Proposition de stage de Master 2
Génotypage de Variants Structuraux avec des données de séquençage de type linked-reads

Equipe d'accueil : Equipe Genscale, Inria Rennes Bretagne Atlantique / IRISA
https://team.inria.fr/genscale/

Adresse où se déroulera le stage : Inria/IRISA, Campus de Beaulieu, 35042 Rennes

Responsable du stage : Claire LEMAITRE, CR Inria, claire.lemaitre@inria.fr
http://people.rennes.inria.fr/Claire.Lemaitre/

Co-encadrement : Claire Mérot, CR CNRS, UMR 6553 ECOBIO, Rennes

Mots clés résumant les méthodes et techniques à utiliser au cours du stage :
Variants de Structure (SV), données de séquençage haut débit, séquençage linked-reads, Haplotagging, mapping de lectures, graphe de variations, algorithmique du texte, programmation en python et/ou C++

Résumé du projet de stage :

Une question fondamentale en biologie est de détecter et d'interpréter les variations entre les génomes d'individus d'une même espèce. Ces variations peuvent être des mutations ponctuelles d'un seul nucléotide (SNP), ou bien peuvent impliquer des segments d'ADN plus longs qui peuvent être dupliqués, supprimés, inversés ou déplacés dans le génome. Ces Variants de Structure (SV), par leur taille et leur association à des séquences répétées, sont beaucoup plus difficiles à détecter et génotyper dans les génomes que les variants ponctuels. Grâce aux progrès du séquençage, on réalise aujourd'hui que les SVs représentent une part considérable mais méconnue de la diversité génétique. Ils couvrent 5 à 10 fois plus de bases dans le génome que les mutations ponctuelles analysées couramment. C'est l'essor du séquençage de 3ème génération (lectures longues) qui a permis ces 5 dernières années d'enfin caractériser et cataloguer toute la gamme de SVs dans de nombreux génomes. Mais cette technologie reste chère et cette caractérisation est limitée souvent à un faible nombre d'individus ou à des espèces modèles, telles que l'homme. Une nouvelle technologie de séquençage, l'Haplotagging [1] (2021), permet de baisser les coûts en séquençant simultanément des centaines d'individus avec une approche « linked-reads » (lectures courtes associées à des barcodes donnant une information longue distance jusqu'à 30-50 Kb).

L'objectif de ce stage est de **développer une méthode** de génotypage de SVs adaptée au données de type linked-reads, d'**appliquer** cette méthode sur des données réelles et de **comparer** l'efficacité de la méthode et de la technologie de séquençage Haplotagging par rapport à l'utilisation de lectures longues.

Pour les développements méthodologiques, le/la stagiaire s'appuiera sur des développements existants dans l'équipe, pour l'étude des SVs et/ou l'analyse des données de type linked-reads. SVjedi et SVJedi-graph[2] sont des outils développés en python pour le génotypage de SVs avec des données de lectures longues (PacBio ou Nanopore). D'autre part, la librairie C++ LRez[3], ainsi que les outils MTG-Link et Leviathan, sont spécifiquement dédiés au traitement et à l'analyse par barcode des

données de type linked-reads. Enfin, nous disposons d'un prototype de logiciel en python pour le génotypage de grandes inversions avec des données linked-reads, qui a permis d'établir la preuve de concept que les informations longue distance apportées par les barcodes permettent le génotypage de ce type de SV. Cette méthode utilise un graphe de variation pour représenter le génome de référence et ses variants. Après avoir mappé les lectures sur ce graphe, le principe du génotypage repose sur la sélection et le comptage des barcodes communs portés par des lectures mappées de part et d'autre des points de cassure des inversions. Il s'agira donc dans un premier temps d'évaluer cette méthode sur des données réelles, d'identifier les limites de la méthode, de proposer de nouvelles solutions algorithmiques et de l'étendre aux autres types de SVs.

Concernant les applications sur données réelles, le stage s'effectuera en collaboration avec Claire Mérot, chercheuse dans le laboratoire EcoBio à Rennes, qui étudie l'impact des SVs sur l'adaptation au changement climatique des insectes. Sur l'espèce de mouche du varech *Coelopa frigida*, plusieurs grandes inversions ont déjà été caractérisées et sont associées à des traits phénotypiques d'intérêt (taille, fécondité, etc.), des données de re-séquençage par Haplotagging (linked-reads) et par lectures longues sont simultanément disponibles. Ces données sont inédites et exceptionnelles, car elles permettent une évaluation fine et réaliste des performances de la méthode développée ainsi qu'une comparaison avec la stratégie par lectures longues.

Ce stage pourra déboucher sur un doctorat dont le sujet portera sur les problématiques de détection et d'analyse des variants de structure dans les organismes non modèles.

Références :

[1] Haplotype tagging reveals parallel formation of hybrid races in two butterfly species, J Meier et al, *PNAS*, 2021, doi:10.1073/pnas.2015005118
[2] SVJedi-graph: improving the genotyping of close and overlapping structural variants with long reads using a variation graph, Sandra Romain et Claire Lemaitre. Bioinformatics, 2023, doi:10.1093/bioinformatics/btad237
[3] LRez: a C++ API and toolkit for analyzing and managing Linked-Reads data. P Morisse, C Lemaitre, F Legeai. *Bioinformatics Advances*, 2021, doi :10.1093/bioadv/vbab022

Montant des indemnités de stage : gratification