COMPTE RENDU VIROMATHON, LIRMM, MONTPELLIER, 18 OCTOBRE 2018.

Journée organisée par Annie Château et Marie-Agnès Petit, dans le cadre du réseau phage.fr

Participants : Cédric Midoux (Irstea), Guillaume L'Hostis (Pherecydes-Pharma), Eric Rivals (LIRMM), Marie-Agnès Petit (INRA), Annie Château (LIRMM).

Les diapos des 4 exposés présentant les résultats détaillés sont disponibles à cet endroit.

Rappel du principe du Viromathon lancé à la dernière journée bio-info : Cinq jeux de données d'assemblage de reads Hiseq 'VP17' (il s'agit d'un virome de feces de porc purifié et séquencé dans l'équipe Phages de Micalis, à l'INRA) ont été générés par les participants, envoyés à Annie Château (LIRMM, Montpellier), qui a procédé à leur comparaison, avec l'aide d'un stagiaire M2, Peter Bock. En 'référence' des long reads pacbio, produits à partir du même échantillon d'ADN ont été fournis à Annie.

Données de départ :

- les reads eux-mêmes (voir topo readPacbioVP17, de Marie-Agnès Petit, 1ère partie): 50 millions de reads, 2x 150 bp (c'est beaucoup, mais vu la redondance, pas inutile). Fait à la plateforme INRA de Toulouse, génotoul (demandait 600 ng d'ADN, étiquetage Trueseq, de nos jours on s'en sort à moins). Jeu de données très redondant, dû à une amplification génomiphi des échantillons, nécessaire pour produire 600 ng, les échantillons avaient entre 75 et 400 ng au Qbit). Pas de pb particulier relevé de contamination / faible qualité des reads. En revanche gros pb de redondance.

- trois manières différentes de prétraiter les reads (voir topos estimating et methods d'Annie Château): avec le script normalize-by-median.py dans khmer, ne retient que 4% des reads (ce script revient à dérépliquer les reads, et a permis de faire tourner Metaspades en 30 min contre …infini sans cette étape), BBduk pour enlever les reads contaminés par les étiquettes, et la méthode de Mende et Bork pour bien trimmer les bords. Pas vu d'effet massif lié à ces divers prétraitements, sauf que metaspades fait une kyrielle de petits contigs (on passe de 12 000 à 44 000 contigs si on ne normalise pas les reads).

- trois assembleurs dans la course : CLC, Megahit, et Meta spades, tous basés sur les graphes de Bruijn. CLC ne fait pas varier k. Il le fixe en fonction du nombre de reads.

Conclusions très résumées sur les comparaisons d'assembleurs

Voir diapos de Cédric Midoux et Annie Château (topo 2).

Tous les assembleurs ont bien travaillé, beaucoup de contigs communs. CLC plutôt moins bon que les deux autres. Plus grand contig trouvé 110 kb dans tous les cas, et 1000 contigs de plus de 1 kb. Assignation des contigs générés avec Kaiju : le virome de porc est riche en Gokushoviridés de Bacteroides. Plusieurs phages tempérés d'espèces du gut. Un long read d'Apicomplexe intriguant.

Analyse des reads pacbio

Voir topo de Marie-Agnès Petit (deuxième partie).

22 377 reads, taille moyenne 7 kb. Deux séries de problèmes :

- taux d'erreur élevé (jusqu'à 40%), donc usage des reads illumina pour corriger ces erreurs, avec le logiciel Lordec .
- chimères abondantes (sur les reads les plus longs). Pb non résolu.

Conclusion sur les reads pacbio : il reste encore du travail pour qu'ils soient utilisables comme référence. Actuellement, seuls 1% de ces reads s'alignent avec les contigs illumina (ce qui représente environ 2/3 des contigs illumina de taille supérieure à 1 kb).

Le travail continue, on cherche à repérer (1) les contigs issus d'assemblages qui seraient chimériques, en prenant les longs reads « nettoyés » comme référence (2) les longs reads Pacbio mais repérés comme phagiques (Virsorter) qui n'auraient pas de « petit frère » dans les contigs issus d'assembleurs, pour cause de difficulté d'assemblage. Ce projet a donc une suite , et ses participants souhaitent en tirer une publication.