Validation de variants de structure à l'aide de données Linked-Reads

Stage de Master 1 IL - ISTIC - 2020-2021

Contexte

Depuis le milieu des années 2000, et l'émergence des séquenceurs à très haut débit, la biologie doit faire face au traitement d'un large volume de données numériques. Ces données sont formées par des millions de courtes séquences, de 50 à 300 caractères, appelées lectures ou *reads*. Dans les années 2010, des séquenceurs de troisième génération, permettant de séquencer des lectures longues de plusieurs milliers de caractères, au prix d'un taux d'erreurs plus élevé, ont émergé. Plus récemment, une nouvelle technologie, permettant de combiner la haute qualité des lectures courtes à l'information longue distance portée par les lectures longues, a été développée. Ces données, appelées *Linked-Reads*, reposent sur un identifiant appelé *barcode*, permettant de reconstruire l'information longue distance en déterminant l'origine commune des lectures.

L'exploitation de ces données a rendu possible une large variété d'applications, et notamment l'étude de variants de structure. Ces derniers correspondent à une modification de la structure d'un chromosome d'un individu, par rapport à une génome de référence. Ces variations peuvent être associées à des maladies génétiques, peuvent correspondre à de multiple types de modifications, tels que des délétions, des duplications, des inversions, des insertions, ou des translocations.

De nombreuses méthodes algorithmiques permettant de détecter ces variants, et tirant parti des spécificités de chaque type de lecture ont été développées. Manta [1] et SvABA [4] permettent par exemple d'identifier des variants à l'aide de lectures courtes classiques, tandis que LinkedSV [2] et VALOR2 [3] permettent d'identifier des variants à partir de données *Linked-Reads*. Ces méthodes, bien que permettant d'identifier convenablement un certain nombre de variants, retournent aussi des faux-positifs, pouvant altérer les analyses sous-jacentes.

Sujet

Le sujet de ce stage consiste à appliquer des méthodes lectures courtes classiques afin de détecter des variants, puis à utiliser en complément l'information longue distance portée par les données *Linked-Reads*, afin de détecter les variants correctement identifiés. Une telle approche de validation représente un axe novateur d'exploitation des données *Linked-Reads*, et pourrait permettre de grandement réduire l'impact des faux-positifs.

D'un point de vue informatique, le stage inclura notamment de la programmation en Python, de la rédaction de scripts permettant de parser et d'exploiter l'information contenue dans des fichiers d'alignements, ainsi que du développement d'algorithmes permettant de rechercher et combiner différents signaux, provenant à la fois de données lectures courtes et de données *Linked-Reads*.

Encadrant

Pierre Morisse

References

[1] Xiaoyu Chen et al. "Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications". In: *Bioinformatics* 32.8 (2016), pp. 1220–1222.

[2] Li Fang et al. "LinkedSV for detection of mosaic structural variants from linked-read exome and genome sequencing data". In: *Nature Communications* 10.1 (2019).

[3] Fatih Karaoğlanoğlu et al. "VALOR2: characterization of large-scale structural variants using linked-reads". In: *Genome Biology* 21.1 (2020).

[4] Jeremiah A Wala et al. "SvABA: genome-wide detection of structural variants and indels by local assembly". In: *Genome Research* 28.4 (2018), pp. 581–591.