Spatially regularized NMF for population genetic applications

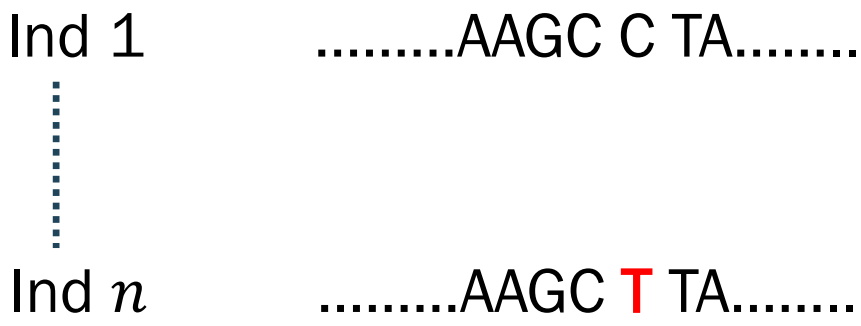
Kevin Caye
Université de Grenoble Alpes

Outline

- Method to estimate individual ancestry coefficients from population genetic and spatial data
- Graph regularized non-negative matrix factorization
- Alternating least squares algorithm
- Application to plant data

Genotypic data

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



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

Our data

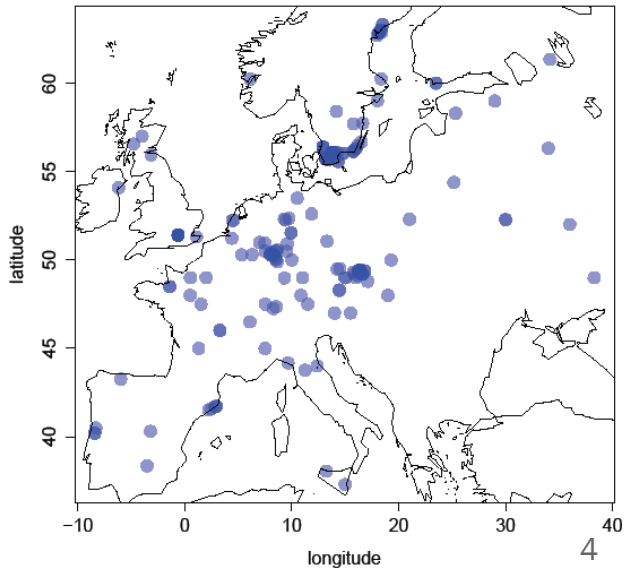
› Genotypic matrix for diploid individuals: number of mutations observed for each individual and locus (0, 1 or 2)

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

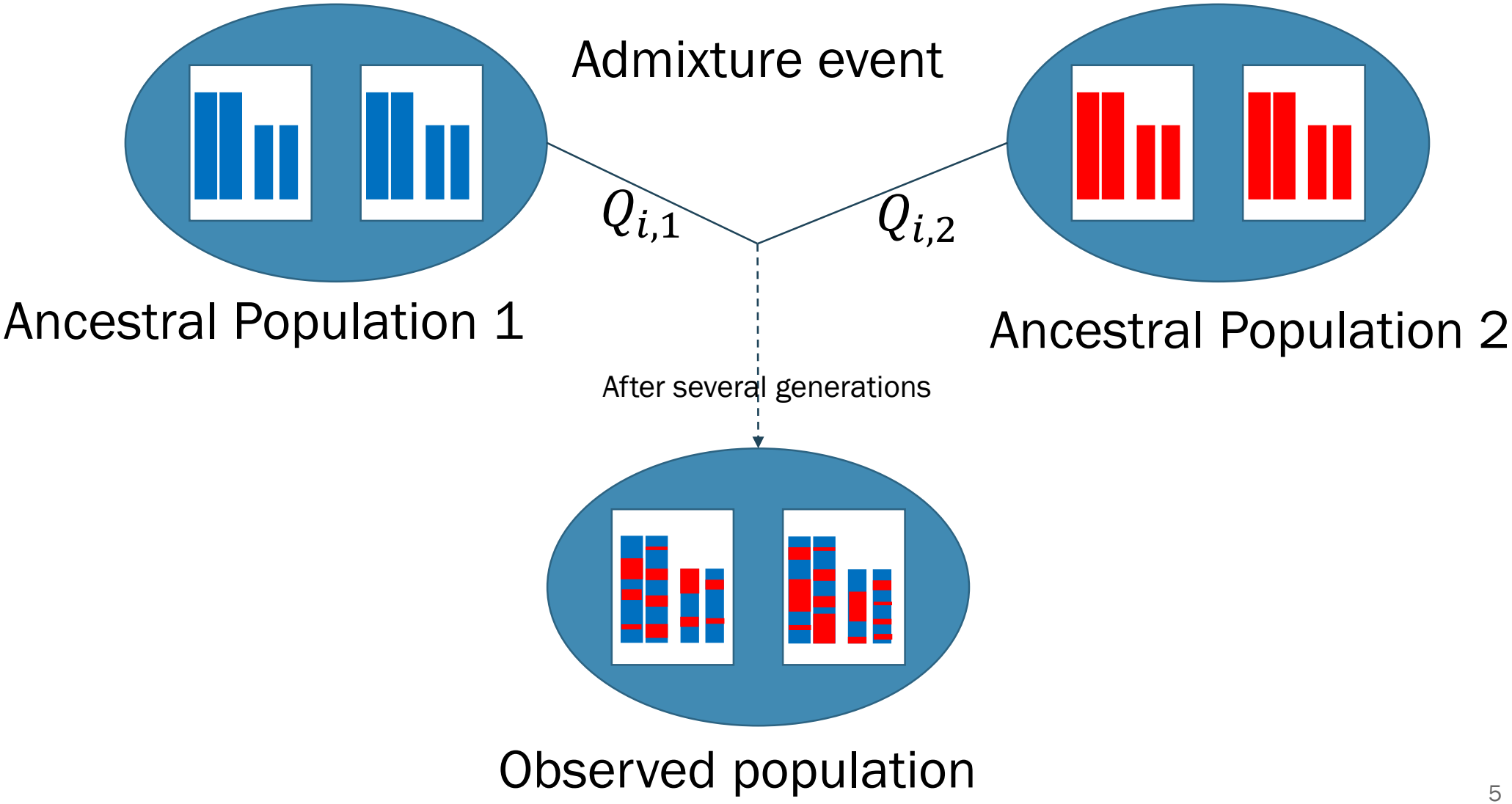
n ind

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

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

$Q_{i,k}$ = 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)

Least square minimization

› Graph regularized NMF (Cai et al. 2011)

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

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

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

› Where Γ is the Cholesky decomposition of the graph Laplacian matrix

Numerical algorithm

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

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

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

- Computing Q by solving

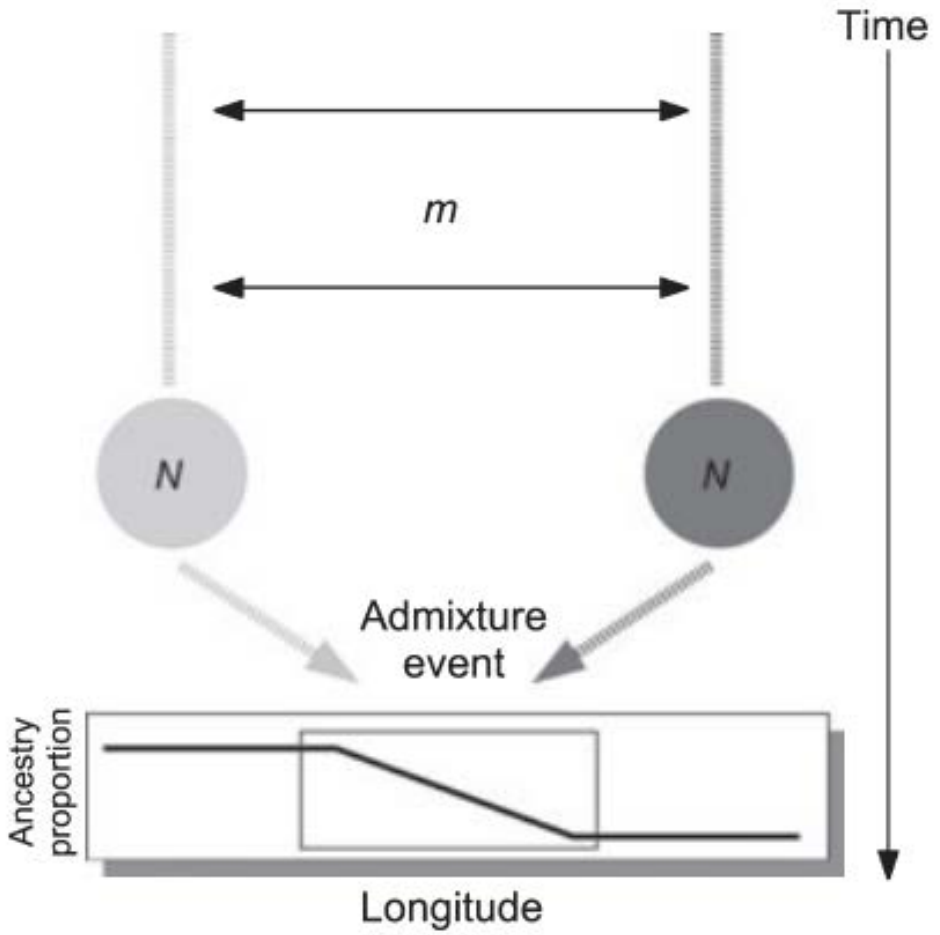
$$\min_{\mathrm{Q}\geq 0}\|X-QF^{T}\|^{2}+\ \alpha\|\Gamma Q\|^{2}$$

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

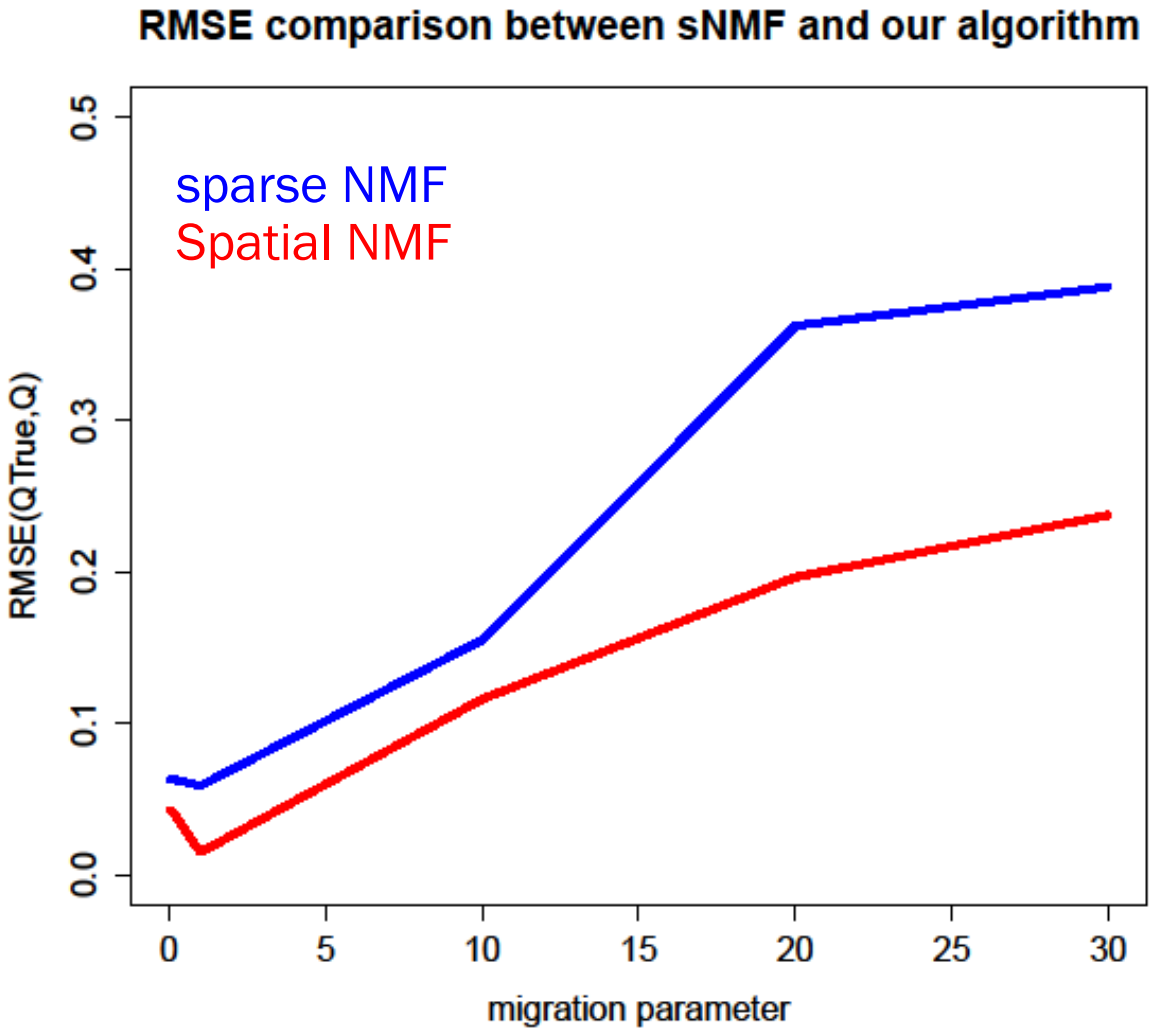
Simulation study

› Simulation of 2 populations with an admixture event

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



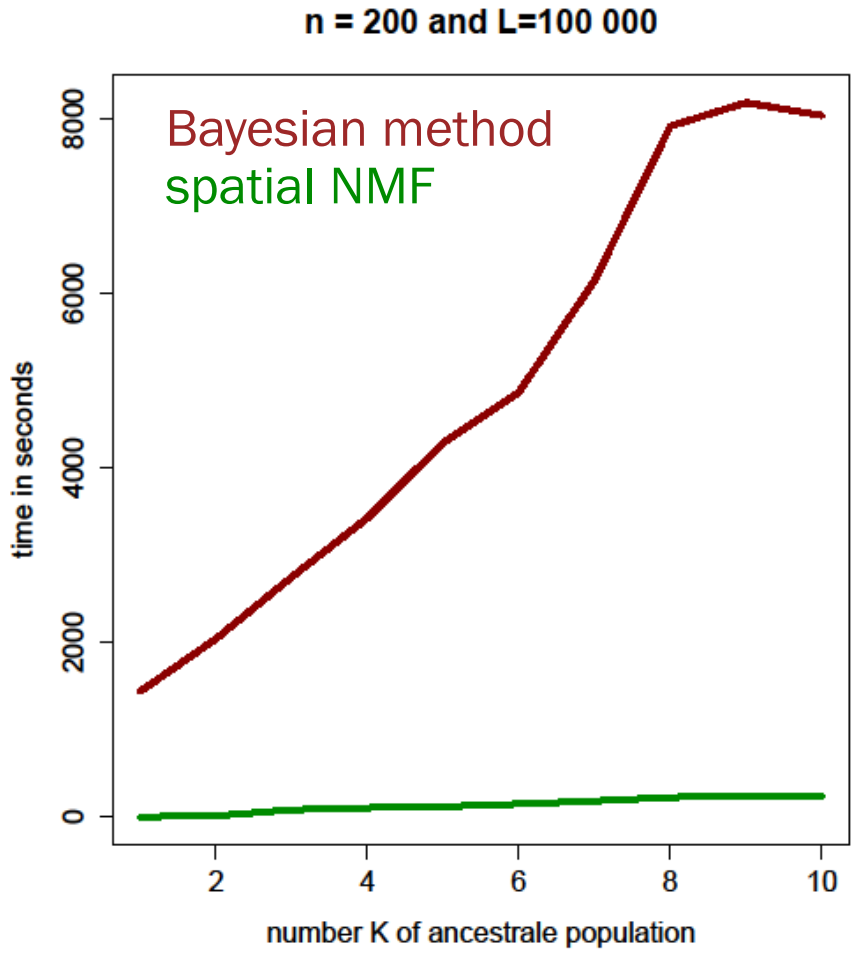
Benefit of including spatial information



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

Benefit of the least square approach

› Run time analysis: spatial NMF about 10-100 fold faster

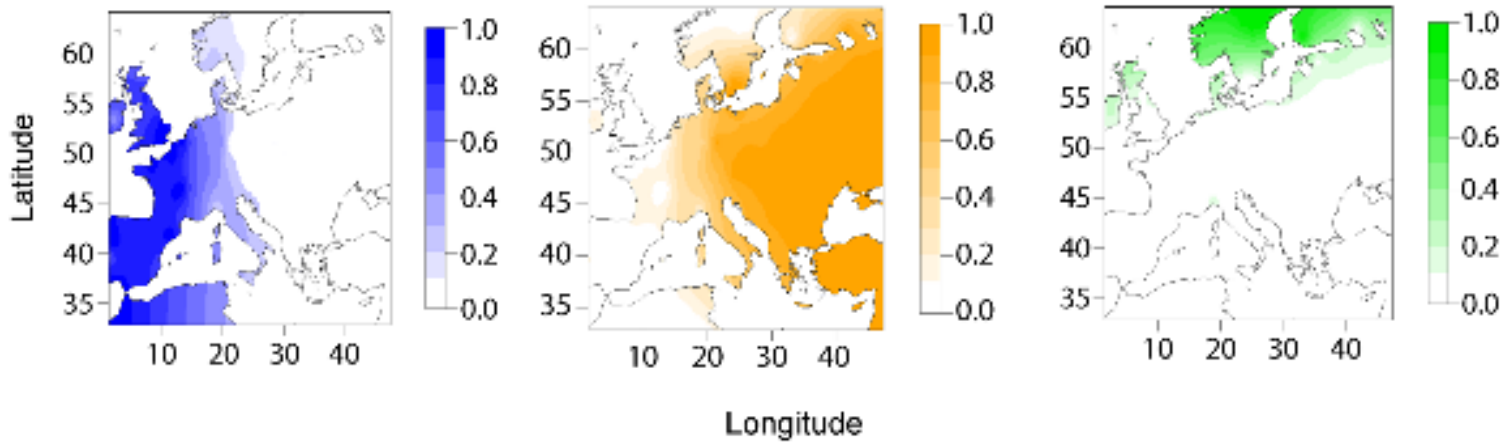


Analysis of *Arabidopsis thaliana* data

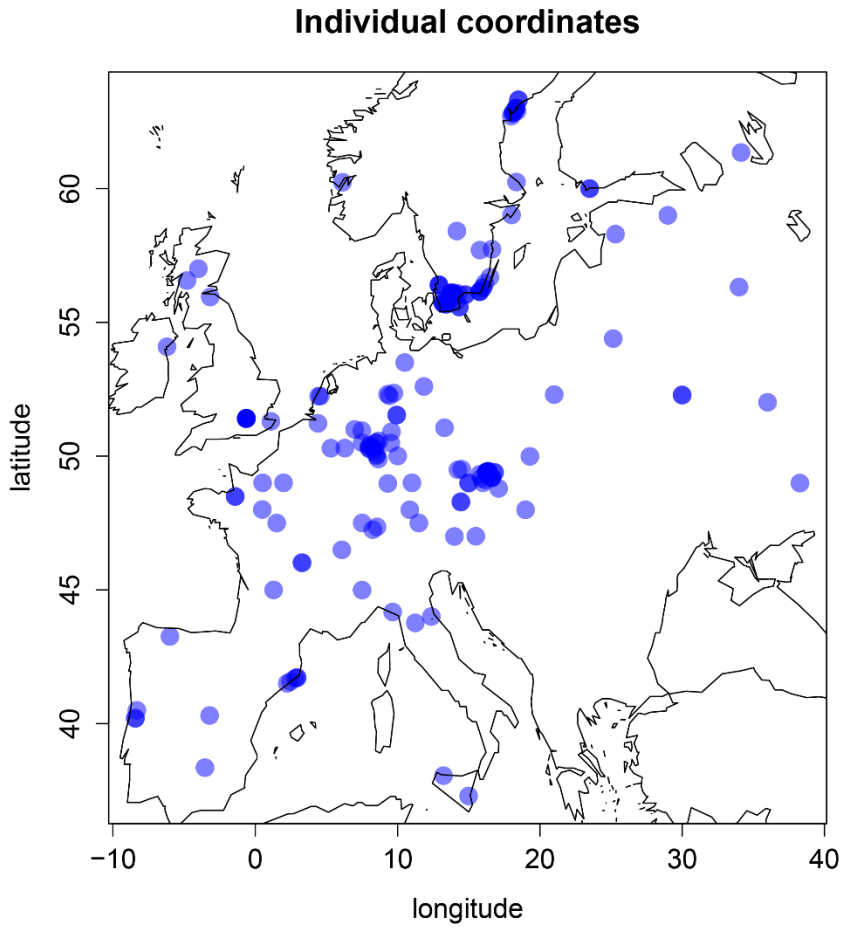
› Popular model organism in plant biology

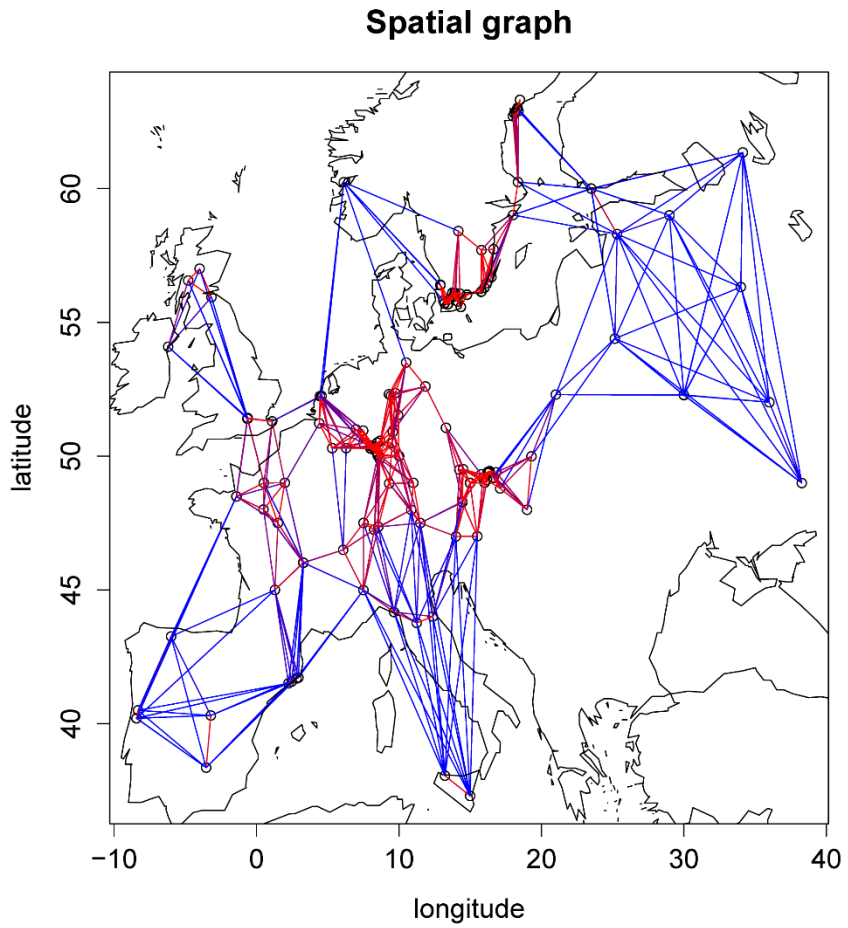
› 170 European individuals genotyped at 230 000 loci (Atwell et al. 2010)

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



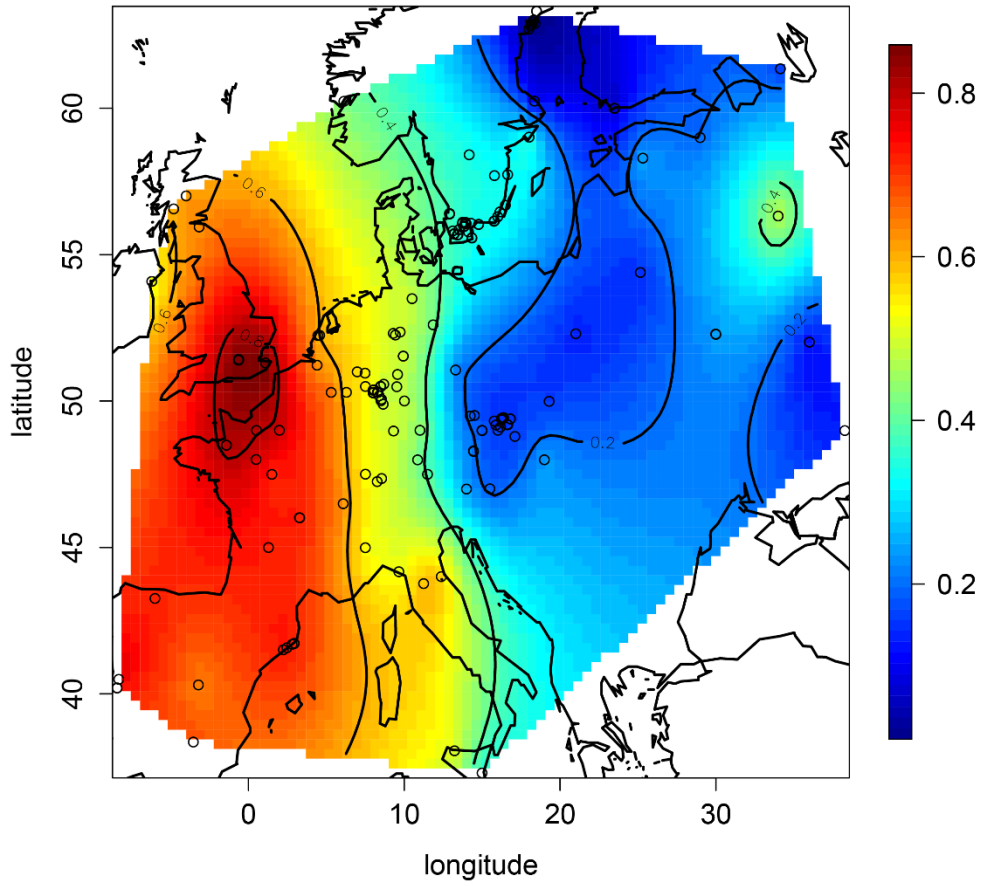
Sampling design and graph used in spatial NMF



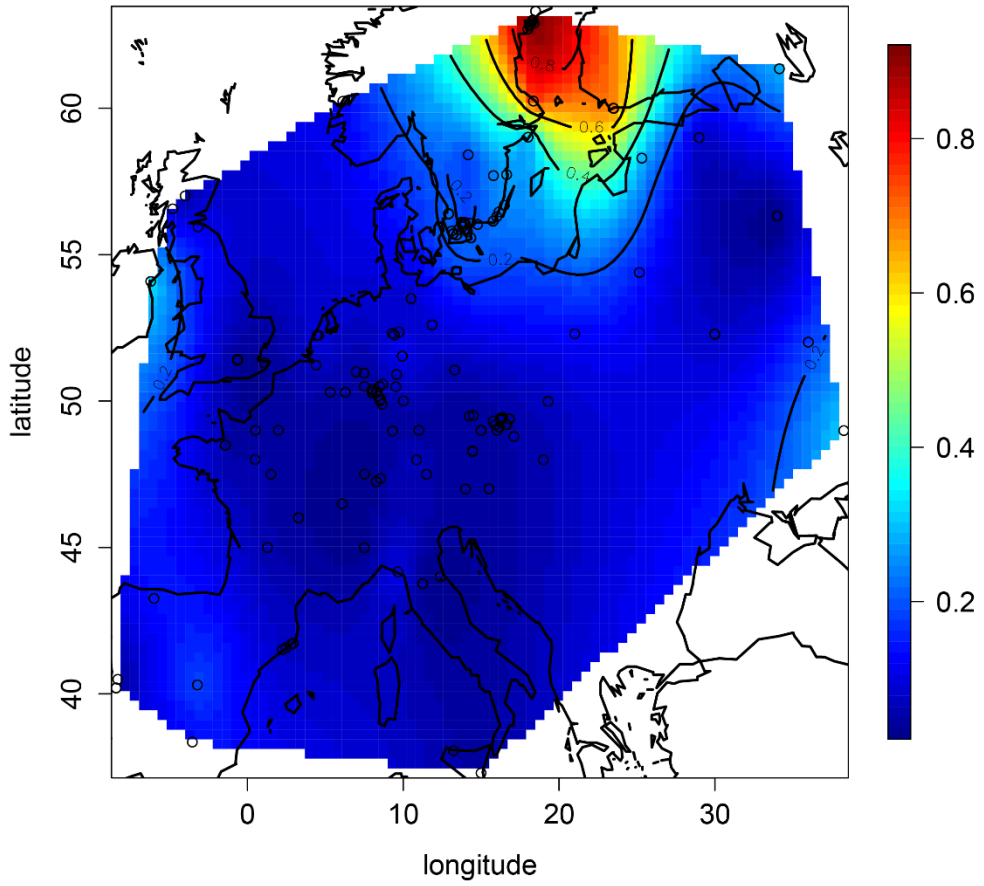


Ancestry map results (K = 3)

Eastern ancestral population coefficient



Scandinavian ancestral population coefficient



Discussion

- Graph regularized NMF combines spatial and genetic data
- We developed a new ALS algorithm for graph NMF
- We observed improved statistical performance compared to sparse NMF in spatially explicit population genetic simulations
- 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