Structure d'accueil : ASTRE

Localisation : Cirad , Campus international de Baillarguet. 34398 Montpellier

Contexte du stage

L'évolution récente des approches métagénomiques permet aujourd'hui l'étude de communautés virales complexes. Ce type d'approche appliqué aux arthropodes vecteurs de maladies permet l'acquisition de connaissances clés dans la prévention et le contrôle des agents pathogènes au travers d'une meilleur compréhension des interactions au sein de ces communauté virale et avec leurs environnements. Notre équipe se concentre sur l'étude des communautés virales chez certaines espèces de moustiques, provenant d'Afrique et de France, qui sont vecteurs de maladies humaines et vétérinaire. L'objectif à court terme est de mieux appréhender la diversité des virus eucaryotes ainsi que les différents facteurs qui peuvent structurer cette diversité (espèce de moustique, continent, habitat, etc.). Ces connaissances seront ensuite utilisées pour étudier la potentielle influence de ces communautés virales sur la transmission d'agents pathogènes.

Pour permettre ce type d'analyses de métagénomique, l'équipe a mis au point un protocole de préparation de libraire à haut débit qui permet un séquençage en parallèle de nombreuses librairies à un coût limité. De plus, un pipeline bioinformatique en snakemake, 'Snakevir', a été développé pour les analyses de métagénomique virale. Ce pipeline suit les principales étapes dans les pipelines pour la métagénomique virale et inclut des optimisations pour notre approche. Dans un premier temps les reads vont être nettoyés et filtrés. Ces reads sont ensuite assemblés à l'aide des outils Megahit & Cap3. Afin d'améliorer la comparaison entre les échantillons, un 'méta-assemblage' est réalisé en utilisant tous les reads de toutes les librairies. Une recherche d'homologie entre les contigs et les séquences issues de base donnée publique est ensuite réalisée.

Objectifs et contenu du stage :

Cependant, la méthode de 'méta-assemblage' avec megahit et Cap3 , bien que souvent utilisée dans la littérature, a toutefois quelques problèmes. En effet, des anomalies d'assemblage sont rencontrées. Certains contigs correspondant à des virus connus ont des séquences répétées aux extrémités. D'autres séquences ont leurs extrémités inversées par rapport aux séquences des bases de données. Un travail de vérification des contigs, très chronophage, est donc requis pour nettoyer ces chimères avant la soumission des séquences virales identifiées aux bases de données publiques. L'objectif de ce stage sera de limiter la production de chimères par une optimisation de l'assemblage.

Pour cela l'étudiant sera amené à :

(1) Tester différents paramètres d'assemblage pour l'outil déjà étudié au sein de l'équipe (Megahit)

(2) Chercher, tester et intégrer d'autre assembleur de métagénomique virales tels que metaSPAdes.

(3) Si l'étudiant le souhaite, il pourra aussi être amené à améliorer et ajouter de nouvelle fonctionnalité telle que la création d'un rapport, la création de fichier nécessaire à la soumission de séquences sur NCBI, etc.

Compétences demandées :

Formation recherchée : M1 ou première année d'école d'ingénieur avec parcours bio-informatique.

- Maîtrise du langage bash
- Bonnes pratiques de développement de code.
- Des connaissances en analyses métagénomiques serait appréciées
- Des connaissances en langages de programmation tels que python ou snakemake serait un plus.

Contact :

* Florian CHARRIAT : florian.charriat@cirad.fr

* Antoni EXBRAYAT : antoni.exbrayat@cirad.fr

* Serafin GUTIERREZ : serafin.gutierrez@cirad.fr