Algorithm For The Ancestry Coefficient Estimation In Spatial Populations

48èmes Journées de Statistique de la SFdS

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

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

13 février 2017

Genotypic Data

DNA Sequencing Technologies :

- SNPs array (*Arabidopsis thaliana* RegMap lines [Horton et al., 2012] : 200k 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

Spatial Data

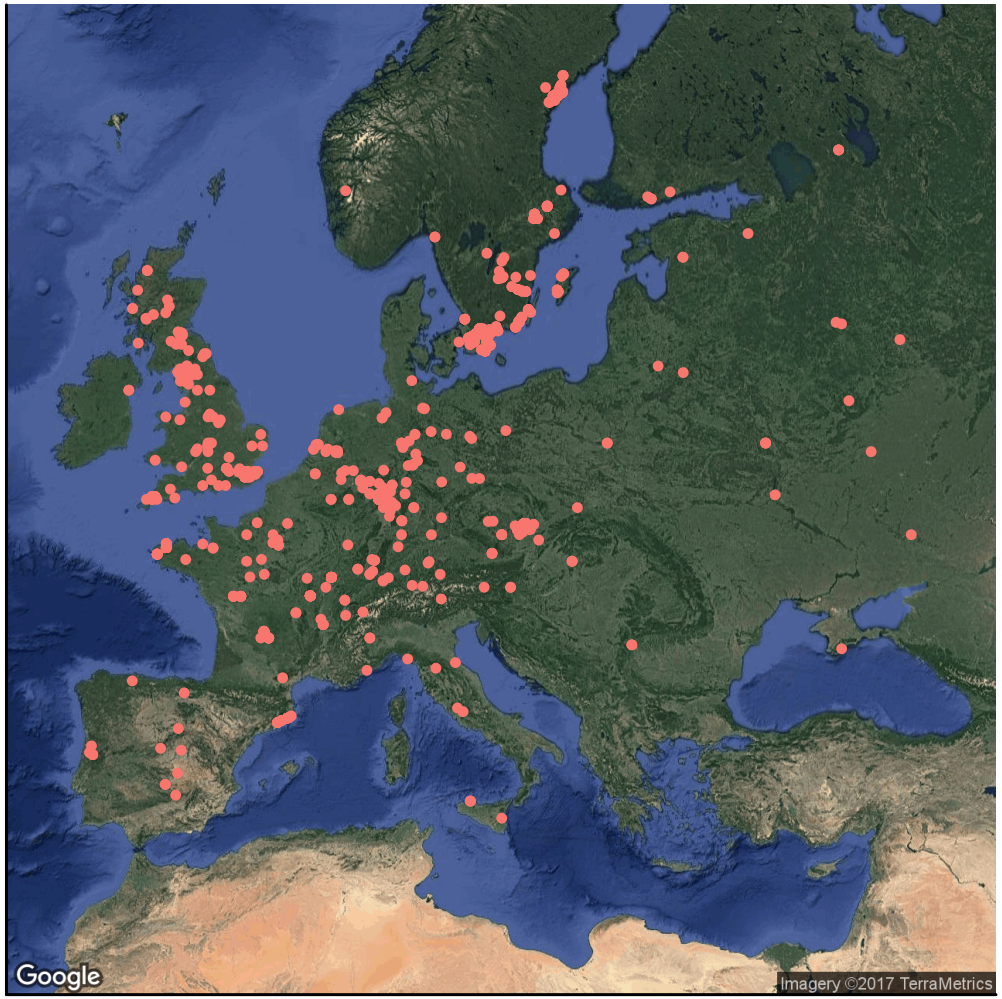


Figure 1 – Individual spatial coordinates of *Arabidopsis thaliana* RegMap Lines dataset.

Goal : Estimating Individual Ancestry Coefficients

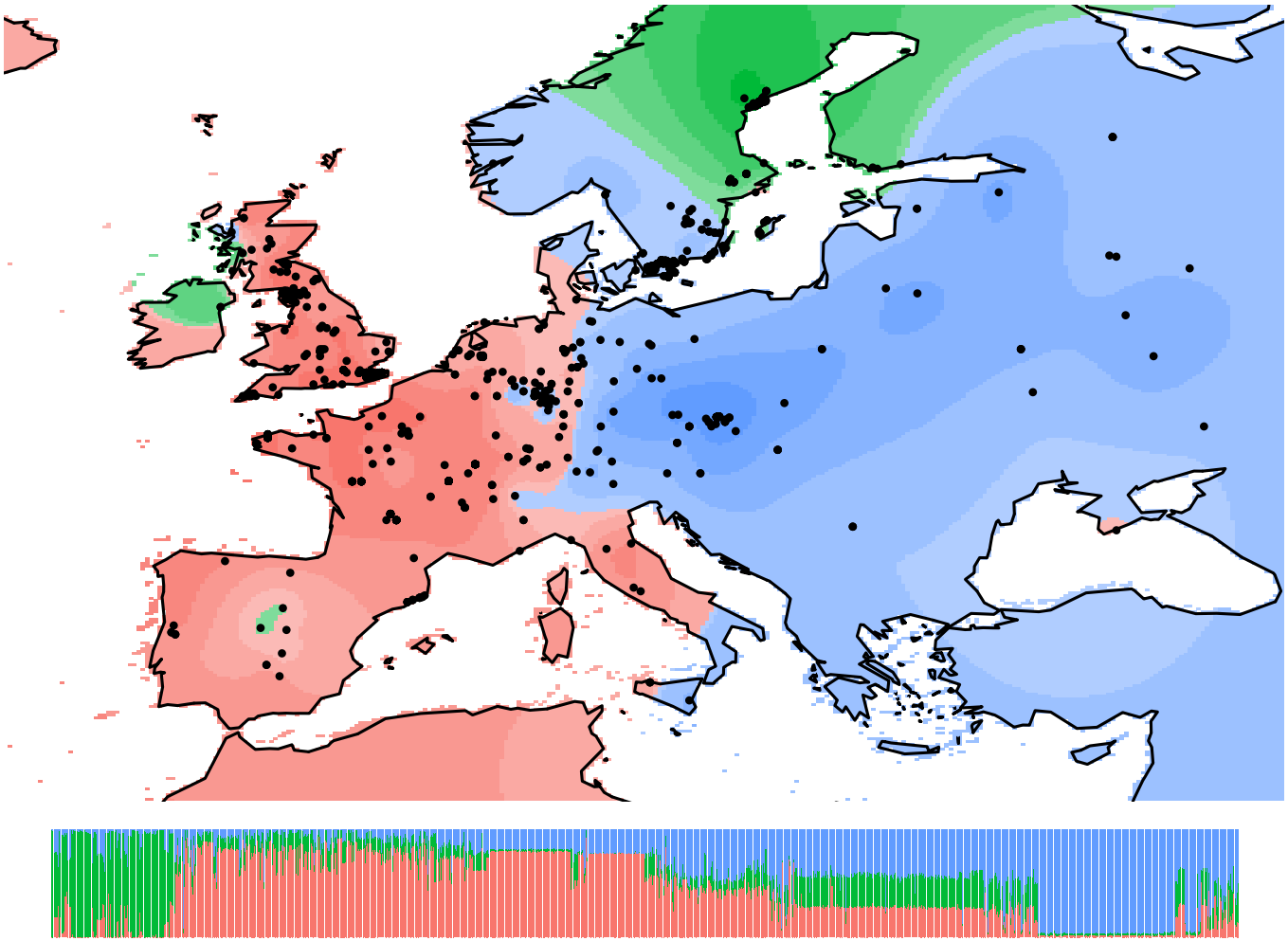


Figure 2 – Individual ancestry coefficients of *Arabidopsis thaliana* RegMap Lines dataset calculated for $K = 3$ ancestral populations.

Model

We write G the genomic matrix.

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

where Q is the ancestry coefficient matrix and F the ancestral genotype frequency matrix.

Optimisation Problem

Optimisation problem to estimate Q and F of sNMF method [Frichot et al., 2014] :

$$\begin{aligned} \min_{Q,F} \quad & ||X - QF^T||^2 \\ \text{tel que} \quad & Q \succeq 0, \ F \succeq 0 \\ & \sum_{k=1}^{K} Q_{i,k} = 1, \ \forall i \in \{1, ..., n\} \\ & \sum_{j=0}^{d} f_{k,\ell}(j) = 1, \ \forall \ell \in \{1, ..., L\}, \end{aligned}$$

where X is a binary matrix which encode absence or the presence of each genotype at each locus.

Graph 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 TESS3 method [Caye et al., 2015] 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.

Block-coordinate Descent Scheme

- 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**

$$F \leftarrow \arg\min_{F} f_F(Q, F)$$

$$Q \leftarrow \arg\min_{Q} f_Q(Q, F)$$

end for

Alternated Quadratic Programming (AQP)

AQP Pseudo Algorithm

for $it \in 1, .., itMax$ **do**

  # Quadratic programming problem for each locus

  for $l \in 1, .., L$ **do**

    $Vec(F^T_{(d+1)l..(d+1)l+d,.}) \leftarrow \underset{f \in \Delta_F}{\arg\min} - 2cf + f^T Df$

  end for

  # Quadratic programming problem of size $n \times K$

  $Vec(Q^T) \leftarrow \underset{q \in \Delta_Q}{\arg\min} - 2cq + q^T Dq$

end for

Alternated Quadratic Programming (AQP)

- Advantage :
  Such algorithm is guaranteed to asymptotically provide a stationary point Bertsekas [1999].
- Drawback :
  The Q optimizarion step is a quadratic programming problem of size $n \times K$.

TESS3 Algorithm : Non Negative Matrix Factorization (NMF)

TESS3 Pseudo Algorithm

for $it \in 1, .., itMax$ **do**

  # Non-negativite least squares

  for $j \in 1, .., (D+1)L$ **do**

    $F_{j,.}^T \leftarrow \underset{f \succeq 0}{\arg\min} - 2cf + f^T Df$

  end for

  # Projection onto F polygon of constraints

  $F \leftarrow \mathcal{P}_F(F)$

  # Non-negativite least squares

  $Vec(Q^T) \leftarrow \underset{q \succeq 0}{\arg\min} - 2cq + q^T Dq$

  # Projection onto Q polygon of constraints

  $Q \leftarrow \mathcal{P}_Q(Q)$

end for

TESS3 Algorithm : Non Negative Matrix Factorization (NMF)

- Advantage :
  The NMF problems are solved with efficient active-set like method Kim and Park [2011]
- Drawback :
  The algorithm is not garenteed to asymptotically provide a stationary point.
  The Q optimizarion step is a non negative least squares problem of size $n \times K$.

Alternated Projected Least Squares (APLS)

APLS Pseudo Algorithm

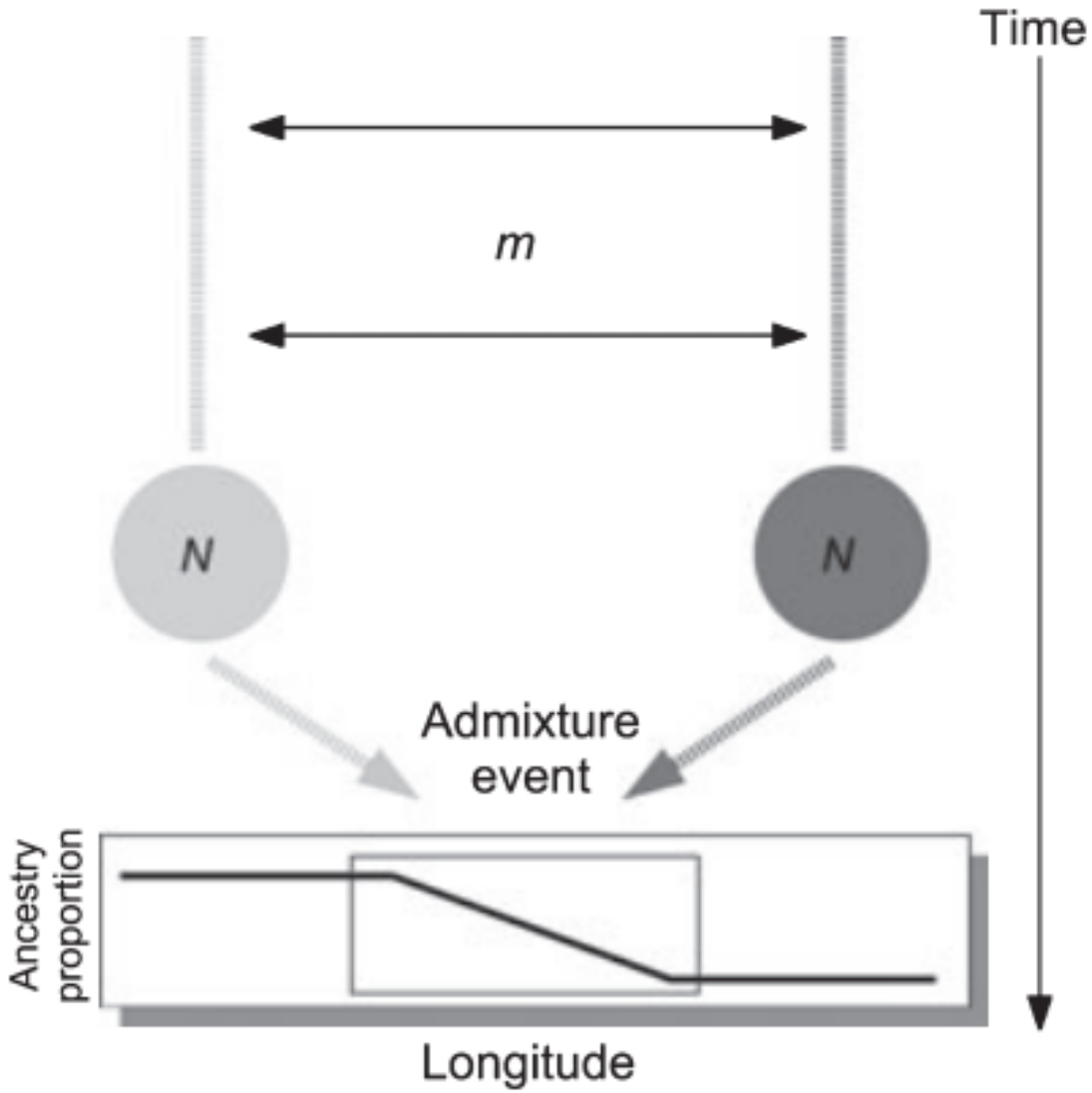
for $it \in \{1, .., itMax\}$ **do**
  # Least squares problems
  for $j \in \{1, .., (D+1)L\}$ **do**
    $F_{j,.}^T \leftarrow \arg\min ||Vec(X^j) - Qf||^2$
  end for
  # Projection onto F polygon of constraints
  $F \leftarrow \mathcal{P}_F(F)$
  # ℓ_2-regularized least squares problems
  for $i \in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

Alternated Projected Least Squares (APLS)

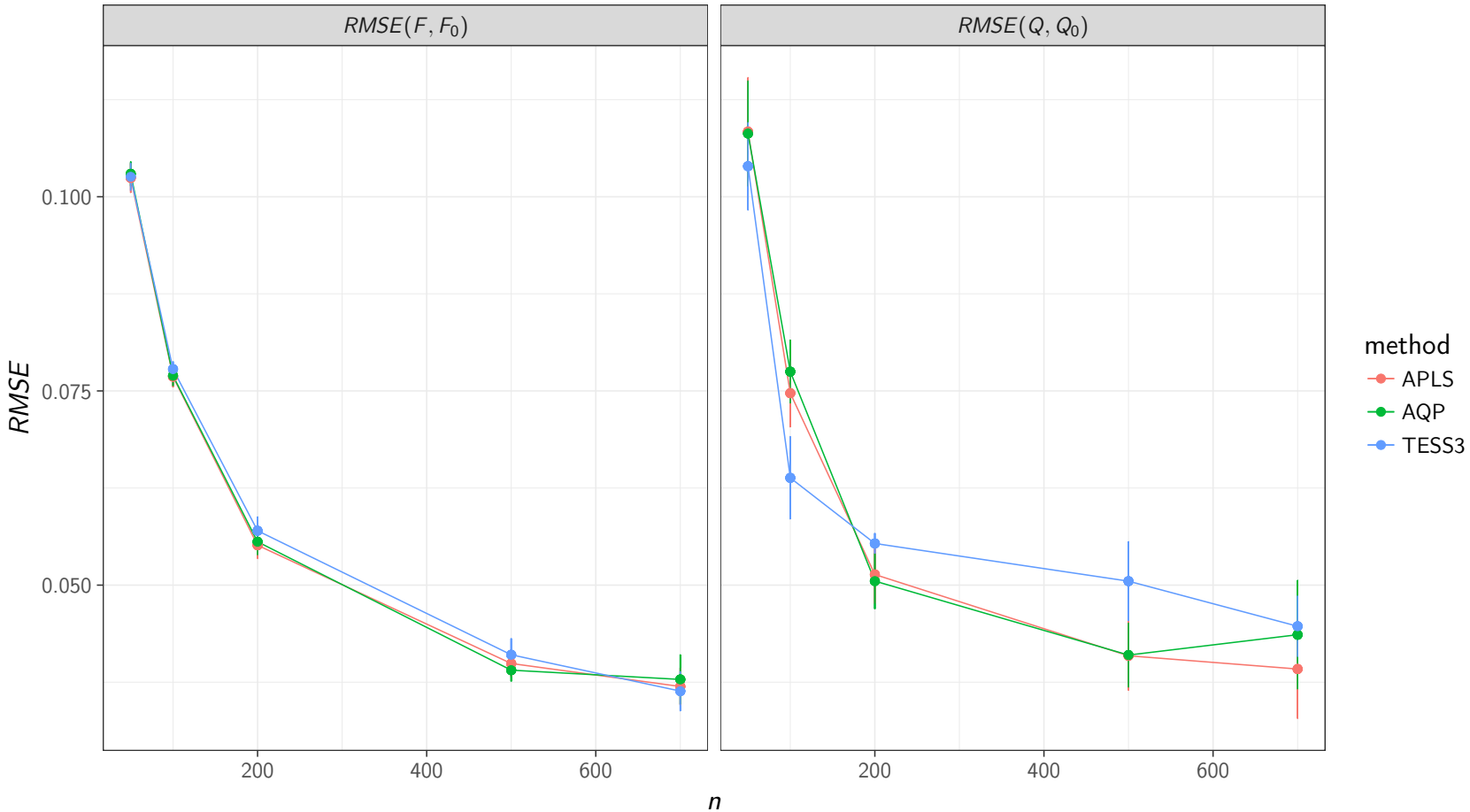
- Advantage :
  The Q optimisation step require to solve n ℓ_2-regularized least squares problems of size K.
- Drawback :
  The algorithm is not garenteed to asymptotically provide a stationary point.

Simulations

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

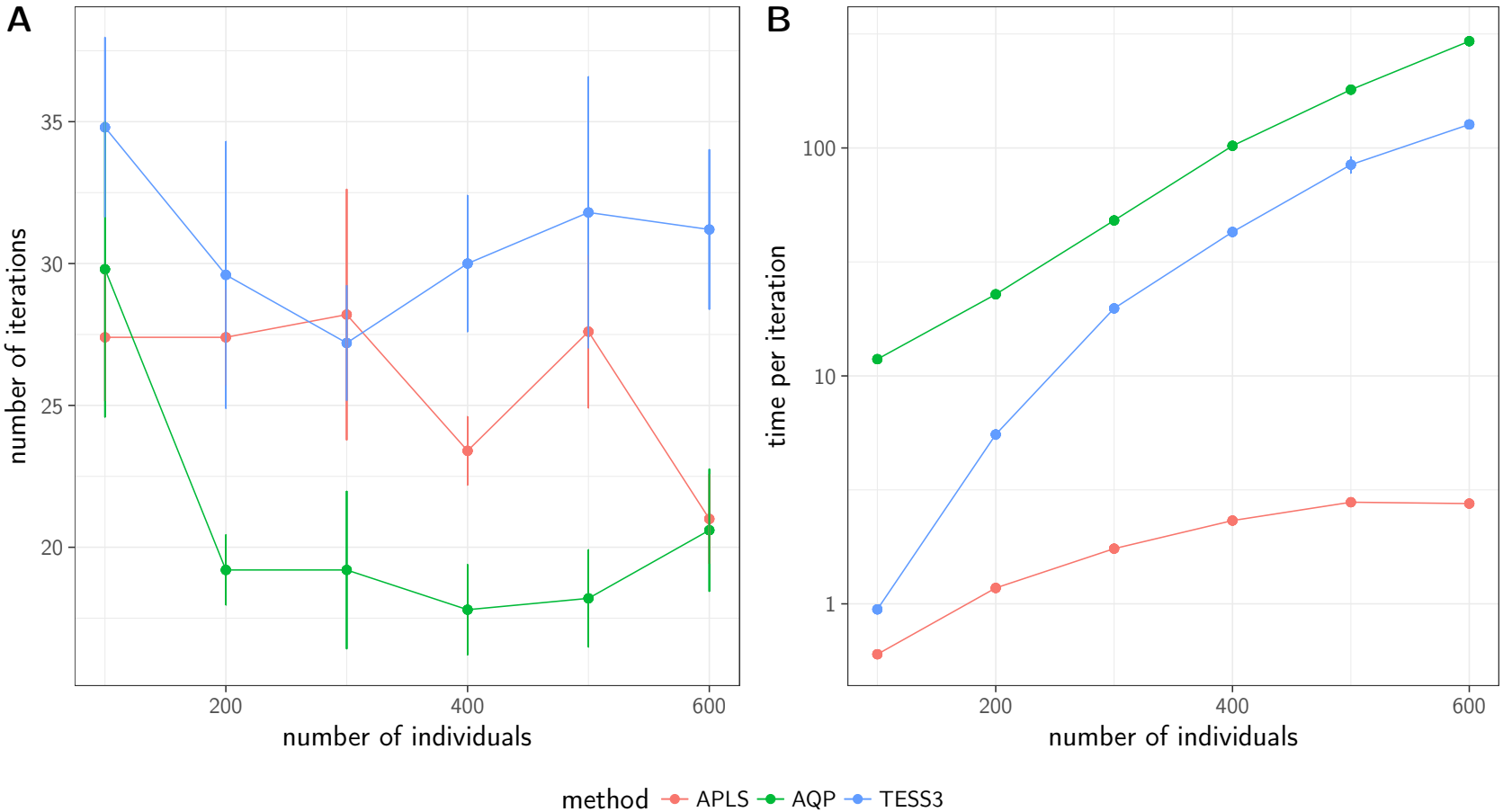


Estimation Error Comparisons



Runtime Performance Comparisons

We used *Arabidopsis thaliana* RegMap lines dataset to sample dataset with different number n of individuals.



Conclusion

- On considered dataset, the three algorithms converge and provide same estimation errors.
- APLS algorithm have a better complexity in n the number of individuals.

Thank you for your attention.

References

Dimitri P Bertsekas. Nonlinear programming. 1999.

Kevin Caye, Timo M Deist, Helena Martins, Olivier Michel, and Olivier François. Tess3 : fast inference of spatial population structure and genome scans for selection. *Molecular ecology resources*, 2015.

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

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