tess3r: a R package for estimating spatial population structure

Jobim 2016

Kevin Caye[1], Olivier Michel[2], Olivier Francois[1]

[1] TIMC-IMAG, [2] GIPSA-lab

June 28, 2016

Genotypic Data

DNA Sequencing Technologies:

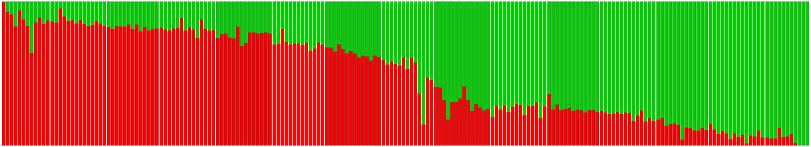
- SNPs array (*Arabidopsis thaliana* RegMap lines [Horton et al., 2012] : 200*k* loci of 1 307 individuals)
- next generation sequencing (1000 Genome project [Consortium et al., 2015] : whole genome of 2504 individuals)

	chr: 1 pos: 657	chr: 1 pos: 3102	chr: 1 pos: 4648
02B6	1	1	1
09A3	1	0	1
12A1	1	1	1
13B5	0	0	0

Goal: Estimating Ancestral Population Structure

We assume that the genome of each individual come from K ancestral populations. We want to estimate:

- ancestral genotype frequencies for each locus
- ancestral coefficients for each individual



Several method exist:

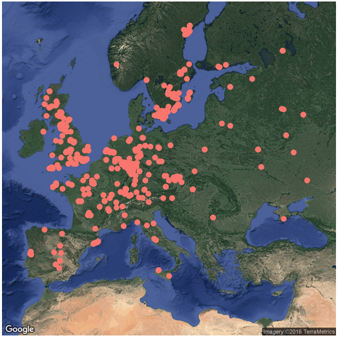
- Bayesian: structure [Pritchard et al., 2000]
- optimisation based: sNMF [Frichot et al., 2014]

These methods are crucial for demographic analysis, medical genetics, conservation genetics or landscape genetics.

Spatial Data

Several methods incorporate geographic data in prior distributions:

- TESS 2.3 [Durand et al., 2009]
- BAPS [Corander et al., 2008]



We propose a new method to estimate spatial population based on an optimizarion problem.

Model for the Genotypic Matrix

We write G the genomic matrix

$$\begin{aligned} P(G_{i,\ell} = j) &= \textstyle\sum_{k=1}^{K} Q_{i,k} f_{k,\ell}(j), \\ P &= QF^T, \end{aligned}$$

where Q is the ancestry coefficient matrix and F the ancestral genotype frequency matrix. There are such as:

$$Q \succeq 0, \ \sum_{k=1}^{K} Q_{i,k} = 1, \ \forall i \in \{1, ..., n\}$$

$$F \succeq 0, \ \sum_{j=0}^{d} f_{k,\ell}(j) = 1, \ \forall \ell \in \{1, ..., L\}.$$

Optimisation Problem

We construct a weighted graph using spatial data:

$$W_{i,j} = e^{-\frac{||z_i - z_j||^2}{\sigma}},$$

where z are geographic positions.
The loss function is:

$$\begin{aligned} Loss(Q,F) = & \quad ||X - QF^T||^2 + \lambda \sum_{i,j}^{n} W_{i,j} ||Q_{i,} - Q_{j,}||^2 \\ = & \quad ||X - QF^T||^2 + \lambda \text{trace}(Q^T \Lambda Q), \end{aligned}$$

where Λ is the graph Laplacian matrix and X is the data matrix.

Block-coordinate Descent Scheme

- The optimisation problem is not convex.
- It is convex with respect to one of the variables Q or F when the other one is fixed.
- We can use a block-coordinate descent scheme:

```
for $it \in 1, .., itMax$ do
    $F \leftarrow \arg\min_{F} f_F(Q, F)$
    $Q \leftarrow \arg\min_{Q} f_Q(Q, F)$
end for
```

Alternating Projected Least Squares

F optimisation step :

1. Solving least squares problems for F rows
2. Projecting onto the polygon of F constraints

Q optimisation step :

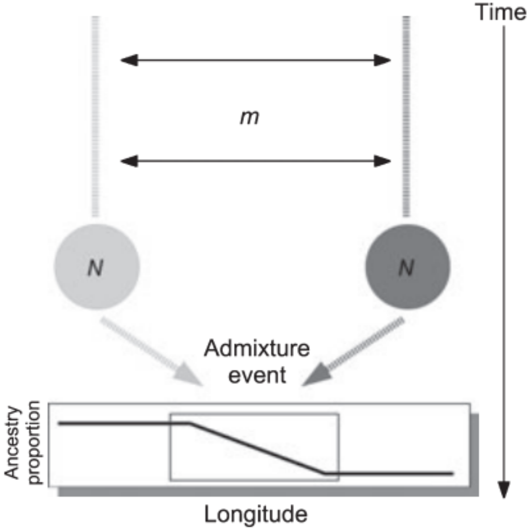
1. Solving L_2-regularized least squares problems for Q rows
2. Projecting onto the polygon of Q constraints

Alternating Projected Least Squares

- Advantage:
  The F optimisation step requires to solve L least squares problems of size K.
  The Q optimisation step requires to solve n L_2-regularized least squares problems of size K.
- Drawback:
  The algorithm is not guaranteed to provide a stationary point.

Simulations

Simulation of admixed population dataset [François and Durand, 2010]



Comparison with TESS 2.3

On these simulation TESS3 algorithm performs 30 time better than TESS 2.3

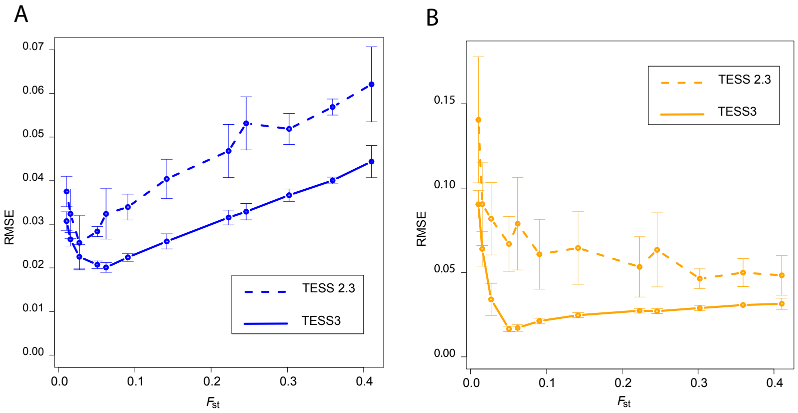
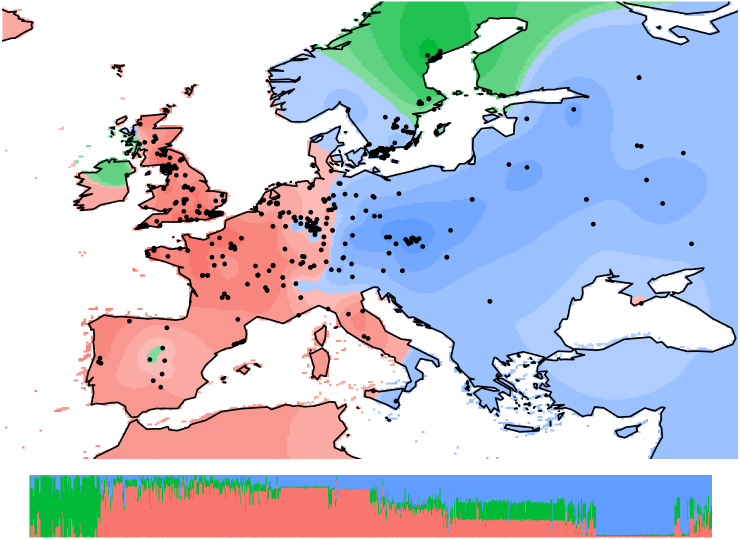


Figure 1: A) RMSEs of F estimates. B) RMSEs of Q estimates.

Arabidopsis thaliana RegMap Lines Dataset

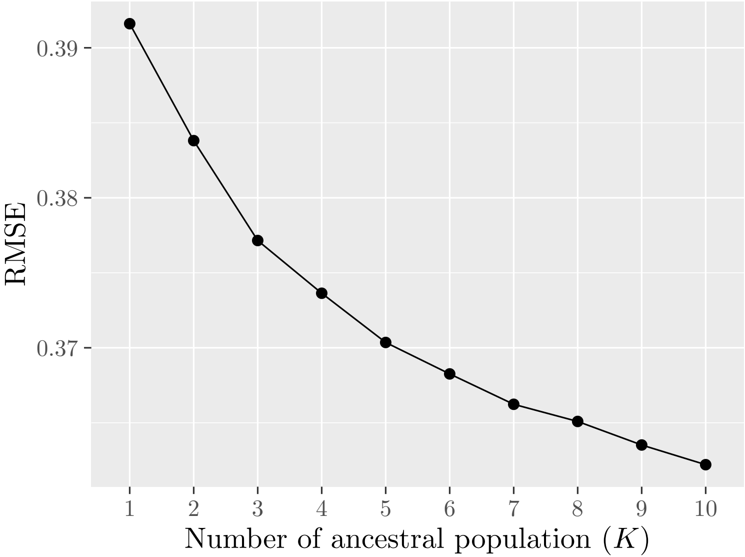
- 1,307 accessions of *A. thaliana* have been genotyped using the Affymetrix Arabidopsis 250k - SNP chip [Horton et al., 2012]
- Geographic coordinates for 1,193 of these samples [Anastasio et al., 2011]

Ancestry Coefficients with $K = 3$ Ancestral Populations



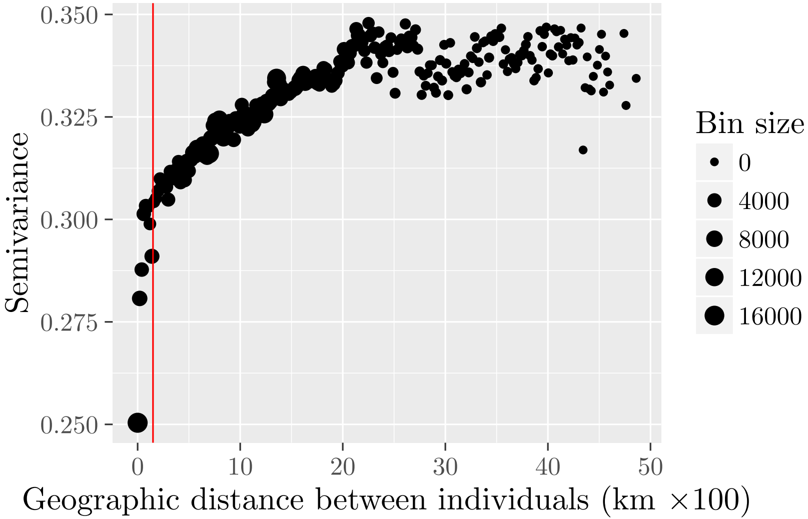
Selection of the Number of Ancestral Populations

We choose $K = 6$ ancestral populations.

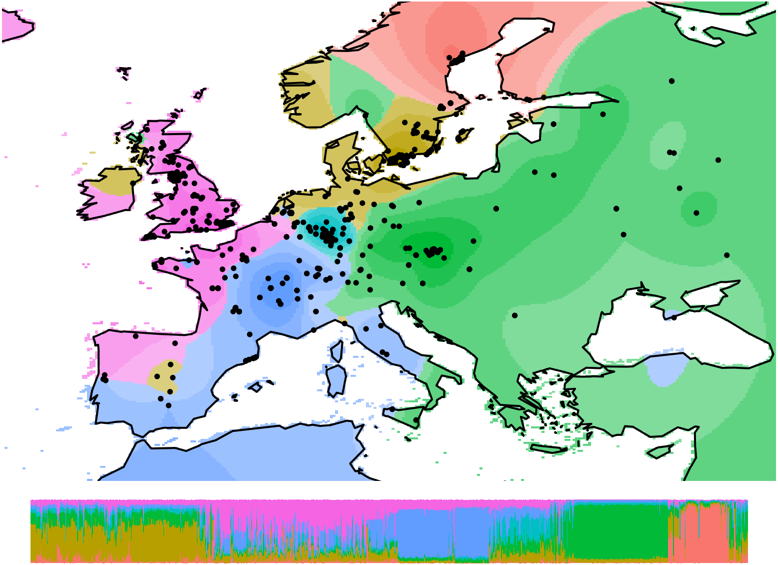


Selection of the Spatial Autocorrelation Scale

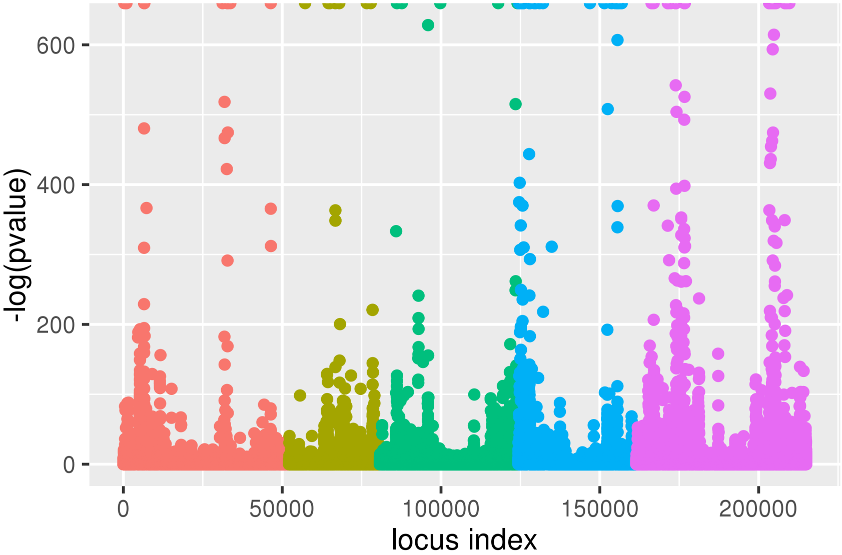
We choose $\sigma = 1.5$ for the spatial autocorrelation scale.



Ancestry Coefficients for $K = 6$ and $\sigma = 1.5$



Ancestral Frequency Manhattan Plot for $K = 6$ and $\sigma = 1.5$



Detecting Adaptive Loci (Preliminary Results)

- Flowering-related genes detected in the top list: FRIGIDA, FLOWERING LOCUS C (FLC), DELAY OF GERMINATION 1 (DOG1) [Horton et al., 2012].
- Statistical overrepresentation of genes involved in metabolic processes: 1.08 fold enrichment, p-value $< 1e^{-6}$ (PANTHER database).

Conclusion

- Estimation and visualisation of spatial population structure.
- Local adaptation detection.
- Beta Version available on github.

```
devtools::install_github("cayek/TESS3_encho_sen@master")
```

Acknowledgments

- Timo Deist, Eric Frichot, Olivier François, Olivier Michel
- This Ph.D is funded by the labex Persyval-lab

Thank you for your attention.

Alison E Anastasio, Alexander Platt, Matthew Horton, Erich Grotewold, Randy Scholl, Justin O Borevitz, Magnus Nordborg, and Joy Bergelson. Source verification of mis-identified arabidopsis thaliana accessions. *The Plant Journal*, 67(3): 554–566, 2011.

1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

Jukka Corander, Jukka Sirén, and Elja Arjas. Bayesian spatial modeling of genetic population structure. *Computational Statistics*, 23(1):111–129, 2008.

Eric Durand, Flora Jay, Oscar E Gaggiotti, and Olivier François. Spatial inference of admixture proportions and secondary contact zones. *Molecular Biology and Evolution*, 26(9):1963–1973, 2009.

Olivier François and Eric Durand. Spatially explicit bayesian clustering models in population genetics. *Molecular Ecology Resources*, 10(5):773–784, 2010.

Eric Frichot, François Mathieu, Théo Trouillon, Guillaume Bouchard, and Olivier François. Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4):973–983, 2014.

Matthew W Horton, Angela M Hancock, Yu S Huang, Christopher Toomajian, Susanna Atwell, Adam Auton, N Wayan Muliyati, Alexander Platt, F Gianluca Sperone, Bjarni J Vilhjálmsson, et al. Genome-wide patterns of genetic variation in worldwide arabidopsis thaliana accessions from the regmap panel. *Nature genetics*, 44(2):212–216, 2012.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.