Stage M2 - Comparaison de procédures statistiques pour des analyses d'enrichissement dans des études transcriptomiques

Description

Les données transcriptomiques nécessitent une modélisation statistique particulière car très peu d'individus sont disponibles pour étudier des milliers voire millions de variables simultanément. Lorsque les effets individuels des gènes sont faibles, des analyses d'enrichissement sont effectuées en tenant compte de groupes de gènes prédéfinis. Les méthodes existant dans la littérature reposent sur deux grandes familles d'approches, la première testant une sur-représentation de certains groupes de gènes connus pour interagir ensemble, la seconde agrégeant des scores individuels (par exemple GSEA: gene set enrichment analysis). Parmi les packages R disponibles pour ce type d'analyse, le package clusterProfiler (Yu et al., 2012, doi: 10.1089/omi.2011.0118; Wu et al., 2021, doi: 10.1016/j.xinn.2021.100141) présente une bonne introduction à ce type d'analyses dans sa vignette.

Le but principal de ce stage de M2 sera d'évaluer l'influence du choix de la métrique dans les analyses de type GSEA, par des analyses de données réelles et des simulations. Un objectif secondaire sera d'améliorer les pipelines de la plateforme bilille pour les analyses d'enrichissement.

Activités principales

- Bibliographie des packages R réalisant des analyses d'enrichissement
- Comparaison de différentes métriques dans les analyses type GSEA (analyses de données réelles, mise en place de simulations, interprétation des résultats, comparaison des résultats)
- Tests d'outils de visualisation de catégories enrichies pour améliorer les rendus aux biologistes
- Proposition d'une procédure automatisée pour la plateforme bilille, avec un soin particulier donné à la documentation

Compétences requises

- Programmation courante en R
- Maîtrise des techniques statistiques exploratoires multivariées et de plusieurs outils d'apprentissage statistique en grande dimension (plus de variables que d'individus)
- Connaissances liées à des études biostatistiques
- Langue anglaise: B2 (cadre européen commun de référence pour les langues) anglais technique du domaine
- Goût pour l'interdisciplinarité

Encadrement et conditions d'accueil

Le ou la stagiaire sera accueilli.e au sein de de la plateforme bilille (UAR 2014 PLBS, https://wikis.univ-lille.fr/bilille/). Il ou elle sera encadré.e principalement par Guillemette Marot, avec des interactions fortes avec les autres ingénieurs de la plateforme.

Le stage sera réalisé en présentiel dans les bureaux de la plateforme, avec la possibilité d'un jour de télétravail par semaine. La localisation principale du stage sera le campus hospitalo-universitaire de Lille, mais des rotations hebdomadaires l'amèneront aussi à travailler sur les campus de l'Institut Pasteur de Lille et de la Cité Scientifique à Villeneuve d'Ascq.

La date de démarrage du stage est à fixer en fonction des contraintes de calendrier du Master du ou de la stagiaire.

Contact

Guillemette Marot, Pierre Péricard et Jimmy Vandel, responsables scientifiques de la plateforme bilille

bilille@univ-lille.fr