

# Génétique des populations pour la paléogénétique : analyse de données temporelles *Population genetics: time matters*

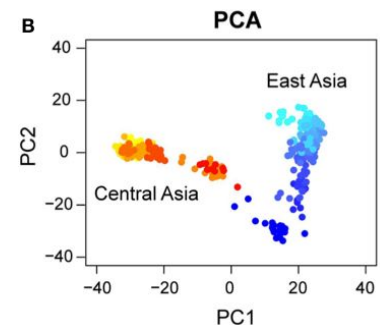
Flora Jay (LRI, Orsay) [flora.jay@lri.fr](mailto:flora.jay@lri.fr) ; collaboration avec Olivier François (TIMC-IMAG, Grenoble)

Stage de M2, 5 ou 6 mois (éventuellement M1 si au moins 3 ou 4 mois). Contactez F. Jay  
Début possible dès février/mars 2018

## Contexte

La paléogénétique est en grand développement depuis une dizaine d'années, et les techniques sont telles que l'on a maintenant accès aux **génomés d'échantillons anciens (datant d'il y a plusieurs milliers d'années, eg. Homme de Néandertal)**. Ces données sont extrêmement précieuses puisqu'elles ouvrent une fenêtre directe sur l'histoire passée. Par exemple, pour étudier les différentes transitions culturelles qui ont eu lieu au cours de l'Histoire humaine, les chercheurs séquentent des individus ayant vécu avant/pendant/après une transition. Pour comprendre le mécanisme de la transition (remplacement complet d'une population par une autre, métissage entre locaux et nouveaux arrivants, ou pure diffusion culturelle) on peut alors évaluer **la diversité génétique des individus séquencés et détecter une potentielle structuration en groupe génétiquement proches** (sorte de clustering non supervisé des individus en fonction de leurs données génétiques).

Une méthode couramment utilisée pour explorer les patrons de variation dans un groupe d'individus est l'**Analyse en Composante Principale (ACP)**. Cette technique classique d'apprentissage permet en effet de **réduire la dimension des données génétiques** (milliers de marqueurs) à quelques axes principaux et de visualiser le regroupement ou la dispersion des individus en fonction de leurs génomes.



## Enjeu du stage

Souvent les méthodes pour décrire la structure génétique des populations ne tiennent pas compte explicitement dans leur modèle de l'**hétérogénéité temporelle d'échantillons anciens**. Or des travaux récents ont démontré l'impact de ce facteur temporel sur l'ACP classique, ce qui conduit fréquemment à une interprétation incorrecte des résultats (eg mauvaise reconstruction de l'histoire passée d'une population). Ces travaux ont de plus modélisé la relation entre différenciation temporelle et génétique [1].

- (1) **L'enjeu de ce stage est donc d'intégrer cette information dans une méthode statistique de type ACP pour corriger l'analyse et éviter les biais causés par l'autocorrélation temporelle des échantillons.** On s'inspirera pour cela d'une approche proposée précédemment pour tenir compte de l'autocorrélation spatiale [2]
- (2) Une fois la nouvelle méthode implémentée, une série de tests sera effectuée sur des données génétiques simulées avec des logiciels existants. La campagne de simulations s'attachera à représenter des scénarios correspondant à des histoires démographiques variées.

- (3) Enfin l'application aux données génétiques de centaines d'individus de populations humaines ayant vécu en Eurasie d'il y a 10000 ans à nos jours permettra de mieux documenter les transitions ayant eu lieu à partir de la période Néolithique.

**Travail demandé :**

*Comprendre l'ACP corrigée [2], programmer la nouvelle méthode (à partir du script développé par [2]), planifier les simulations (qui reposent sur des logiciels existants), tester la méthode, traiter des données réelles. En fonction du temps et de son profil le stagiaire sera libre de contribuer aux packages développés au sein de l'équipe de TIMC-IMAG (implémentation, documentation, test) / de privilégier le dialogue avec nos collaborateurs paléogénéticiens et l'interprétation biologique/démographique des analyses. Il ne sera PAS demandé de dériver les formules mathématiques de dérive génétique.*

**Compétences techniques recherchées :**

- Maîtrise d'au moins un langage de programmation (eg R et/ou Python) et de shell (écrire, lancer des scripts, automatiser des simulations etc.)
- Connaissances en apprentissage statistique
- Connaissances en génétique/génétique des populations (non requises mais avantageuses)
- Anglais (compréhension des articles)

[1] Duforet-Frebourg, Nicolas, and Montgomery Slatkin. 2016. "Isolation-by-Distance-and-Time in a Stepping-Stone Model." *Theoretical Population Biology* 108 (April): 24–35.

[2] Frichot, Eric, Sean Schoville, Guillaume Bouchard, and Olivier François. 2012. "Correcting Principal Component Maps for Effects of Spatial Autocorrelation in Population Genetic Data." *Frontiers in Genetics* 3. (figure copyright)