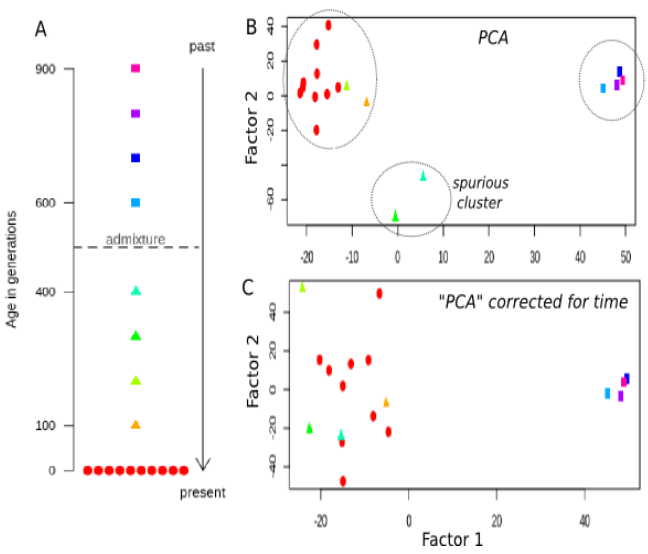
Réduction de dimension et imputation pour la paléogénomique

Stage M2 (ou long M1) 3 à 6 mois

Flora Jay (LRI, Paris saclay) flora.jay@lri.fr
Olivier François (TIMC-IMAG, Grenoble) olivier.francois@univ-grenoble-alpes.fr

Depuis l'apparition des premières extractions ADN à partir de fossiles, les données génomiques d'échantillons anciens ne cessent d'augmenter. Elles sont extrêmement précieuses puisqu'elles ouvrent une fenêtre directe sur l'histoire passée. Par exemple, pour étudier les processus à l'oeuvre dans les différentes transitions culturelles de l'Histoire humaine, les paléogénéticiens séquencent des individus ayant vécu avant/pendant/après une transition et analysent les similarités génétiques entre ceux-ci pour comprendre la structure cachée des données. Les méthodes de réduction de dimension classiquement utilisées en génétique des populations, type PCA, permettent la détection et visualisation de la structure mais ne tiennent pas compte de l'hétérogénéité temporelle des données de paléogénomique.



Or il a été montré que, tout comme l'autocorrélation spatiale, l'autocorrélation temporelle a un impact sur la construction des axes de PCA ou d'analyse factorielle [1]. Pour tenir compte explicitement de l'autocorrélation temporelle lors de la réduction de dimension de données paléogénomiques nous avons donc commencé à développer un nouveau package (*tDR* [3]) regroupant différentes approches reposant sur (1) des modèles de factorisation matricielle avec ou sans contraintes, (2) des modèles mixtes à facteurs latents (LFMM [2]), (3) une extension de l'ACP probabiliste à des modèles à covariance non diagonales (Figure). Cependant nous avons montré que ces méthodes sont peu robuste aux données manquantes, point problématique puisque l'ADN ancien est généralement très dégradé et les génomes séquencés sont loin d'être complets.

Plus précisément les données nous intéressant sont constituées d'environ 1000 génomes d'individus modernes (*1000 Genomes* database) séquencés à 4X en moyenne (plus de 70M de marqueurs génétiques) et d'une cinquantaine d'individus anciens provenant des principales études sur l'Europe et l'Asie pendant les périodes Néolithique, Âge du Bronze et Âge du Fer.

Objectifs

Les objectifs du stage seront de contribuer au développement du package et de :

(1) Déterminer des critères d'évaluation de performance objectifs pour comparer différentes versions de *tDR* et estimer les hyperparamètres (basée sur la vraisemblance, sur les critères d'évaluation d'algorithmes de classification, ...)

(2) Evaluer les performances de *tDR* sur le jeu de données imputé et publié par [4]. C'est-à-dire un jeu de 30M de marqueurs génétiques, où les génotypes anciens manquants sont prédits à partir d'une base de référence de génomes contemporains et du logiciel *GATK*. En particulier évaluer le biais introduit par ce type d'imputation sur les axes de *tDR.*

(3) Implémenter et évaluer des méthodes intégrant les données manquantes. En particulier, comparer les performances d'algorithmes de type IPCA [5] avec des approches venant du *collaborative filtering,* avec ou sans pré-filtrage des individus et marqueurs les moins bien représentés.

(4) Appliquer aux données réelles humaines pour une meilleure compréhension du paysage génétique.

(5) Selon les compétences et propositions de l'étudiant, d'autres méthodes de visualisation / réduction / clustering des données pourront être implémentées et testées.

Profil

Etudiant niveau M2 (ou M1), machine learning, biostatistique, bioinformatique, math/info, ...

Compétences recherchées :

Programmation R et Python

Apprentissage statistique / Statistiques multidimensionnelles

Intérêt pour la biologie et la génétique des populations sont un plus.

Légende

Fig. A. Demographic scenario with ancient and contemporary individuals. A gene flow event occurred 500 generations ago. There is a discontinuity between squares and triangles but not between triangles and circles. B. A PCA on genetic data wrongly identifies 3 clusters. C. A PCA corrected for temporal drift should show two clusters gathering (i) individuals before admixture (squares) (ii) individuals after admixture (triangles and circles).

Biblio

[1] Duforet-Frebourg and Slatkin. "Isolation-by-Distance-and-Time in a Stepping-Stone Model." *TPB* 108 (2016): 24–35. [2] Frichot et al. "Testing for associations between loci and environmental gradients using latent factor mixed models." MBE 30.7 (2013): 1687-1699. [3] Liegeois et al. "Dimension Reduction Adapted to Paleogenomics." Poster at JDSE2018 Paris-Saclay. [4] Martiniano et al. "The population genomics of archaeological transition in west Iberia: Investigation of ancient substructure using imputation and haplotype-based methods." PLoS genetics 13.7 (2017): e1006852. [5] Josse, J. & Husson, F. Handling missing values in exploratory multivariate data analysis methods. JSFS 79–99 (2012)