PERSYVAL-Lab

UNIVERSITÉ GRENOBLE ALPES

Spatially regularized NMF for population genetic applications

Kevin Caye
Université de Grenoble Alpes

Outline

› Method to estimate individual ancestry coefficients from population genetic and spatial data

› Graph regularized non-negative matrix factorization

› Alternating least squares algorithm

› Results

Genotypic data

› Single nucleotide polymorphism (SNP)
  – single nucleotide variation occurring commonly within a population

Ind 1 .....AAGC C TA.....

⋮

Ind n .....AAGC T TA.....

› Data matrix: L loci for n individuals ($n \sim 10^2 - 10^3$, $L \sim 10^6 - 10^7$)

Our data

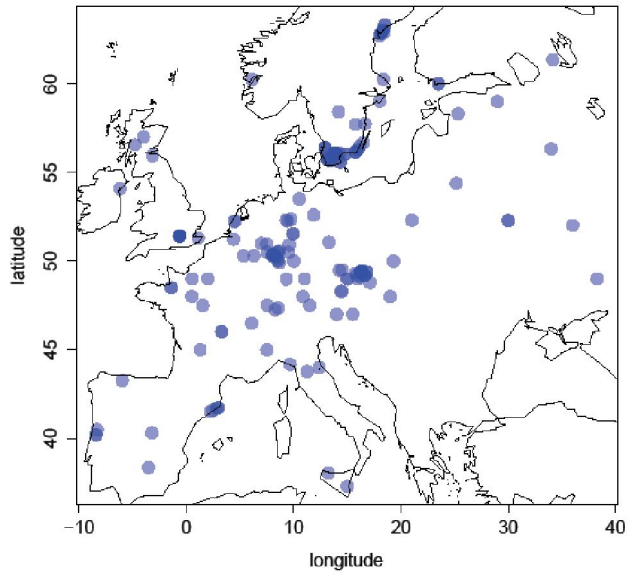
› Genotypic matrix for diploid individuals: number of mutations observed for each individual and locus (0, 1 or 2)

$$G = \begin{pmatrix} 1 & \cdots & 0 \\ \vdots & X_{i,l} & \vdots \\ 2 & \cdots & 1 \end{pmatrix}$$

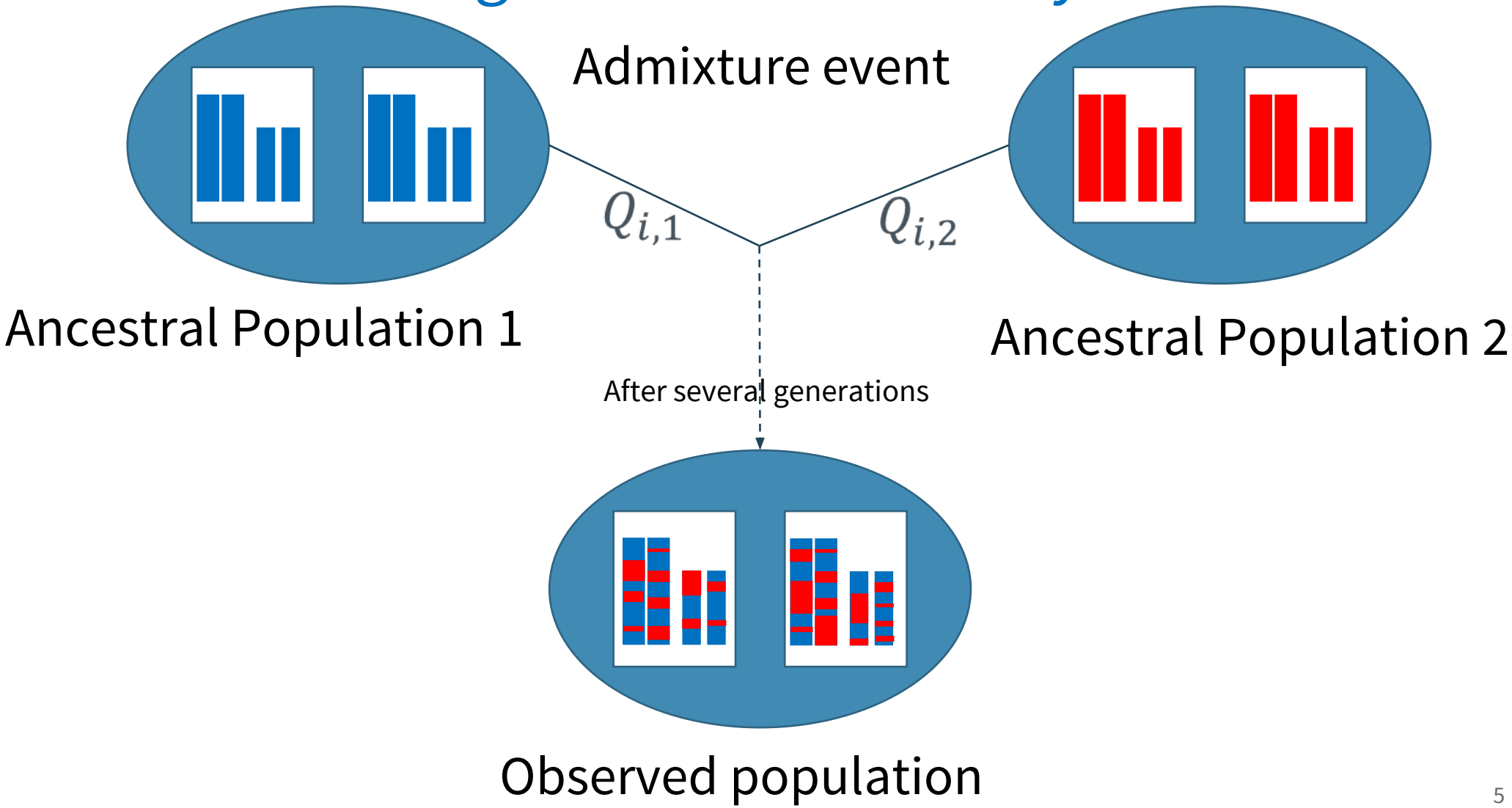
n ind

L loci

› Geographic data for each individual



Goal: Estimating individual ancestry coefficients



Definition of ancestry coefficients

› We assume there are K ancestral populations (K unknown)

› The observed allele frequencies are a convex combination of ancestral frequencies

$$P(G_{i,l}= j) = \sum_{k=1}^{K} Q_{i,k} F_{k,l}(j)\,, \qquad \forall i, l, j$$

$Q_{i,k}$ = the fraction of individual i's genome that originates from ancestral population k

State of the art

› Estimation of ancestry coefficients without spatial information:

  – Bayesian method: Structure (Pritchard et al. 2000)

  – sparse NMF: sNMF (Frichot et al. 2014)

› With spatial information:

  – Bayesian method: Tess (Durand et al. 2009)

Non negative matrix factorisation

- X : zero/one values depending on the absence or the presence of each genotype at each locus

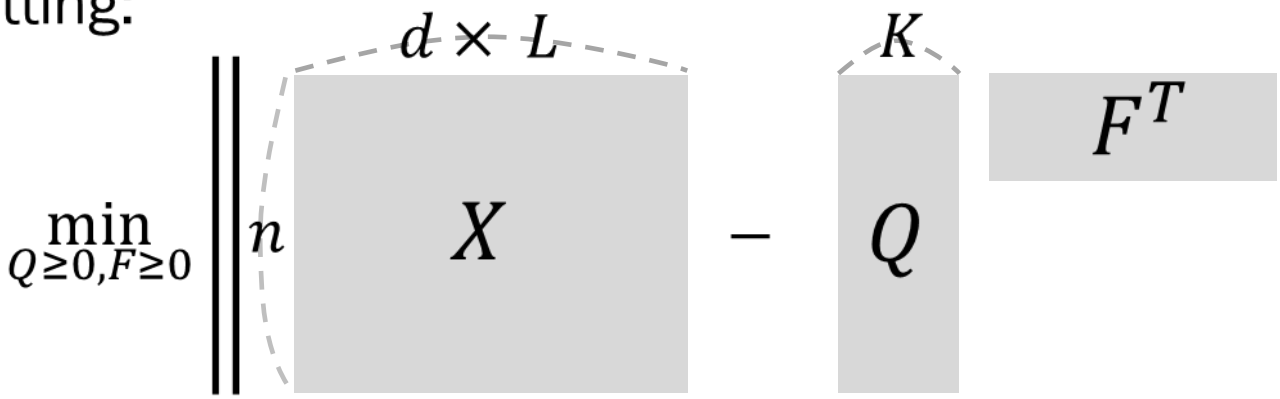
$$X_{i,d\times l+j} \sim Bernoulli\left(P(G_{i,l} = j)\right)$$

Donc

$$X_{i,d\times l+j} \sim Bernoulli\left(\sum_{k=1}^{K} Q_{i,k} F_{k,l}(j)\right)$$

Non negative matrix factorisation

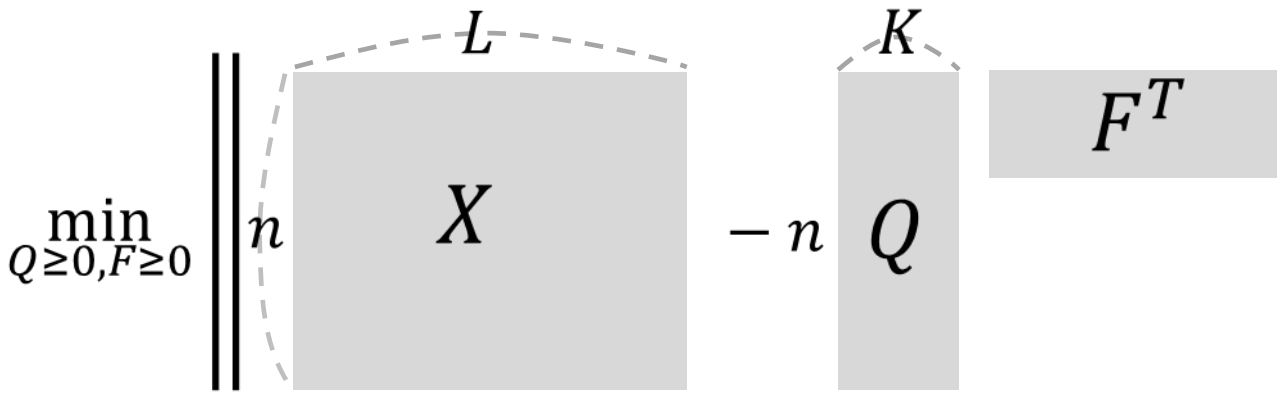
› Model fitting:



› Additional constraints

$$\sum_{k=1}^{K} Q_{i,k} = 1,\ \sum_{j=0}^{d} F_{l,k}(j) = 1, \quad \forall i, l, k$$

› sNMF (Frichot et al. 2014)

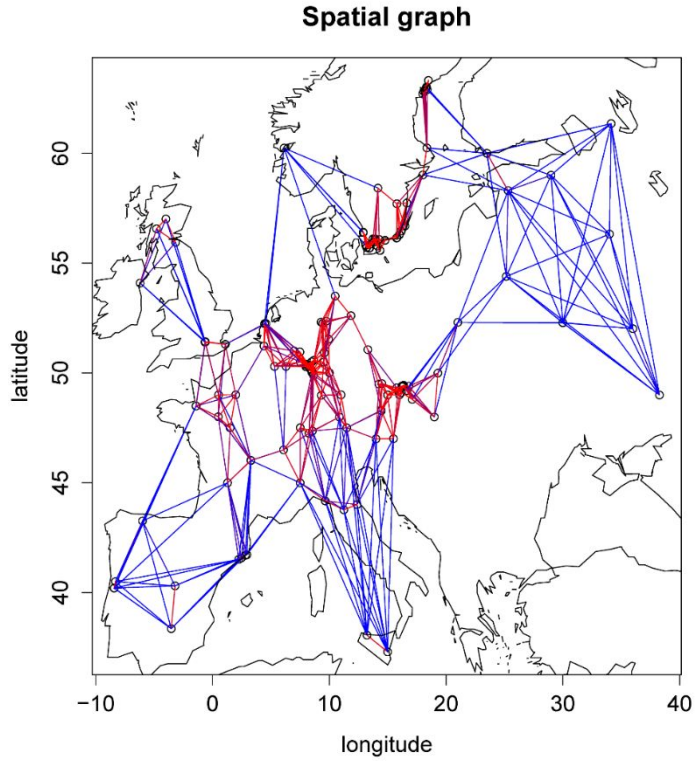

$\min_{Q \geq 0, F \geq 0}$
n
L
X
$- n$
K
Q
F^T

Least square minimization

› Graph regularized NMF (Cai et al. 2011)

$$\min_{Q\geq 0, F\geq 0} \|X - QF^T\|^2 + \alpha \frac{1}{2}\sum_{m,r}^{N} \|Q_{m,:} - Q_{r,:}\|^2 W_{m,r}$$

$W \in \mathbb{R}^{N\times N}$: weight coefficients



› Additional constraints

$$\sum_{k=1}^{K} Q_{i,k} = 1,\ \sum_{j=0}^{2} F_{l,k}(j) = 1, \quad \forall i, l, k$$

Our approach

› Rewriting the error functional as follows

$$\|X - QF^T\|^2 + \alpha\|\Gamma Q\|^2$$

$\Gamma \in \mathbb{R}^{N\times N}$: Cholesky decomposition of the graph Laplacian matrix

› Rewriting the error functional to use Alternating least squares

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

Numerical algorithm

- Alternating non-negativity-constrained least squares using the active set method (Kim and Park 2011)
- Computing F by solving

$$\min_{F \geq 0} \left\| X - QF^T \right\|^2$$

$$\sum_{j=0}^{2} F_{l,k}(j) = 1$$

- Computing Q by solving

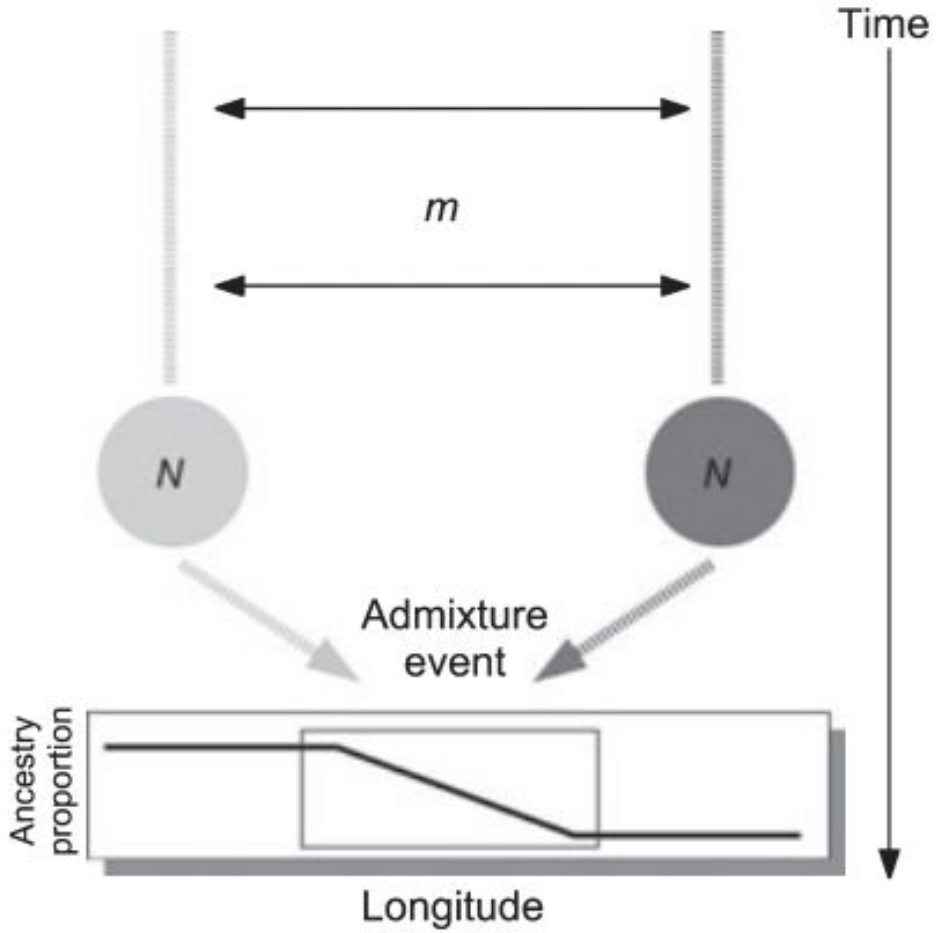
$$\min_{Q \geq 0} \left\| \begin{pmatrix} Vec(X^T) \\ 0 \end{pmatrix} - \begin{pmatrix} Id \otimes F \\ \sqrt{\beta}(\Gamma \otimes Id) \end{pmatrix} Vec(Q^T) \right\|^2$$

$$\sum_{k=1}^{K} Q_{i,k} = 1$$

Simulation study

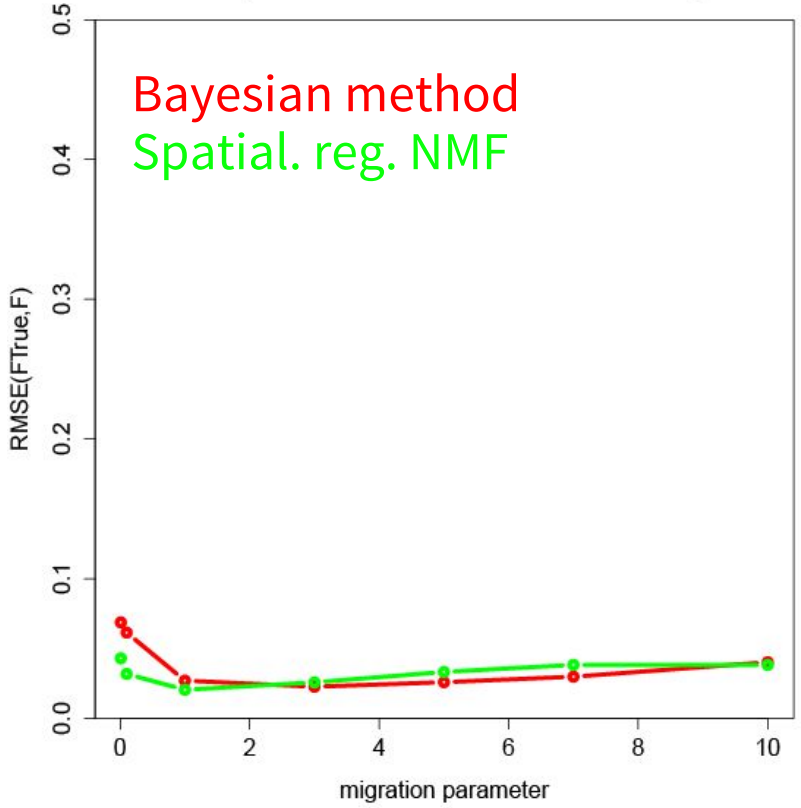
› Simulation of 2 populations with an admixture event

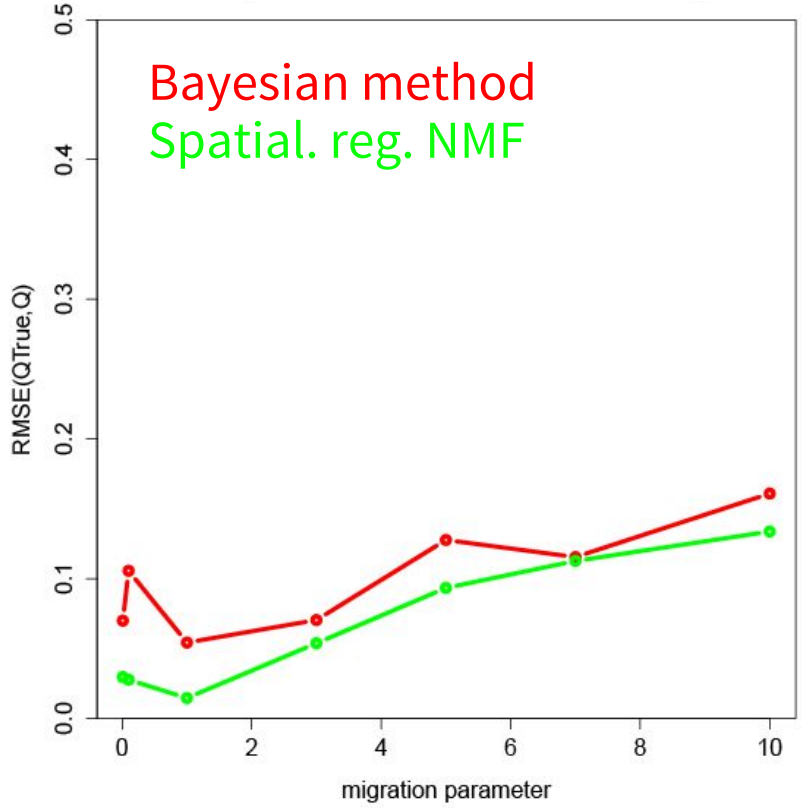
$$n = 200$$
$$L = 10^5$$



Statistical error comparison with Tess

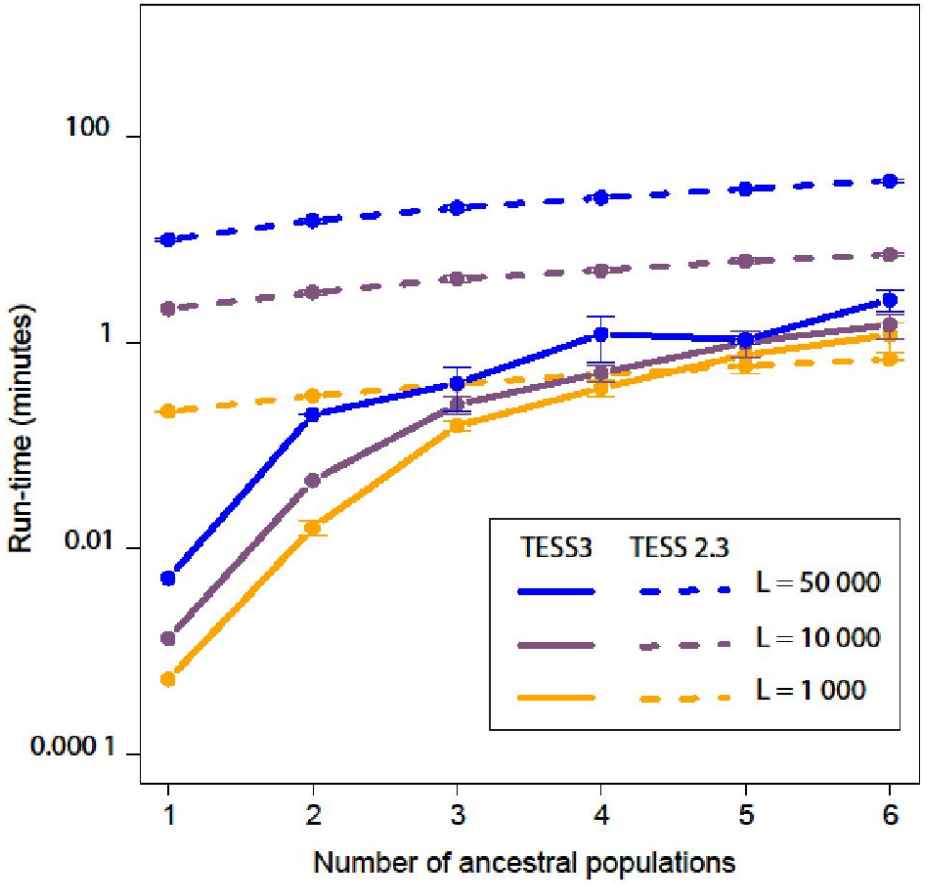
$$RMSE(Q^{TRUE}, Q) = \sqrt{\frac{1}{nK}\sum_{i,k}(Q_{i,k}^{TRUE} - Q_{i,k})^2}$$



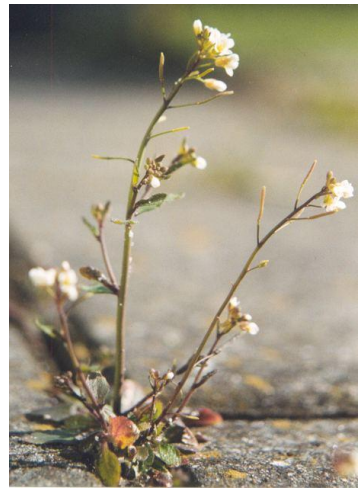


Benefit of the least square approach

› Run time analysis: TESS3 about 10-100 fold faster than TESS 2.3



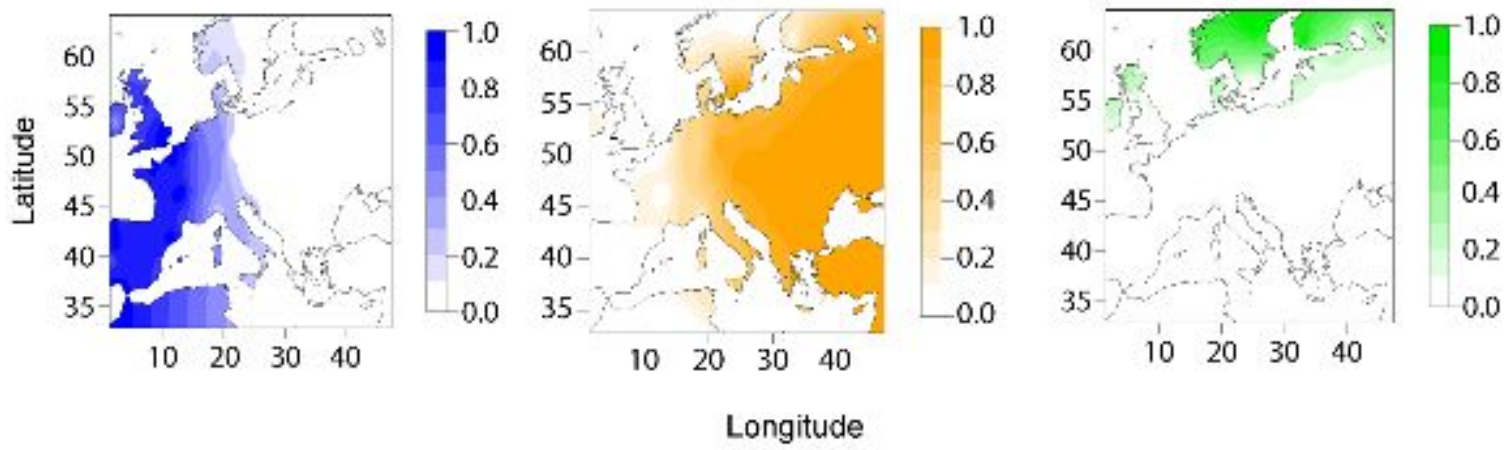
Analysis of *Arabidopsis thaliana* data



› Popular model organism in plant biology

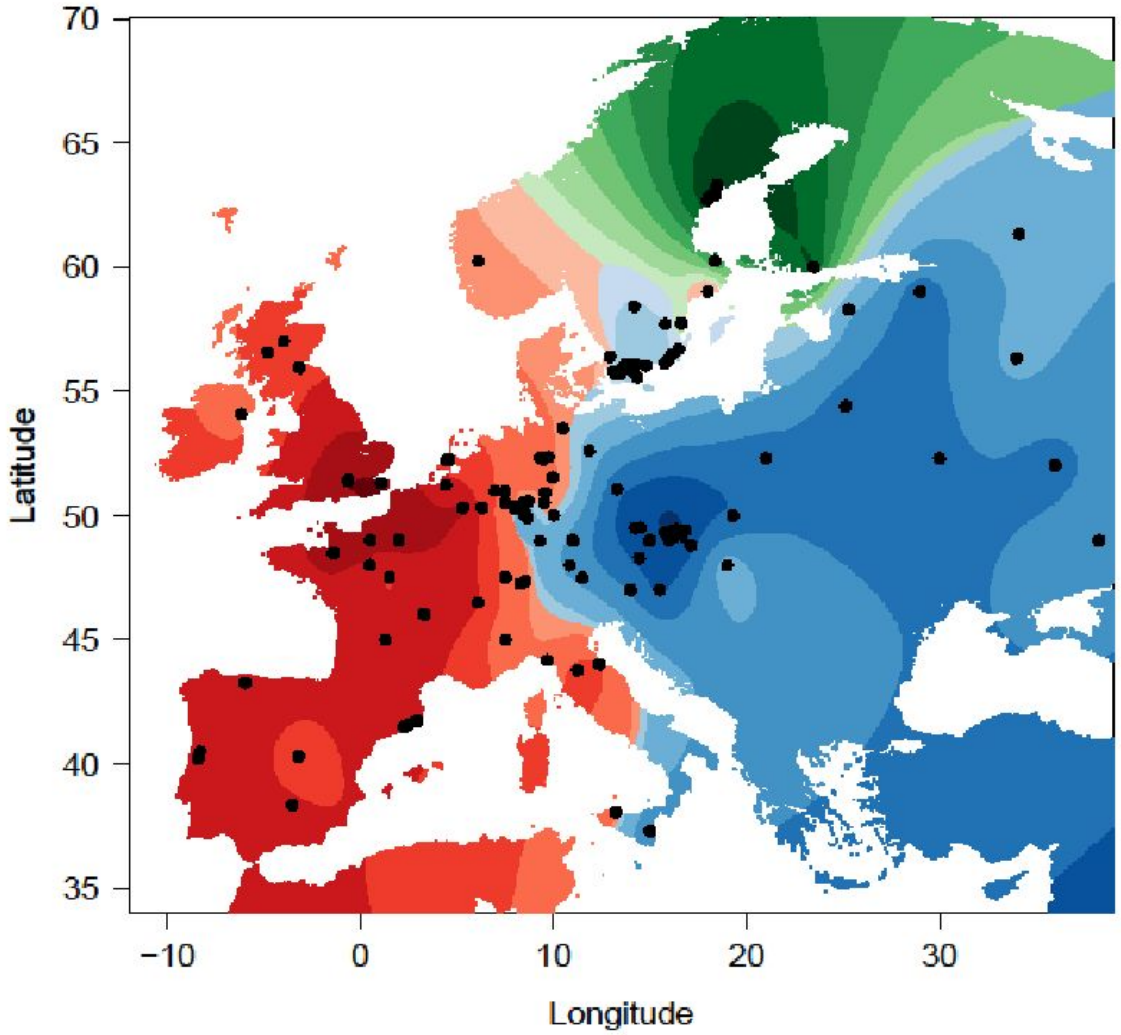
› 170 European individuals genotyped at 230 000 loci (Atwell et al. 2010)

› Three spatially consistent ancestral populations in Europe (Francois et al. 2008):

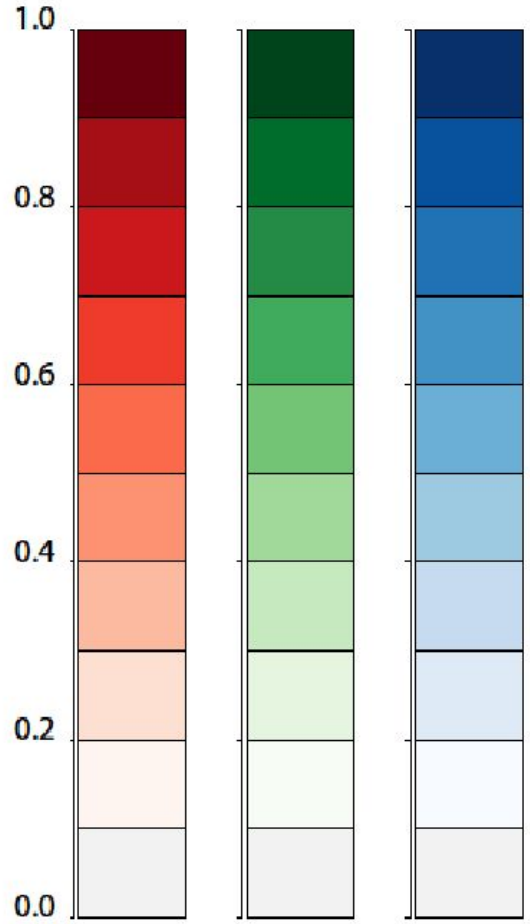


Analysis of *Arabidopsis thaliana* data

Arabidopsis thaliana Ancestry Coefficient Map with TESS3



Ancestry Coefficients



Discussion

› Graph regularized NMF combines spatial and genetic data

› We rewrote the problem of graph NMF to use ALS algorithm

› The algorithm is much faster than Bayesian methods

Acknowledgments

› Timo Deist, Eric Frichot, Olivier Francois, Olivier Michel

› This Ph.D is funded by the labex Persyval-lab