PRINCIPALES CONCLUSIONS DE LA JOURNEE « ASSEMBLAGES DE VIROMES »

- Les méthodes permettant d'estimer la qualité des assemblages sont validées uniquement sur données simulées. Sur les viromes « naturels », il n'existe que des mesures pas forcément pertinentes sur ces assemblages (N50).
- Ce qui est sûr c'est qu'il y a au moins deux raisons qui rendent ces assemblages particulièrement difficiles pour les séquences de phages :
    - la grande microdiversité des séquences. A ce propos le papier Martinez-Hernandez et al (Nature Comm 2017) est particulièrement éclairant (notamment figure 6).
    - le mosaïcisme des phages tempérés, c'est-à-dire la présence de blocs d'ADN quasi-identiques entre génomes de phages, ce qui revient au même que le problème des répétitions dans les génomes.
- Une question soulevée par François Enault : cette microdiversité est-elle le résultat d'un niveau élevé de mutagénèse spontanée, ou de la diversité des phages combinée à la grande taille des populations de phages ? Pas de réponse tranchée dans le groupe…
- Une méthode prometteuse, présentée par Sébastien Halary (eq. C. Desnues à Marseille) qui tient compte du problème de microdiversité, consiste à renoncer à assembler, mais à créer plutôt des clusters de petits contigs très probablement chevauchants (reconnus par une analyse de kmers), et à représenter les résultats sous forme de graphes.
- Une autre méthode, décrite dans Nielsen et al (Nature Biotechno, 2014), et poursuivie par MA Petit, consiste à construire un catalogue de gènes viraux le plus complet possible de l'environnement étudié, puis à faire l'analyse de la matrice d'abondance des reads sur ce catalogue, en cherchant des groupes de gènes co-abondants (CAGs), pour finalement tenter l'assemblage du sous-ensemble des reads correspondant à ces CAGs.
- Nous proposons un « viromathon » à partir d'un jeu de données réel, un virome de feces de porc qui a été séquencé en Hiseq (pair-end) et Pacbio (les reads pacbio serviront de référence). Les reads Hiseqs seront délivrés soit au format brut (chaque participant nettoie les données comme il l'entend) soit au format nettoyé par la méthode de Mende *et al.* 2012 (méthode qui permet d'éliminer des reads les bases issues des premiers cycles Illumina qui sont toujours biaisés). Il n'y a pas besoin de développer une nouvelle méthode pour participer, c'est intéressant de comparer aussi tous les logiciels existants (spades, mira, velvet, CLC, minia…)
- Chaque participant tente un assemblage/clustering/descriptif du contenu en phages de ce virome. Le fichier de sortie doit être un multifasta des contigs (ou clusters) générés, avec leur profondeur dans le header. Annie Château récupèrera les résultats pour en faire la comparaison.
- Date limite pour s'inscrire : 20 novembre 2017 (début du meeting phage.fr) en envoyant un mail à marie-agnes.petit@inra.fr
- Date pour rendre le travail : 31 décembre 2017. En plus du fichier multi-fasta, rédiger le mat et meth correspondant (en vue d'une publication).
- Sous réserve qu'un financement soit trouvé, un stagiaire de M2 prendra en charge l'analyse, sous la supervision d'Annie Château (lirm, Montpellier), qui rendra ses conclusions en juin 2018.