US1426 GeT-Plage
Plateforme GeT de Genotoul
Centre INRAE, Auzeville CS52627, 31326 Castanet-Tolosan Cedex
http://get.genotoul.fr
@get_genotoul

Directrice : C. Donnadieu
Directeur Scientifique : D. Milan
get-plage.direction@genotoul.fr

Stage Ingénieur Bioinformatique M2

Développement d'un pipeline d'analyse de données PacBio Iso-Seq

Lieu de travail : Campus INRAe d'Auzeville (31)
Durée : 6 mois
Contacts : claire.kuchly@inrae.fr
celine.vandecasteele@inrae.fr

Environnement :

- L'activité s'exercera au sein de l'unité GeT-PlaGe du Centre de Recherche INRAe de Toulouse. Get-PlaGe est le site principal de la plateforme GeT de Genotoul (get.genotoul.fr). Constituée d'une trentaine de personnes, elle a pour vocation de permettre l'accès de la communauté scientifique à des équipements performants, et de favoriser les transferts de savoir-faire dans les domaines de la génomique et de la transcriptomique.
- La plateforme est depuis 2010 une infrastructure d'avenir dans le cadre du programme France Génomique. Elle est labélisée IBiSA (Infrastructures en Biologie Sante et Agronomie) depuis 2008 et est certifiée ISO9001-2015 et NFX50-900. Elle fait partie des plateformes stratégiques CNOC (Commission Nationale des Outils Collectifs) d'INRAe. GeT-PlaGe a un partenariat historique avec la plateforme bioinformatique de GenoToul pour la gestion, le stockage et l'analyse primaire des données de séquençage nouvelle génération (NGS) .
- Le stage se déroulera dans l'équipe de (bio-)informatique de l'unité (8 personnes) en interaction avec les personnes impliquées dans le séquençage nouvelle génération et la plateforme Bioinformatique de Genotoul.

Contexte :

La plateforme est équipée de plusieurs machines de séquençage dont le Sequel2 de Pacific Biosciences. Utilisée principalement pour du séquençage de génome complet, nous aimerions développer d'autres applications sur cette technologie. L'approche Iso-Seq (Isoform sequencing) de PacBio utilise les longues lectures pour séquencer les transcrits pleine longueur permettant ainsi d'améliorer l'annotation et de mieux identifier les différents isoformes. Dans le cadre d'un projet de recherche en collaboration, nous allons mettre en place ce protocole sur 2 espèces modèles. C'est dans ce contexte que nous proposons un stage M2 de 6 mois dans l'équipe de bioinformatique de la plateforme.

Objectif du stage :

L'objectif du stage sera de mettre en place une chaine de traitement de ces données Iso-Seq pour mettre à disposition de nos chercheurs des résultats d'analyse sur notre interface NG6 (ng6.toulouse.inra.fr). Le travail consistera principalement à développer un pipeline sur un cluster de calcul sous SLURM à l'aide d'outils bioinformatiques existants. Pour cela, il sera nécessaire de réaliser un benchmarking de ces outils.

Compétences requises :

L'étudiant(e) devra connaître l'environnement Linux et être familier avec un gestionnaire de workflows (de préférence nextflow) ainsi que les différents formats classiques de fichiers bioinformatiques associés (fastQ, BAM/SAM, BED, VCF, …). Ce travail nécessitera l'utilisation de langages de programmation pour la manipulation (Shell, Python, Perl) et l'analyse statistique (R) des données. Ce travail sera fait en interaction avec les biologistes et les équipes de recherche impliquées dans ce développement. Une bonne capacité de communication et de travail en équipe est donc essentielle pour ce stage. Il sera amené à présenter ses travaux à l'ensemble de la plateforme et devra rédiger des documents permettant l'utilisation de la solution mise en place.