Estimation of Spatial Population Structure

Kevin Caye[1], Olivier Michel[2], Olivier Francois[1]
[1] TIMC-IMAG, [2] GIPSA-lab

Introduction

Geography and landscape are important determinants of genetic variation in natural populations, and several ancestry estimation methods have been proposed to investigate population structure using genetic and geographic data simultaneously. Those approaches are often based on computer-intensive stochastic simulations, and do not scale with the dimensions of the data sets generated by high-throughput sequencing technologies. There is a growing demand for faster algorithms able to analyze genome-wide patterns of population genetic variation in their geographic context.

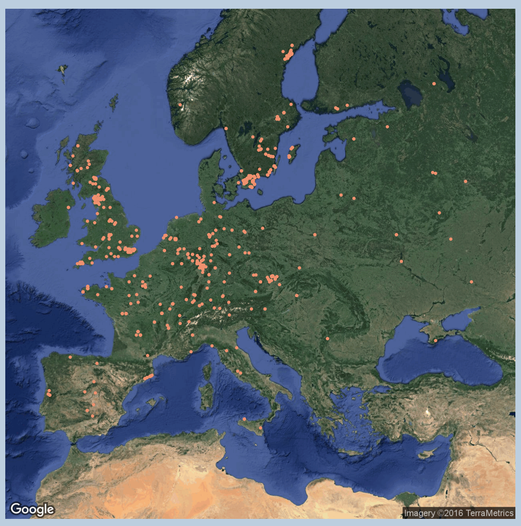
Input Data

Genotypic data:
DNA Sequencing Technologies :

- SNPs array (*Arabidopsis thaliana* RegMap lines [4] : $200k$ loci of 1 307 individuals)
- next generation sequencing (1000 Genome project [2] : whole genome of 2504 individuals)

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

Spatial data:
Individual spatial coordinates of *Arabidopsis thaliana* RegMap Lines dataset.



Model

Ancestry coefficients and ancestral population definitions:
We write G the genomic matrix.

$$P(G_{i,\ell} = j) = \sum_{k=1}^{K} Q_{i,k} f_{k,\ell}(j),$$

$$P = QF^T,$$

where Q is the ancestry coefficient matrix and F the ancestral genotype frequency matrix.
Optimisation Problem:
Optimisation problem to estimate Q and F of sNMF method [3]:

$$\min_{Q,F} \quad ||X - QF^T||^2$$

$$\text{such as} \quad Q \succeq 0,\ F \succeq 0$$

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

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

where X is a binary matrix which encode absence or the presence of each genotype at each locus.
Graphe based regularization:
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 introduce in GNMF method [1] is:

$$||X - QF^T||^2 + \lambda \sum_{i,j}^{n} W_{i,j} ||Q_{i,} - Q_{j,}||^2$$

$$||X - QF^T||^2 + \lambda \text{trace}(Q^T L Q),$$

where L is the graph laplacian matrix.

Algorithm

- The TESS3 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**
  # Least squares problems
  for $j \in \{1, .., (D+1)L\}$ **do**
    $F^T_{j,.} \leftarrow \arg\min ||Vec(X^j) - Qf||^2$
  end for
  # Projection onto F polygon of constraints
  $F \leftarrow \mathcal{P}_F(F)$
  # ℓ_2-regularized least squares problems
  for $i \in \{1, .., n\}$ **do**
    $Q^T_{Ri,.} \leftarrow \arg\min ||X^T_{R,i} - Fq||^2 + \lambda \mu_i ||q||^2$
  end for
  # Projection onto Q polygon of constraints
  $Q \leftarrow \mathcal{P}_Q(R^T Q_R)$
end for

References

[1] Deng Cai et al. Graph regularized nonnegative matrix factorization for data representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 33(8):1548–1560, 2011.

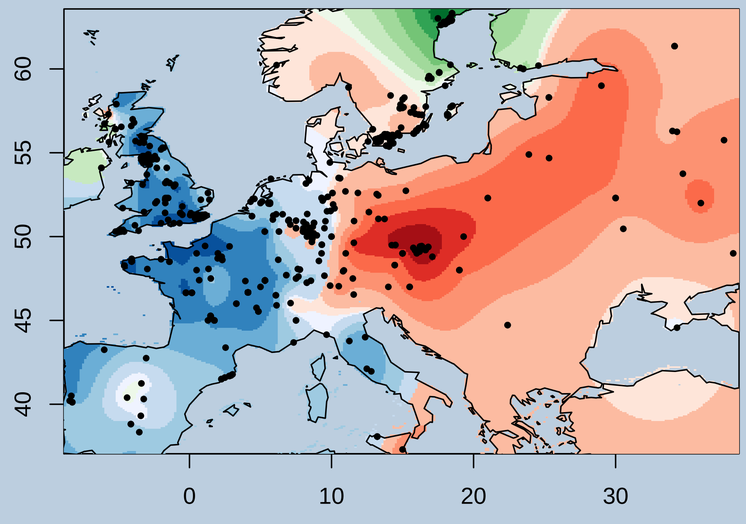
[2] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68–74, 2015.

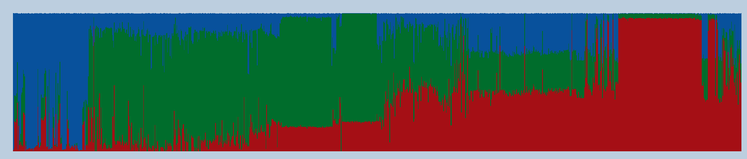
[3] Eric Frichot et al. Fast and efficient estimation of individual ancestry coefficients. *Genetics*, 196(4):973–983, 2014.

[4] Matthew W Horton et al. Genome-wide patterns of genetic variation in worldwide arabidopsis thaliana accessions from the regmap panel. *Nature genetics*, 44(2):212–216, 2012.

Results

We computed population structure of the *Arabidopsis thaliana* RegMap Lines dataset with $K = 3$ ancestral populations.
Individual ancestry coefficients: Ancestry coefficients can be projected on a map.





Ancestral populations: A F_{st} statistic is calculated to represent the genotype distribution differentiation between ancestral populations.