Offre de thèse sur

« Intégration de données multi-omiques appariées à grande échelle »

Contexte et résumé : Ce projet de thèse s'inscrit dans le cadre du projet cible AgroDiv du PEPR Agroécologie et Numérique qui ambitionne une caractérisation fine de la variabilité génétique sur la base d'une collection étendue d'espèces botaniques et animales qui pourraient présenter un intérêt pour l'agriculture, particulièrement dans un contexte de mutation environnementale forte et de changement climatique.

Un grand nombre de méthodes existent pour réaliser la détection d'associations entre variants génétiques et des caractères moléculaires (e.g. expression quantitative trait loci, eQTL) à l'échelle d'une population. Cependant, une question importante est de comprendre la stabilité ou la spécificité des effets de ces variants régulateurs à travers plusieurs sous-populations contrastées (e.g., les races chez des animaux domestiques ou les écotypes chez les végétaux). Les données multi-omiques (génomique, transcriptomique) collectées sur plusieurs sous-populations offrent une bonne opportunité pour répondre à cette question, mais à ce jour peu d'approches existent permettant d'identifier des associations spécifiques tout en capitalisant simultanément sur l'ensemble des données. Cette thèse se focalisera sur le développement d'une méthode statistique pour la recherche d'associations variants-expression génique qui permettent de caractériser spécifiquement une sous-population ou qui, au contraire, sont partagés entre plusieurs sous-populations. D'un point de vue méthodologique, il sera nécessaire, pour utiliser au mieux l'ensemble des données disponibles, de traiter cette question globalement car réaliser des études d'association indépendantes dans chacune des sous-populations conduirait à un manque de puissance dans les résultats obtenus. C'est un modèle global de ce type, adapté au cadre de sous-populations animales ou végétales d'une même espèce et permettant un passage à l'échelle du transcriptome entier que cette thèse ambitionne de développer.

Description détaillée du programme de travail : Dans ce cadre, le/la doctorant·e sera en charge de la mise au point de méthodes d'intégration de données omiques pour identifier les relations entre marques génétiques d'intérêt et signaux transcriptomiques, spécifiques ou partagés entre plusieurs sous-populations. Ces méthodes seront la base pour la définition de relations spécifiques à une espèce donnée ou, au contraire, conservées entre espèces et permettront une meilleure caractérisation de la variabilité des phénomènes de régulation.

Pour motiver les développements méthodologiques de cette thèse, le/la doctorant·e s'appuiera sur des données chez le porc qui ont été générées et précédemment analysées[1] dans le cadre du projet H2020 GENE-SWitCH. Brièvement, des données transcriptomiques (RNA-seq) dans trois tissus (muscles, duodenum, foie) couplées avec le séquençage du génome entier ont été collectées pour 3 races commerciales (Large White, Landrace, Duroc), avec *n*=100 animaux par race. L'analyse eQTL présentées dans Crespo-Piazuelo *et al.* (2023) s'est focalisé sur un modèle global pour les 3 races, et n'a pas pu mettre en évidence des associations spécifiques à l'une ou plusieurs d'entre elles.

Après une revue bibliographique approfondie de l'état-de-l'art pour les méthodes intégratives de données multi-omiques appariées, le/la doctorant·e procédera par les trois étapes suivantes :

(1) Construction d'un **plan de simulation réaliste adapté aux données génomiques-transcriptomiques multi-populationnelles** (e.g., incorporant des différences de fréquence allélique et/ou des effets d'association variant d'intensité entre races) pour constituer une base de validation pour des associations spécifiques ou communes.
(2) Définition, implémentation et validation d'une **méthode statistique pour relier des variants génétiques à l'expression de gènes** à partir de données génomiques-transcriptomiques multi-populationnelles.
(3) Identification d'**associations spécifiques à une population d'intérêt ou conservées à travers plusieurs sous-populations**.

Conditions de réalisation de la thèse : La thèse sera réalisée sur le site INRAE de Jouy-en-Josas dans l'Unité GABI (Génétique Animale et Biologie Intégrative). Il sera toutefois possible de la localiser alternativement sur le site INRAE de Toulouse dans l'Unité MIAT (Mathématiques et Informatique Appliqués à Toulouse). Dans les deux cas, l'équipe d'encadrement sera composée d'Andrea Rau (GABI) et de Nathalie Vialaneix (MIAT). Nous nous appuierons notamment sur des réunions hebdomadaires en visioconférence ainsi que des outils collaboratifs (e.g., Mattermost pour des messages quotidiens et les comptes rendus des réunions hebdomadaires, Gitlab pour le partage de scripts R/Python, Nextcloud pour le partage de documents) pour assurer une bonne communication entre le/la doctorant·e et l'équipe d'encadrement. Nous prévoyons également d'organiser ponctuellement mais régulièrement des séjours courts (~1 semaine) dans l'unité complémentaire pour permettre le/la doctorant·e de bénéficier des deux environnements scientifiques.

Formation et compétences recherchées

- Diplôme minimum requis : Master ou équivalent en statistique appliquée, (bio)statistique, ou (bio)mathématiques.
- Formation recommandée : Parcours en mathématiques appliquées ou (bio)statistique.
- Connaissances souhaitées : Statistique (modèles linéaires), analyse de données, programmation (R et/ou Python).
- Expérience appréciée : De l'expérience en analyse de données génomiques et/ou transcriptomique est souhaitable mais pas obligatoire.
- Aptitudes recherchées : intérêt pour la recherche scientifique, capacité de communication en anglais (orale et écrite), rigueur, curiosité. Un goût pour les applications, en particulier en génomique, est absolument indispensable.

Postuler : Envoyer CV, lettre de motivation et notes de master à andrea.rau@inrae.fr et nathalie.vialaneix@inrae.fr.

[1] Crespo-Piazuelo, D., et al. (2023) Identification of transcriptional regulatory variants in pig duodenum, liver, and muscle tissues, GigaScience, Volume 12, 2023, giad042, https://doi.org/10.1093/gigascience/giad042