

## **Sujet stage M2 2018-2019 : Développement d'un programme permettant l'évaluation d'outils bioinformatiques de détection d'éléments transposables polymorphes**

**Responsable** : Emmanuelle Lerat

**Coordonnées** : [emmanuelle.lerat\(AT\)univ-lyon1.fr](mailto:emmanuelle.lerat(AT)univ-lyon1.fr)

Laboratoire « Biométrie et Biologie Évolutive » UMR5558 CNRS Université Lyon 1

Les éléments transposables (ET) sont des séquences répétées dans les génomes, qui ont la capacité de se déplacer et de se dupliquer [1]. Ils jouent un rôle très important dans l'évolution des génomes et dans l'adaptation des espèces [2]. Il est donc important de connaître leur répartition exacte dans les génomes pour mieux étudier leur impact. Cette répartition peut varier d'un individu à un autre [3]. Ainsi, les insertions décrites dans un génome séquencé de référence représentant un individu provenant d'une population donnée ne reflètent pas forcément le contenu en insertion d'individus provenant d'autres populations.

Actuellement, il existe un quinzaine d'outils permettant de déterminer les insertions polymorphes dans des données provenant de séquençage haut débit (NGS) par comparaison à un génome de référence de la même espèce [voir références 4, 5 et 6 par exemple]. Cependant, ces outils ont des performances variables et ne sont pas toujours adaptés à toutes les espèces et tout type de données. Dans le cadre d'un consortium européen DrosEU (dont l'équipe d'accueil fait partie), nous cherchons à identifier les insertions polymorphes et leur impact génomique dans les populations naturelles de *Drosophila melanogaster*. Nous avons à notre disposition les données de séquençage de populations collectées à travers l'Europe. Ces échantillons recouvrent environ 30 localisations différentes prélevés à deux périodes de l'année et sur plusieurs années. Il est donc indispensable de produire un outil d'analyse qui soit performant.

### Le but du stage

1/ mettre au point un programme pour générer des jeux de données théorique permettant de contrôler différents biais connus (composition en GC, mauvaises annotations, tendance des ET à s'insérer les uns dans les autres etc.)

2/ tester plusieurs outils sur les jeux de données théoriques et estimer leur performance vis à vis des différents biais.

3/ tester le(s) meilleur(s) outil(s) sur un jeu de données réel provenant du consortium.

### Compétences requises

Ce stage implique la maîtrise d'un langage de programmation mais aussi d'effectuer des analyses de données NGS et génomiques en utilisant différents outils existants et d'utiliser des tests statistiques. Il faut donc avoir de très bonnes notions en statistique et en programmation, mais aussi des notions en évolution et en génétique des populations pour bien comprendre la problématique biologique.

### Références

1. Wicker T et al. 2007. A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* 8:973–982.
2. Biéumont C, Vieira C. 2006. Genetics: junk DNA as an evolutionary force. *Nature.* 443:521–524.
3. Rishishwar et al. 2015. Transposable element polymorphisms recapitulate human evolution. *Mobile DNA.* 6:21.
4. Kofler R et al. 2012. Sequencing of Pooled DNA Samples (Pool-Seq) Uncovers Complex Dynamics of Transposable Element Insertions in *Drosophila melanogaster*. *PLoS Genet.* 8:e1002487.
5. Zhuang et al. 2014. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res.* 42:6826–6838.
6. Fiston-Lavier et al. (2015) T-lex2: genotyping, frequency estimation and re-annotation of transposable elements using single or pooled next-generation sequencing data. *Nucleic Acids Res.* 43:e22.