University Grenoble-Alpes **GIPSA-lab, GSTATS Data Science Chair**
INTERNSHIP PROJECT **TIMC-IMAG, Computational and Mathematical Biology**

Romain COUILLET & Michael BLUM
2018 - 2019 **Office D1178**

A Random Matrix Analysis of Genetic Variation between individuals

Project context and application

Genetic data convey rich information to decipher human demography and population ancestry. Of utmost interest to researchers in genomics is a natural visualization of genomic data in two-dimensional planes. Long-standing approaches have relied on MDS (multidimensional scaling) [1], or on Principal Component Analysis (PCA) [3,4] to provide an optimal 2D representation of large dimensional data (Figure 1). The more modern t-SNE (stochastic neighbor embedding) approach has been shown to reveal population structure on a finer geographical scale [2].

However, to the exception of the seminal contribution [5], these analyses do not account for the fact that sample covariance matrices based on few individuals but extremely large dimensional gene vectors are typically weak estimators of the true covariance structure and are therefore bound to large errors. The project proposes to exploit recent tools from random matrix theory to study dimension reduction methods in genomics [6]. Using evolutionary and population models, the objective is to get a better grasp on the observations made with dimension reduction method in population genomics.

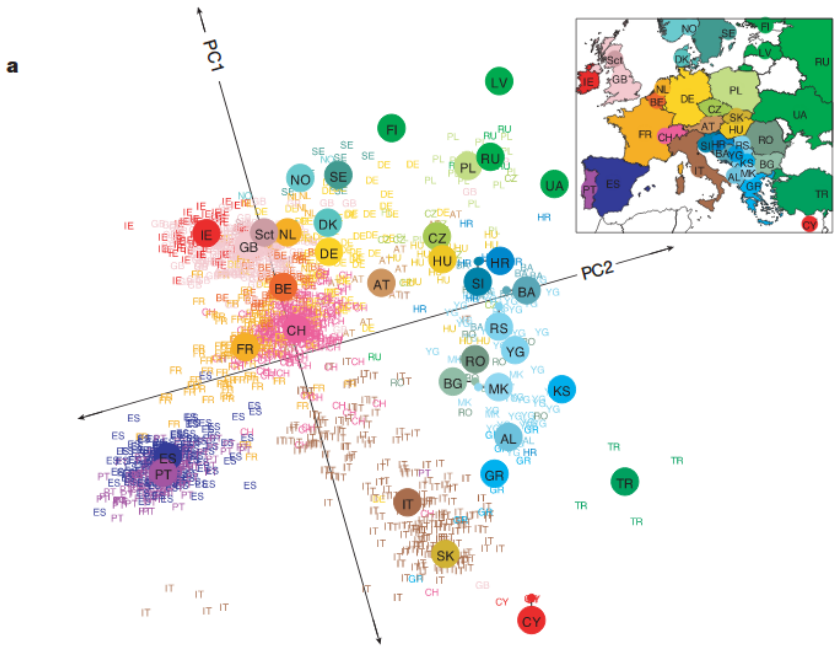

Figure 1 : Principal component analysis of European individuals based on genomic data. The two principal components mimic the geographic map of Europe [4].

Main steps

- Review of the literature on MDS, t-SNE and PCA methods in genetics.
- Implementation and comparison of the methods.
- Theoretical analysis of the performance using random matrix theory.

Requirements: Good coding skill in R, Matlab or Python, knowledge of the basics of large dimensional statistics (optionally random matrix theory), good understanding of general signal processing, statistics, and machine learning concepts.

Location: The internship will take place at GIPSA-lab, University of Grenoble-Alpes, in the Grenoble area.

References

[1] Diaconis, P., Goel, S., & Holmes, S. (2008). Horseshoes in multidimensional scaling and local kernel methods. *The Annals of Applied Statistics*, *2*(3), 777-807.
[2] Diaz-Papkovich, A., Anderson-Trocme, L., & Gravel, S. (2018). Revealing multi-scale population structure in large cohorts. *bioRxiv*, 423632.
[3] McVean, G. (2009). A genealogical interpretation of principal components analysis. *PLoS genetics*, *5*(10), e1000686.
[4] Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature genetics*, *40*(5), 646.
[5] Bryc, K., Bryc, W., & Silverstein, J. W. (2013). Separation of the largest eigenvalues in eigenanalysis of genotype data from discrete subpopulations. *Theoretical population biology*, *89*, 34-43.
[6] Couillet, R., & Debbah, M. (2011). *Random matrix methods for wireless communications*. Cambridge University Press.