Sujet de stage de Master 2
Construction et caractérisation de graphes de variations

Septembre 2022

La disponibilité d'un assemblage de référence pour un grand nombre d'espèces et la démocratisation des technologies de séquençage haut-débit a permis un essor considérable de l'analyse des génomes. Dans cette approche, l'analyse débute généralement par une étape d'alignement qui consiste à aligner l'ensemble des lectures, quelques centaines de paires de bases, à l'assemblage de référence de l'espèce à l'étude. L'alignement permet de caractériser les différences entre l'individu séquencé et l'assemblage de référence et donc les variants génétiques que portent cet individu, à l'état homozygote ou hétérozygote, ce qu'on appelle le génotype d'un individu aux différents variants. Le succès de ces approches a pu faire oublier que l'assemblage de référence ne représentait qu'un seul individu de l'espèce et que cette démarche pouvait donc présenter un biais d'analyse. On observe en effet un biais d'alignement, les lectures qui présentent peu de différences avec l'assemblage de référence sont plus faciles à aligner que des lectures plus distantes, biaisant la quantification des lectures aux positions variables qui sont précisément les positions qui présentent un intérêt pour l'analyse. Ce biais devient même extrême lorsque des fragments génomiques sont absents de l'assemblage de référence, bien que présents dans la population. Il n'est alors pas possible d'aligner les lectures correspondantes sur l'assemblage de réference et les régions associées sont donc exclues de l'analyse.

On s'oriente donc naturellement vers l'abandon d'un unique assemblage de référence au profit d'un pangénome de référence, c'est-à-dire une exploitation simultanée de plusieurs génomes pour caractériser le génome d'une espèce. Le pangénome est généralement encodé sous la forme d'un graphe de variations et l'étape d'alignement consiste alors à aligner les lectures sur ce graphe, c'est-à-dire à les identifier à des chemins du graphe. Cette approche, qui réduit le biais de référence, pose cependant de nouveaux problèmes. Il faut d'abord disposer d'un catalogue de variants fiables pour la construction du graphe de variation, ou d'un ensemble de génomes assemblés. Par ailleurs, le nombre de chemins dans un tel graphe augmente de manière exponentielle avec le nombre de variants ce qui pose des problèmes de calcul. Plusieurs pistes ont été explorées pour aborder ces problèmes. En limitant par exemple le graphe aux variations susceptibles d'avoir le plus grand impact sur le biais [1] ou en développant des heuristiques d'alignement exploitant le fait que tous les chemins du graphe ne sont pas équiprobables [2]. Parmi l'ensemble des chemins possibles, seul un nombre restreint existe réellement dans la population, ils correspondent à ce que l'on appelle en génétique les haplotypes.

Nous proposons dans le cadre de ce stage d'aborder la question de la caractérisation de ces graphes de variations avec applications sur des données réelles d'espèces agronomiques. Les équipes d'accueil disposent de grands jeux de données sur différentes espèces telles que la vache, la chèvre ou le maïs. Ces jeux de données sont composés chacun à la fois de dizaines voire centaines d'assemblages de génomes d'individus différents, et également de catalogues de variants et données de séquençage longues lectures et courtes lectures pour de nombreux autres individus. L'objectif du stage est de caractériser les graphes de pangénomes obtenus avec ces données (en particulier les assemblages), selon plusieurs niveaux, tels que la fréquence des différents types et tailles de variants représentés dans ces graphes, la distribution

le long du graphe des variants et du nombre de chemins possibles, ou encore l'analyse des fréquences alléliques des différents variants. Après avoir construit des pangénomes avec des méthodes existantes, une partie importante du stage consistera à définir les indicateurs permettant la caractérisation du graphe et à implémenter des méthodes pour les calculer (parcours et analyse du graphe de variations). La dernière étape consistera à comparer ces caractéristiques entre les différents graphes construits, c'est-à-dire entre les différentes espèces et/ou entre les différentes méthodes de construction de graphes de pangénomes.

- Niveau de recrutement: Master 2 ou équivalent
- Cursus: informatique ou bio-informatique
- Rémunération: gratification (environ 545€/mois)
- Perspectives : Un financement de thèse est acquis sur cette thématique (projet AGRODIV), la poursuite de ce sujet de recherche en thèse pourrait dont être envisagée.
- Lieu d'accueil: GenPhySE, INRAE Occitanie – Toulouse, 24, chemin de Borde-Rouge, Auzeville-Tolosan *ou* GenScale Team, Inria Rennes Bretagne Atlantique, Campus de Beaulieu, Rennes
- Enacdrant·e·s : Thomas Faraut (INRAE, Toulouse), Claire Lemaitre (Inria, Rennes), Matthias Zytnicki (INRAE, Toulouse)
- Mots-clefs: assemblage, pan-génome, graphe de variations, variants de structure.

References

[1] C. Jain, N. Tavakoli, and S. Aluru. A variant selection framework for genome graphs. *Bioinformatics*, 37(Supplement_1):i460–i467, July 2021. doi:10.1093/bioinformatics/btab302.

[2] J. Sirén, J. Monlong, X. Chang, A. M. Novak, J. M. Eizenga, C. Markello, J. A. Sibbesen, G. Hickey, P.-C. Chang, A. Carroll, N. Gupta, S. Gabriel, T. W. Blackwell, A. Ratan, K. D. Taylor, S. S. Rich, J. I. Rotter, D. Haussler, E. Garrison, and B. Paten. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, 374(6574), Dec. 2021. doi:10.1126/science.abg8871.