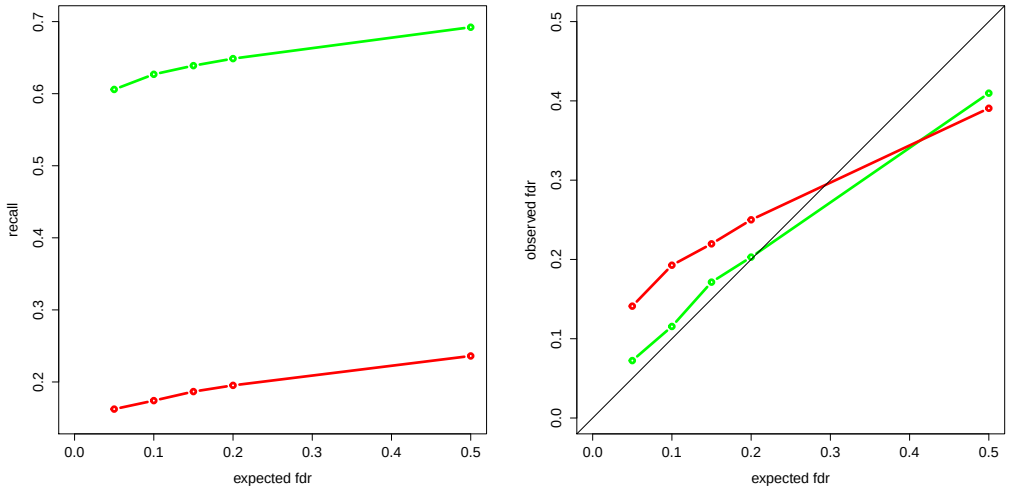

Figure 3: Local adaptation test on simulated data. We plotted the expected false discovery rate (Fdr) against the observed false discovery rate and the proportion of retrieved loci under local adaptation (recall). Each data set was composed of 200 individual and 30 000 loci. (add legend)

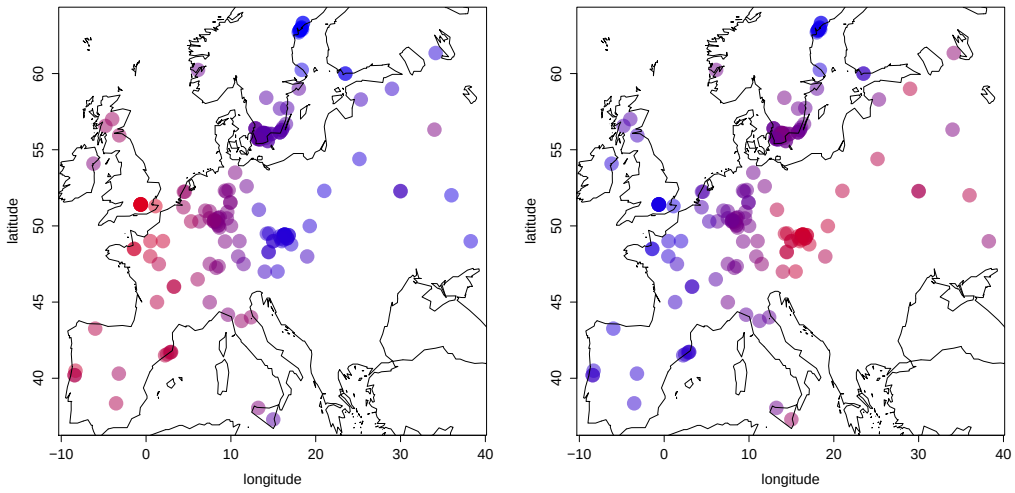

Figure 4: Estimation of ancestry coefficients of Arabidopsis Thaliana data set (170 individuals and 216 130 SNPs) for K = 3 ancestral cluster. ...