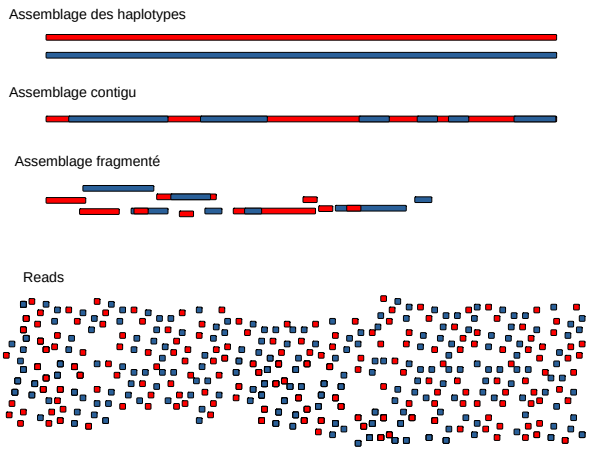
Utilisation d'alignements multiples pour la séparation d'haplotypes à partir de données de séquences d'ADN de troisième génération

25 octobre 2021

Contexte

En 2021, vingt ans après le premier séquençage du génome humain et trois générations de technologies de séquençage, un consortium a réussi à reconstituer les séquences complètes des 23 chromosomes d'un génome humain. Cette performance a été rendue possible notamment par le séquençage de troisième génération, qui produit des longues lectures (ou "long reads"). Ces long reads permettent de distinguer les différentes occurrences des répétitions génomiques et ainsi de reconstituer des séquences génomique a l'échelle des chromosomes. Mais il reste encore un problème non traité de manière satisfaisante : la plupart des génomes sont reconstitués en ignorant la variabilité interchromosomique (hétérozygotie). Or les informations locales (de type SNP, Single Nucleotide Polymorphism) observées entre les allèles d'un même individu ont un impact important sur le fonctionnement du génome. Là encore, les long reads pourraient permettre de progresser par rapport au séquençage de reads courts, en phasant les SNP. Mais, les long reads présentent des taux d'erreurs importants, proches de 10%, avec à la fois des substitutions, des insertions et des délétions. Le traitement de ces reads nécessite donc le développement de nouvelles méthodes bioinformatiques pour en tirer parti efficacement. Pour cela, de nombreuses méthodes s'appuient sur des techniques d'alignement multiple pour détecter les nucléotides erronés et construire une séquence consensus contenant le moins d'erreurs possible. L'objet de ce stage est de généraliser ces approches au cas diploide ou polyploide, pour reconstituer plusieurs séquence consensus représentant les différentes allèles en présence.



L'information issue de séquençage d'un génome peut être représente de différentes manière. L'ensemble des reads est très redondant et très fragmenté. Un assemblage, même fragmenté, permet de s'abstraire de la redondance d'un séquençage. Un assemblage contigu, reconstruit a partir de long reads, permet d'avoir une meilleure vision de la structure du génome mais mélange les séquences des différents haplotypes. La représentation idéale est l'assemblage de représenter chaque haplotype par sa propre séquence.

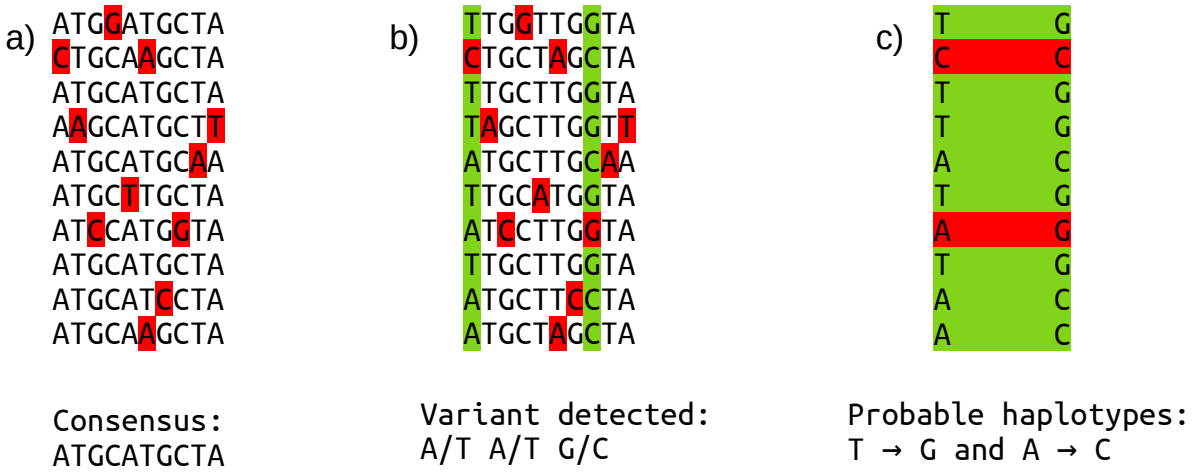


Figure 2 : Trois idées principales utilisée pour la reconstitution d'haplotypes. a) L'utilisation d'alignement multiple pour créer une séquence consensus exempte d'erreurs. b) Détection des positions ou plusieurs nucléotides semblent génomiques.

Projet

De nombreux outils ont recours a des techniques d'alignements multiples entre reads pour détecter et corriger les erreurs de séquençage des séquences bruitées. (voir figure 2a). En utilisant cette approche, il est également possible de profiter de la couverture de séquençage pour identifier les variant présents au sein des séquences (voir figure 2b). L'idée de ce projet est d'aller plus loin et d'essayer de reconstruire les haplotypes en reliant les différents variants a partir de leur co-occurrence dans les reads (voir figure 3c).

L'objectif du stage est ainsi de participer au développement d'une solution algorithmique pour ce problème, en allant jusqu'à la mise en œuvre en Python ou dans un langage compilé (C/C++/Rust . . .). Le stage sera donc l'occasion d'étudier en profondeur des aspects algorithmiques fondamentaux ainsi qu'une expérience en développement. De plus, les méthodes développées seront amenées a être testées sur des données simulées et réelles avec l'objectif d'être intégrés au sein d'outil bioinformatique permettant la séparation d'haplotypes a partir de données de séquençage de troisième génération.

Contact

Antoine Limasset (antoine.limasset@univ-lille.fr)
Hélène Touzet (helene.touzet@univ-lille.fr)