Estimating Ancestry Coefficients using NMF & Spatial Information

A Comparison to Bayesian Methods

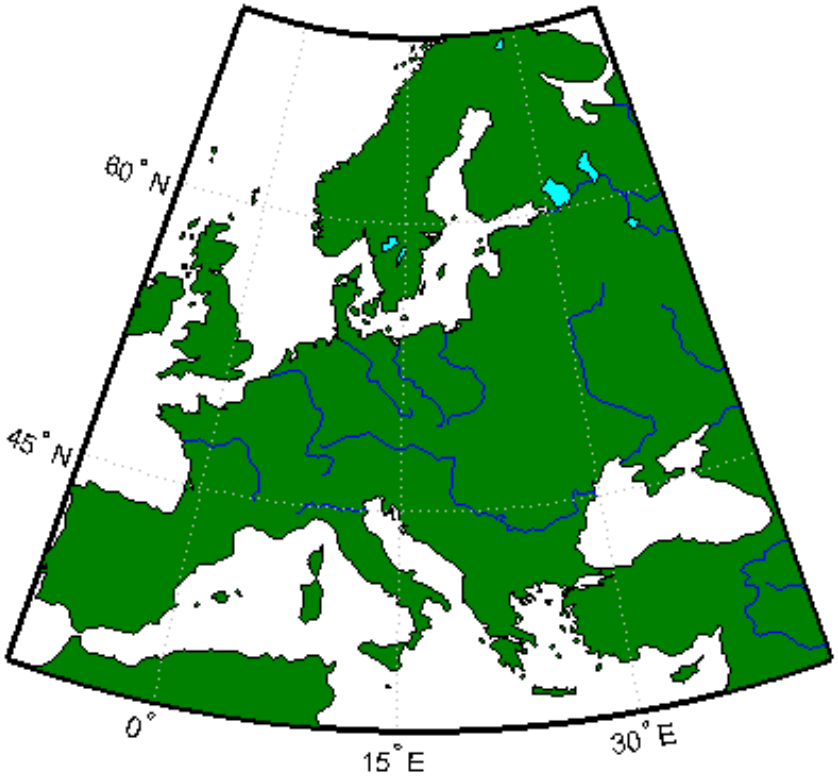
Timo Deist, TIMC-IMAG, Grenoble

Problem Statement

- Population spread over area
- Population structure?

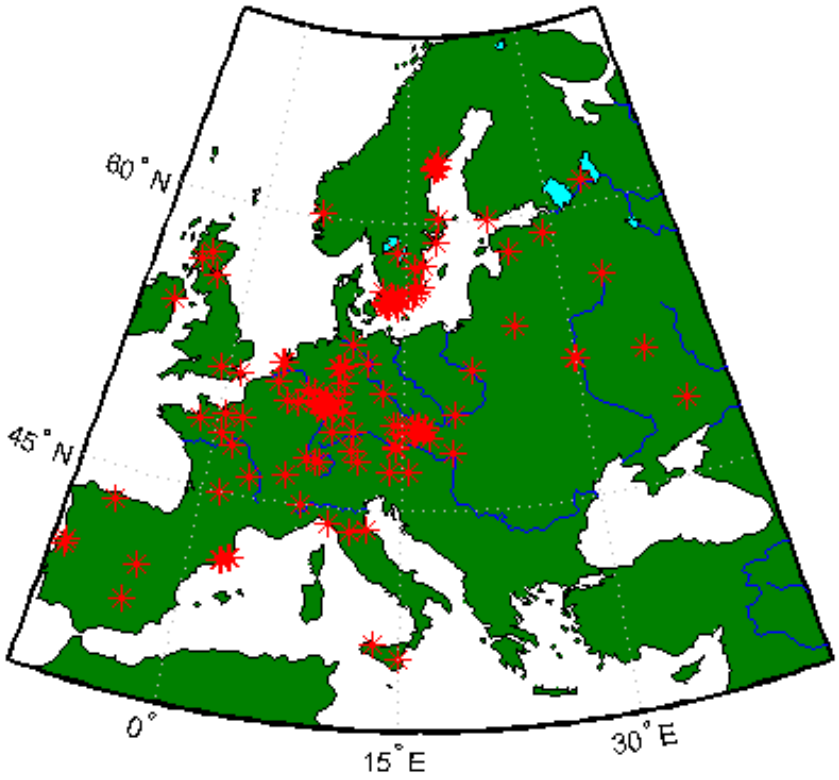
→ Compare genotypes stored in a large matrix M

- M encodes absence/presence of specific alleles
  - binary matrix



Pop. Structure Estimation

- Sample Individuals
- Determine genotype
- Estimate number of subpopulations k
- Estimate ancestry coefficients
  - similar to clustering



Traditional NMF Objective

$$\min ||M - UV^T||_F^2$$

Ancestral Allele Frequencies

Ancestry Coefficients

Minimize squared error in factorization

$$\text{s.t.} \sum_{i \in P_\ell} U_{ij} = 1 \quad \forall \ell \in \{1, ..., p\} \quad \forall j \in \{1, ..., k\}$$

$$\sum_{j=1}^{k} V_{ij} = 1 \quad \forall i \in \{1, ..., n\}$$

$$U \geq 0$$

$$V \geq 0$$

D. D. Lee and H.S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature, 401:788–791, 1999.*

NMF with Spatial Information

$$\min ||M - UV^T||_F^2 + \lambda Tr(V^T L V)$$

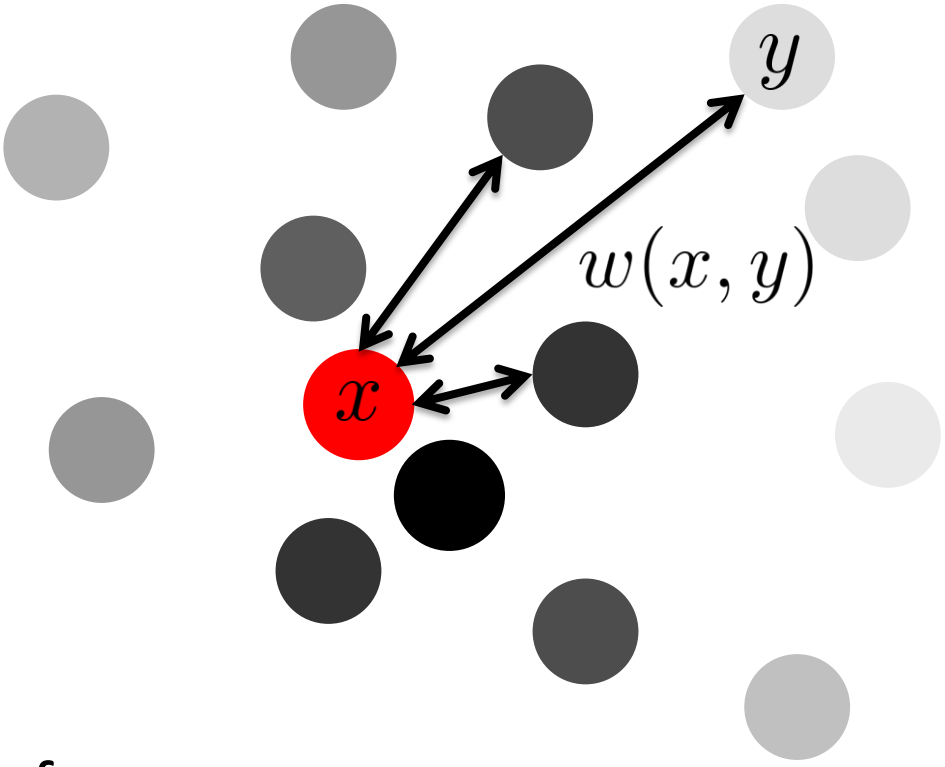
Minimize weighted difference in ancestry coefficients

D. Cai, X. He, T. S. Han, J. Huang. Graph regularized non-negative matrix factorization for data representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* 33:1548–1560, 2010.

Spatial Information

$$w(x,y) = e^{\frac{-d(x,y)}{\sigma}}$$

- $d(x,y)$: Euclidean distance
- σ : determines steepness
  - Set to mean distance of 10 nearest neighbors



Remarks

$$\min ||M - UV^T||_F^2 + \lambda Tr(V^T L V)$$

- Weighted minimization
  - Difficult to set weights
- Objective is non-convex
  - global optimum hard to obtain
  - settle for approximate solutions?

Algorithm

- Alternating Least Squares (ALS)

- Approximate optimal (U, V)
  1) Initialize random (U, V)
  2) Fix one, determine other variable
  3) Vice versa
  4) Repeat until convergence

M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate non negative matrix factorization. *Computational Statistics & Data Analysis, 52:155–173, 2007.*

ALS

$$\min ||M - UV^T||_F^2 + \lambda Tr(V^T L V) + \beta \sum_{ij} V_{ij}$$

- Observe:
  - convex in either U or V

ALS

Repeat until convergence:

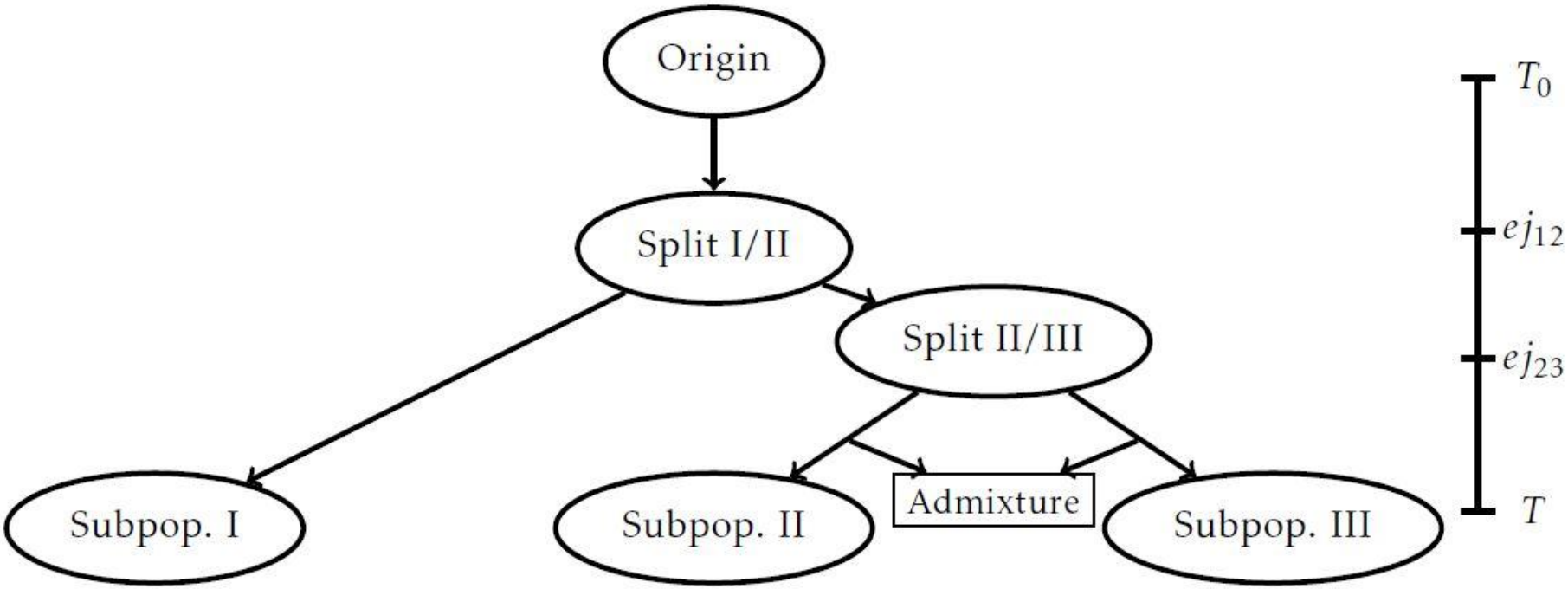
1) Fix V
2) Solve unconstrained problem for U :

$$\min ||M - UV^T||_F^2 + \lambda Tr(V^T LV) + \beta \sum_{ij} V_{ij}$$

3) Remap to feasible space
4) Repeat with fixed U
5) Repeat until convergence

Data

- Two simulation models:
  - 2-island model with migration & admixture
    - 50 data sets
    - 200 individuals, 1000 SNPs
  - divergence model with admixture
    - 45 data sets
    - 300 individuals, 1000 SNPs
- Real data, Atwell et al. (Science, 2010):
  - *Arabidopsis thaliana*
    - 170 individuals, 10000 SNPs


Origin
Split I/II
Split II/III
Subpop. I
Subpop. II
Admixture
Subpop. III
T_0
ej_{12}
ej_{23}
T

Data

- Two simulation models:
  - 2-island model with migration & admixture
    - 50 data sets
    - 200 individuals, 1000 SNPs
  - divergence model with admixture
    - 45 data sets
    - 300 individuals, 1000 SNPs
- Real data, Atwell et al. (Science, 2010):
  - *Arabidopsis thaliana*

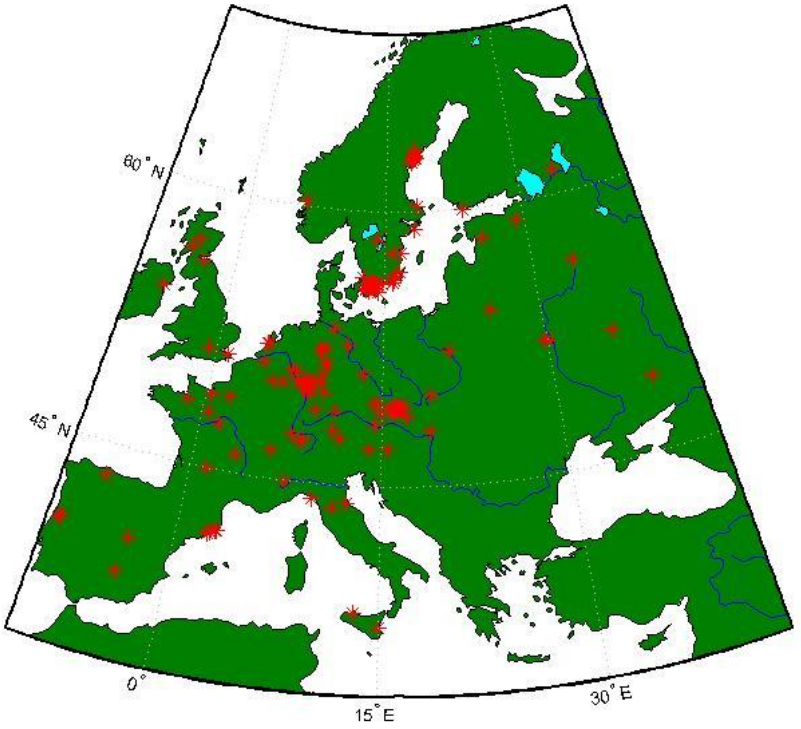
Arabidopsis thaliana

- Model plant with small genome
- Native to Eurasian continent
- Further spread in recent centuries

Reproduced under GNU Free Documentation License

Arabidopsis thaliana

- 170 European samples with 10k SNPs
- Subset of samples from Atwell et al. (Science, 2010) with 210k SNPs



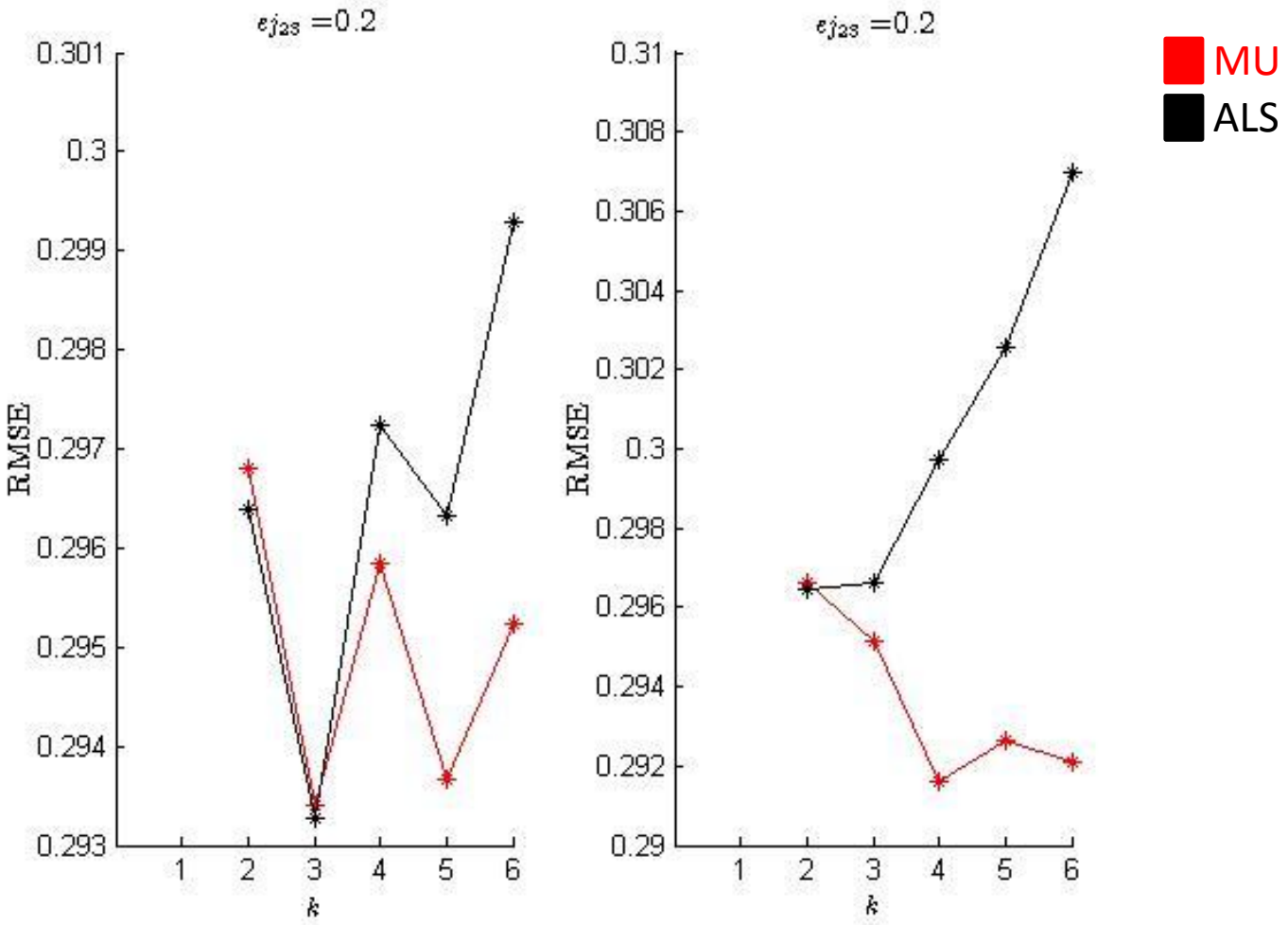
Results

- Comparison between new ALS algorithm and existing Multiplicative Updates (MU) algorithm
  - MU based on Cai et al. (2010)

Estimating the number of subpopulations k

- How many subpopulations k exist?
- Observe estimation error of genotype estimates
  - Root-mean-squared error (RMSE)
- Choose smallest k with significant decrease in RMSE

Estimating the number of subpopulations k



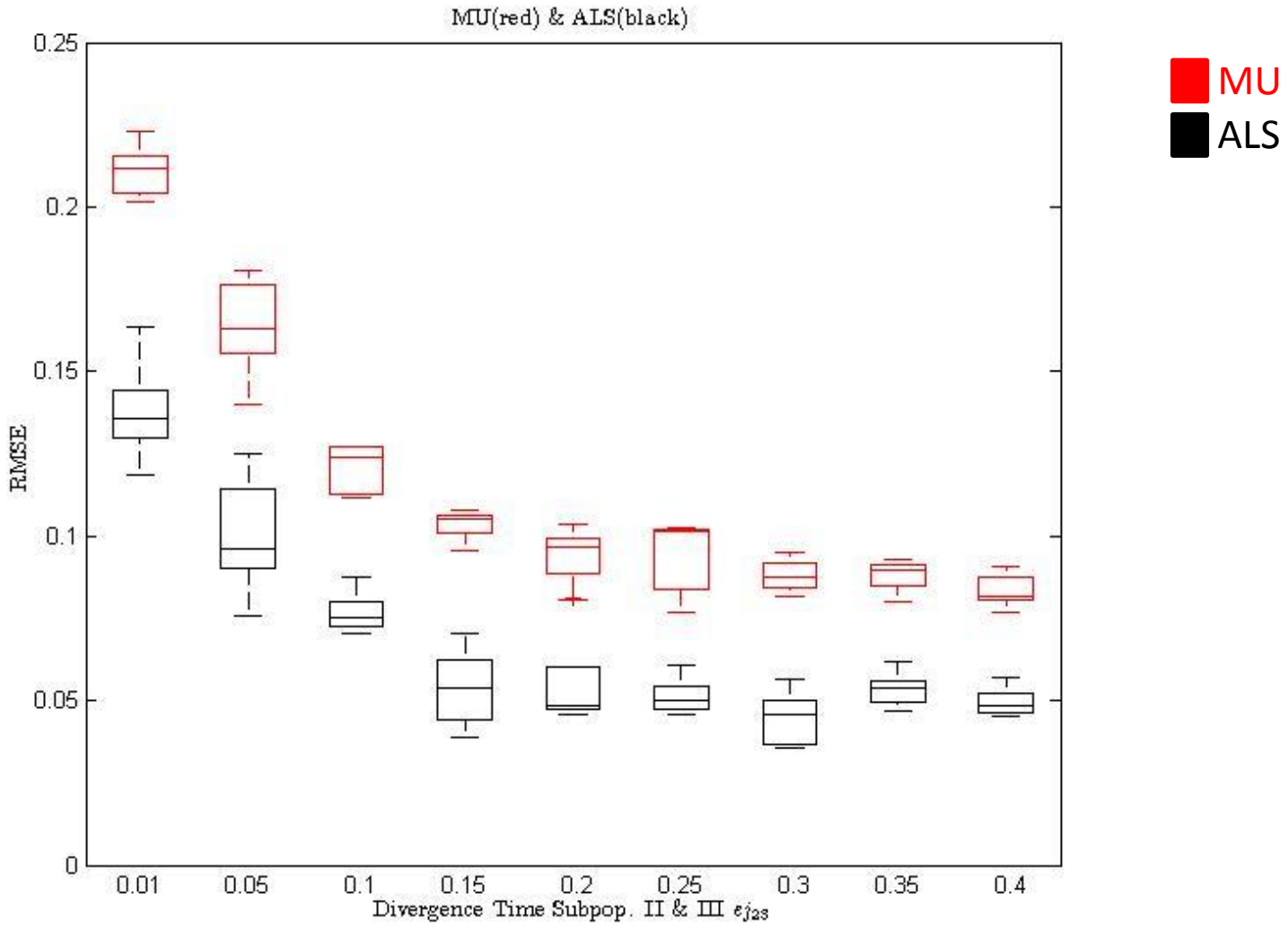
Effect of Regularization Parameters on Ancestry Estimation

- **ALS:** Strictly better estimates when including geographic information

Estimation Error of Ancestry Coefficients

- Ancestry coefficients estimated in V
- Compare RMSE of estimates vs. true values
  - True values are known for simulated data
- Comparison of best parametrizations

Estimation Error



Results – MU versus ALS

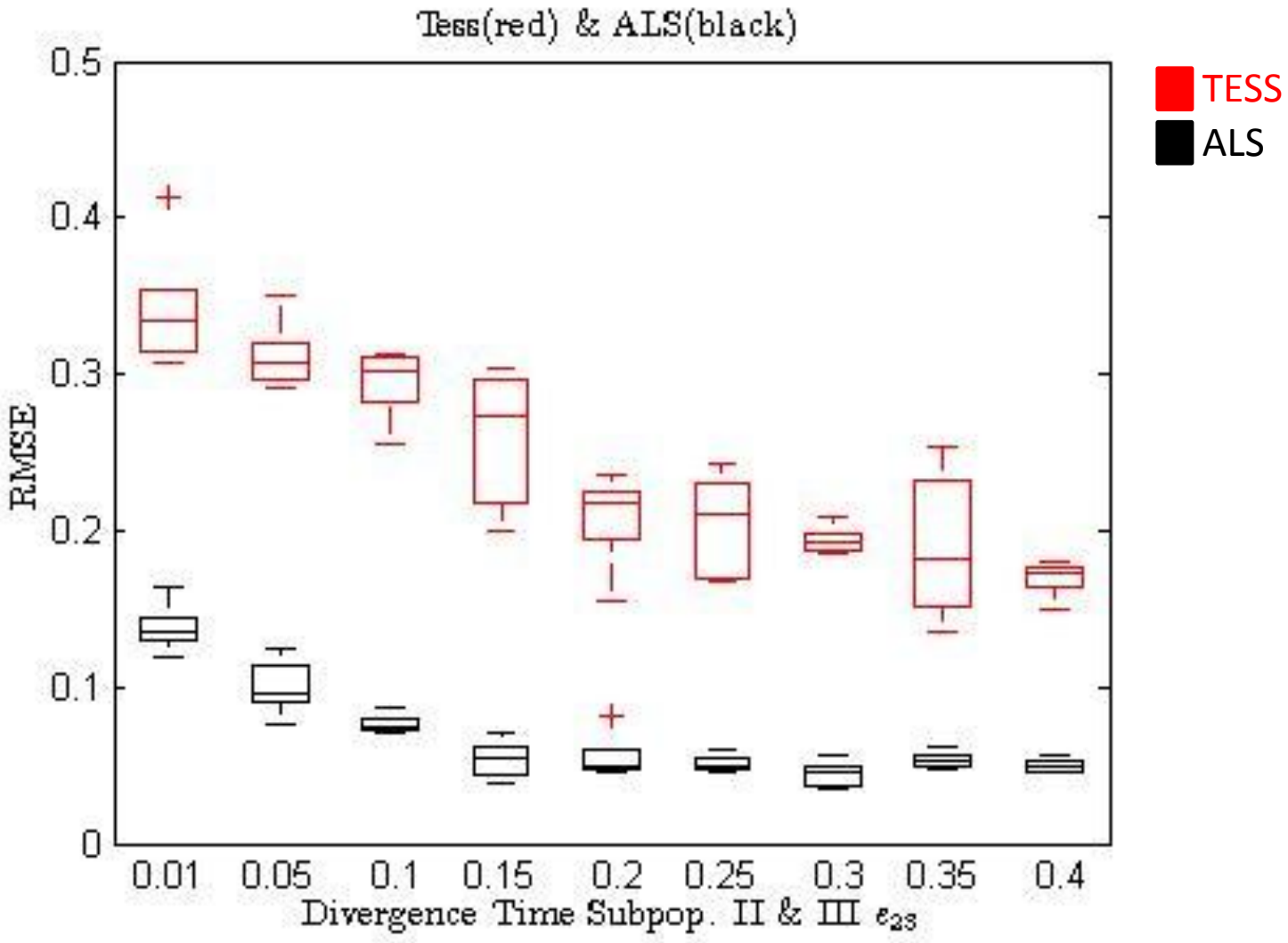
- RMSE is lower (ALS)
- Relatively stable optimal parameterization (ALS): $(\beta, \lambda) = (50, 100)$
- Similar convergence time

Results

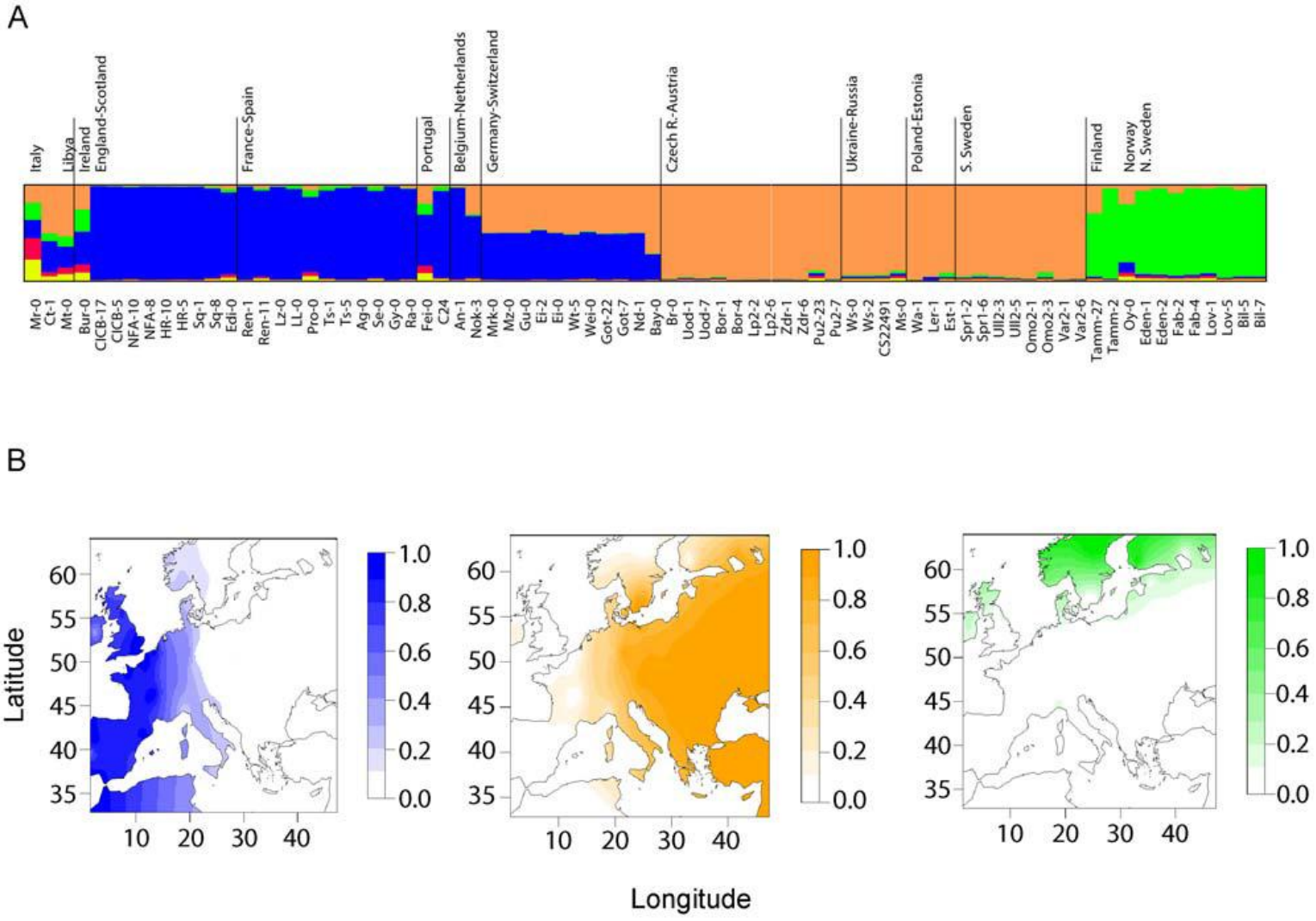
- Comparison with TESS, Durand et al. (MBE, 2009)
  - estimation error (sim. Data)
  - ancestry coefficients for *Arabidopsis thaliana*

- TESS
  - models distribution of ancestry coefficients
  - estimates based on Monte Carlo Markov Chains

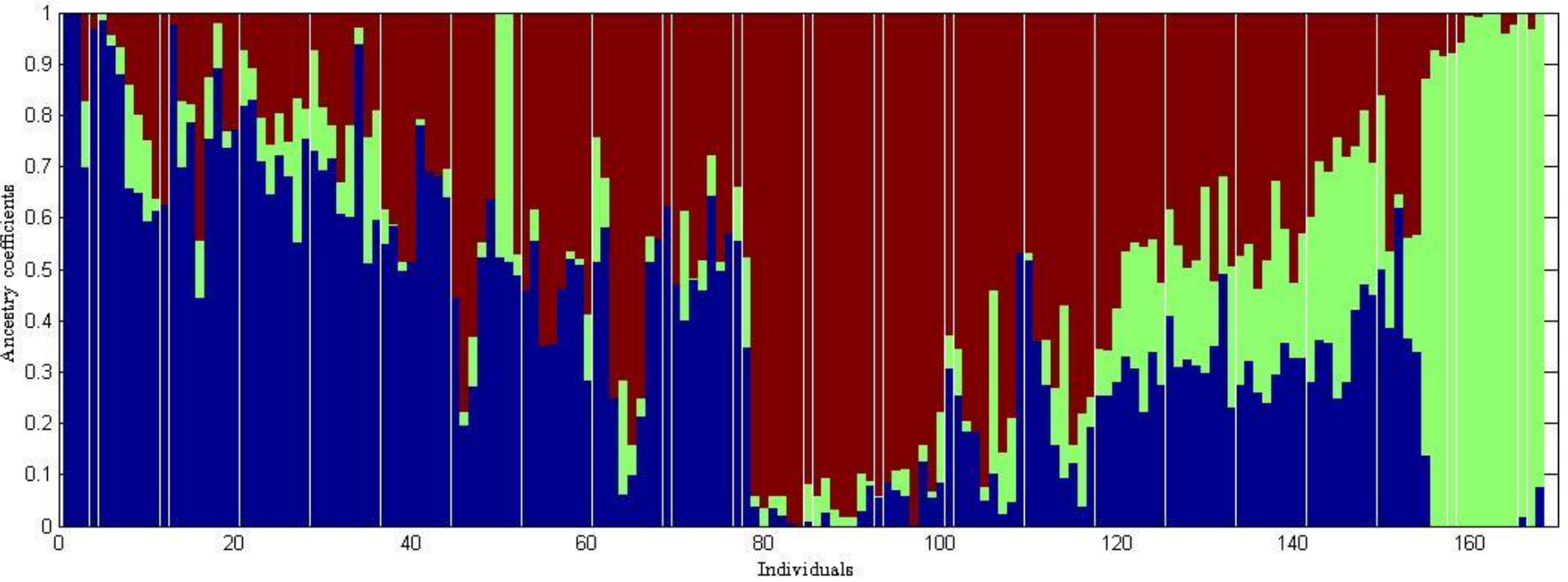
Estimation Error

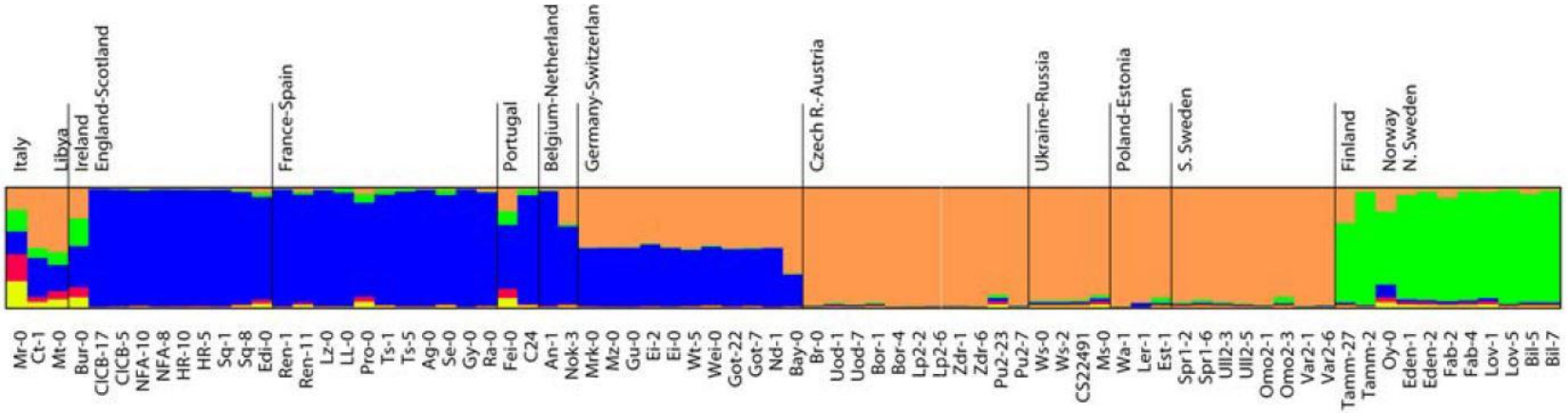


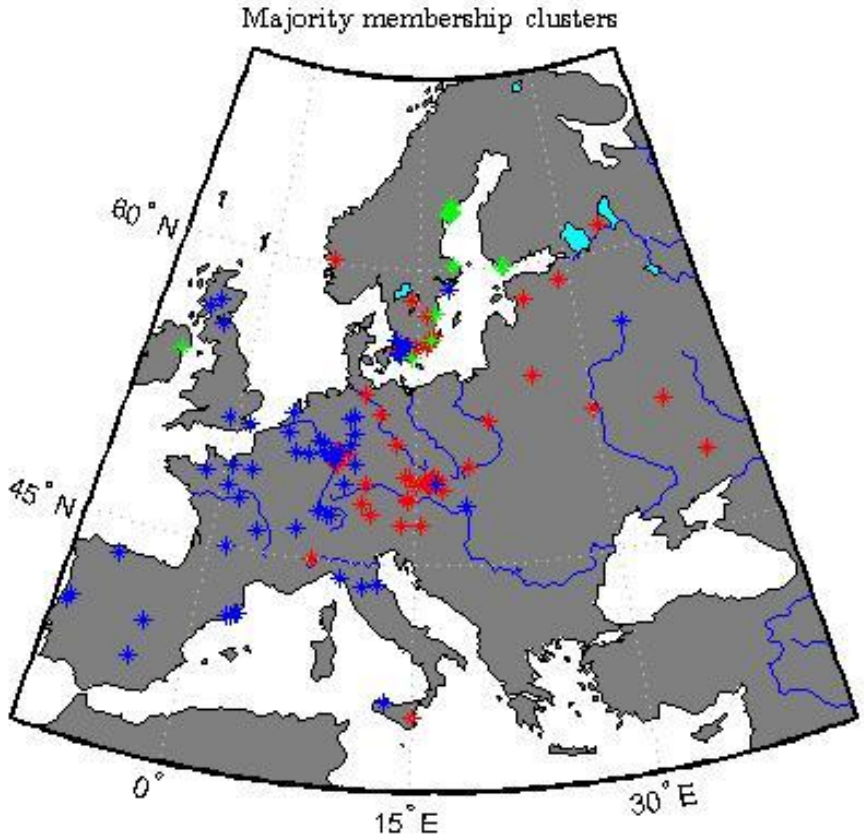
Arabidopsis thaliana

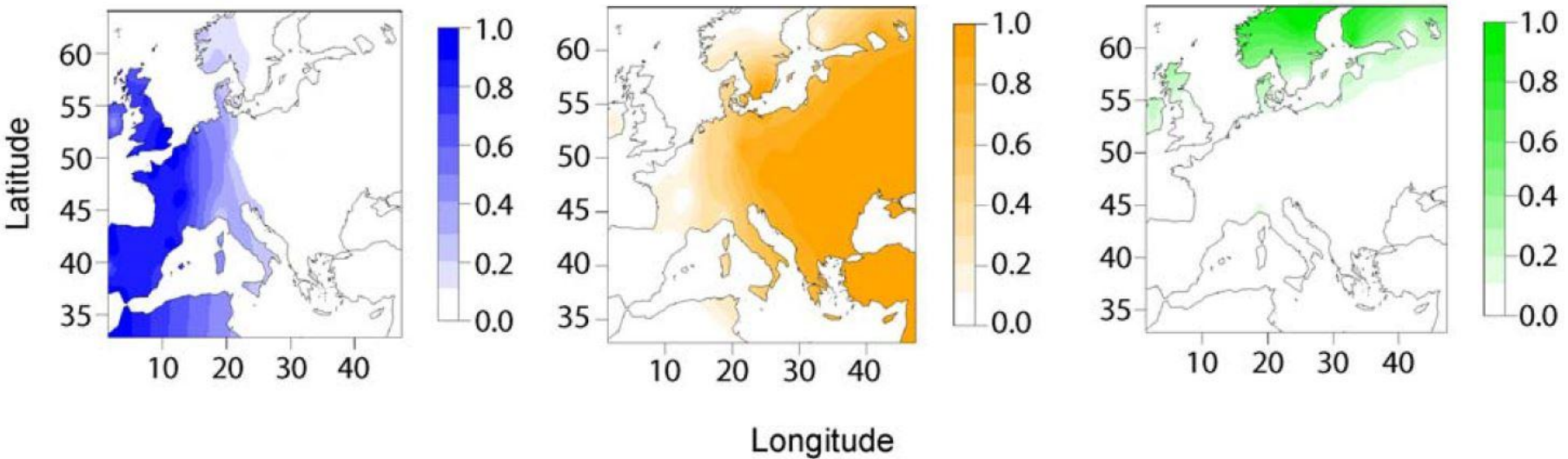


François et al. (2008), Demographic History of European Populations of Arabidopsis thaliana. PLOS Genetics


Ancestry coefficients
Individuals
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
20
40
60
80
100
120
140
160


Italy
Libya
Ireland
England-Scotland
France-Spain
Portugal
Belgium-Netherland
Germany-Switzerlan
Czech R.-Austria
Ukraine-Russia
Poland-Estonia
S. Sweden
Finland
Norway
N. Sweden
Mr-0
Ct-1
Mt-0
Bur-0
CICB-17
CICB-5
NFA-10
NFA-8
HR-10
HR-5
Sq-1
Sq-8
Edi-0
Ren-1
Ren-11
Lz-0
LL-0
Pro-0
Ts-1
Ts-5
Ag-0
Se-0
Gy-0
Ra-0
Fei-0
C24
An-1
Nok-3
Mrk-0
Mz-0
Gu-0
Ei-2
Ei-0
Wt-5
Wei-0
Got-22
Got-7
Nd-1
Bay-0
Br-0
Uod-1
Uod-7
Bor-1
Bor-4
Lp2-2
Lp2-6
Zdr-1
Zdr-6
Pu2-23
Pu2-7
Ws-0
Ws-2
CS22491
Ms-0
Wa-1
Ler-1
Est-1
Spr1-2
Spr1-6
Ull2-3
Ull2-5
Omo2-1
Omo2-3
Var2-1
Var2-6
Tamm-27
Tamm-2
Oy-0
Eden-1
Eden-2
Fab-2
Fab-4
Lov-1
Lov-5
Bil-5
Bil-7


Majority membership clusters
60°N
45°N
0°
15°E
30°E


Latitude
60
55
50
45
40
35
10
20
30
40
1.0
0.8
0.6
0.4
0.2
0.0
Longitude

Results – ALS versus TESS

ALS

- Consistently lower average RMSE (sim. data)

- Lower execution time (real data)
  - ALS: < 6 seconds
  - TESS: 170 seconds

Conclusion

- Geographic information improves estimates
  - Parameterization is difficult
- NMF is fit for large data sets
  - Significantly faster than TESS
- Non-parametric approach with same interpretation as model-based approach