

PROPOSITION DE STAGE M2 (6 mois) 2019-2020

Laboratoire : INRA/GAFL - Avignon

Equipe Résistance Diversité et Durabilité (ReDD)

<https://www6.paca.inra.fr/gafl/>

Nom des Maîtres de stage : Nathalie Boissot (CR) et Jacques Lagnel (IE)

- Tel : 04 32 72 27 23 - email : nathalie.boissot@inra.fr et jacques.lagnel@inra.fr

Projet de stage :

Assemblage hybride long/short reads, carte optique de quatre génomes du puceron du melon et du cotonnier *Aphis gossypii*

Annotations automatiques structurelle et fonctionnelle

Sujet : Malgré le nombre considérable d'espèces d'insectes, peu ont été séquencées, du fait d'abord de leur petite taille. De plus, les génomes d'insectes peuvent être difficiles à assembler en raison de la combinaison d'un polymorphisme élevé, de leur hétérozygotie, la présence de régions répétées, et de la mise en commun d'individus polymorphes pour constituer des banques. Aucun génome hautement résolu n'est disponible pour les pucerons, ravageurs majeurs des plantes cultivées. Nous proposons de relever ce défi pour *Aphis gossypii* ravageur majeur des Cucurbitacées, cotonnier, citrus, ...

Une approche possible la combinaison de données issues de différentes technologies, en associant des lectures courtes (Illumina + 10X genomics) et longues (Oxford Nanopore), aussi bien que des cartes optiques, afin de profiter des avantages de ces méthodes et pallier les inconvénients de chacune (taille des lectures, type d'erreur).

Objectifs du stage :

Dans ce contexte, sur des données de séquençages (10X Genomics et Oxford Nanopore Technology) et de « digital molecules » (carte optique Bionano), il s'agira d'implémenter et de comparer diverses stratégies d'assemblage afin d'obtenir des génomes le moins fragmenté possible (avec les logiciels Canu, Smartdenovo, Flye, RA-assembler, Supernova).

Une annotation structurelle (abinitio gene model prediction) et fonctionnelle sera conduite. Des données RNAseq disponibles au laboratoire seront aussi utilisées pour l'annotation.

Le stagiaire aura accès au serveur Linux de calcul de l'unité GAFL et aux plateformes HPC de l'INRA.

L'étudiant/e sera co-encadré par un ingénieur en bio-informatique et un chercheur en génétique. Il sera en relation étroite avec un doctorant travaillant sur l'interaction plante puceron.

Maîtrise de l'environnement Linux et du Bash, l'utilisation et interprétation de logiciels d'analyse bio-informatique et des bases sur les techniques d'assemblage *de novo* de génome sont demandés. De plus, des notions d'outils de packaging (Singularity) et de management de workflow (Snakemake, Nextflow) seraient un plus. Des connaissances et/ou une pratique du calcul sur cluster seraient appréciées.

Références :

1. Belser, C. *et al.* Chromosome-scale assemblies of plant genomes using nanopore long reads and optical maps. *Nat Plants* **4**, 879–887 (2018).
2. Schmidt, M. H.-W. *et al.* De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *Plant Cell* **29**, 2336–2348 (2017).
3. Gautier, M. *et al.* The Genomic Basis of Color Pattern Polymorphism in the Harlequin Ladybird. *Curr. Biol.* (2018). doi:10.1016/j.cub.2018.08.023