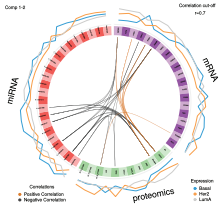
Accompagner l'analyse des données issues des nouvelles bio-technologies

La plateforme de Biostatistique se présente comme un carrefour de compétences en statistique pour la biologie, dédié à la recherche, à la formation et à l'animation scientifique au sein de la communauté scientifique toulousaine. Elle apporte son soutien aux chercheurs en biologie au travers de collaborations et du portage joint de projets appliqués.
La plateforme organise et anime également des formations périodiques ou ponctuelles autour de l'analyse statistique, de l'intégration de données multivariées ou de la maîtrise du langage d'analyse statistique **R**.



Formation

En 2019, la plateforme de Biostatistique a poursuivi ses actions de formation :

- Initiation à la programmation avec R, les 20 et 21 mai 2019.
- Formation mixOmics (analyses multivariées et intégration de données), du 4 au 6 juin 2019.
- Initiation à la statistique avec R, les 9 et 10 septembre 2019.

Les formations récurrentes ou les formations ponctuelles que nous assurons continuent à évoluer à partir des sollicitations que nous recevons et des retours des participants. En particulier, les formations à R intègrent de plus en plus les nouveautés liées au *tidyverse*, un ensemble de *packages* dédié à la science des données dont la popularité est grandissante dans la communauté (bio-)statistique. Le cycle de formations 2020 est d'ores et déjà accessible sur notre site internet.

Un réseau étendu

Les personnes relais de la plateforme de biostatistique sont localisées dans plusieurs laboratoires toulousains :

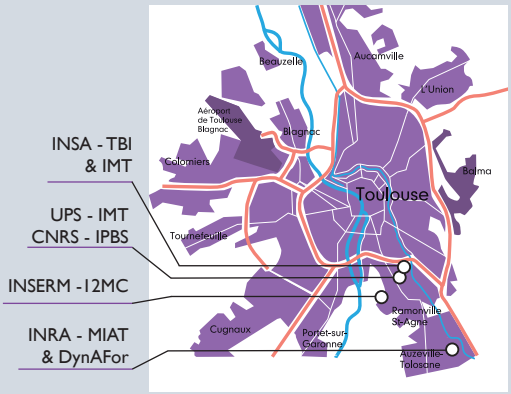
- Marion Aguirrebengoa, ingénieure d'études, CNRS, *Center of Integrative Biology* (CBI),
- Sébastien Déjean, ingénieur de recherche, UPS, Institut de Mathématiques de Toulouse (IMT),
- Jason Iacovoni, ingénieur de recherche, Inserm, Institut de Maladies Métaboliques et Cardio-vasculaires (I2MC),
- Cathy Maugis-Rabusseau, maître de conférences, INSA, IMT,
- David Rengel, ingénieur de recherche, CNRS, *Institute of Pharmacology and Structural Biology* (IPBS),
- Magali San Cristobal, directrice de recherche, INRA, Unité Dynamique et Écologie des Paysages Agriforestiers (DynAFor),
- Nathalie Vialaneix, directrice de recherche, INRA, Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT).

Animateurs :
Sébastien Déjean, David Rengel, Nathalie Vialaneix

Contact :
biostat@math.univ-toulouse.fr

Site web :
https://perso.math.univ-toulouse.fr/biostat

Localisation des équipements

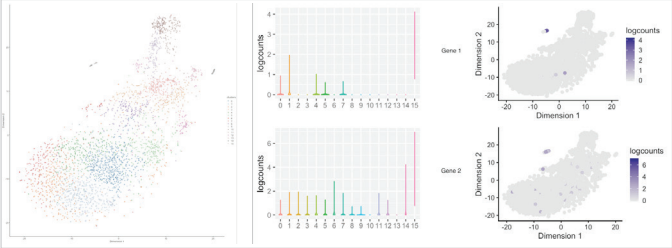


Genotoul Biostat

Le fait marquant scientifique :

L'analyse statistique des données SingleCell RNA-seq

Le SingleCell RNA-seq (scRNAseq) est une technique de séquençage à haut-débit, qui permet de séquencer séparément chaque cellule d'un même échantillon, contrairement au « bulk » RNA-seq qui mesure l'expression des gènes du mélange des différentes cellules. Cette technologie est implantée sur le site toulousain depuis deux ans, en particulier sur la plateforme GeT de Génotoul. Les données scRNAseq ont des points communs avec les données « bulk » RNA-seq mais aussi bien des différences. Il en résulte que les traitements bio-informatiques et les outils statistiques développés dans le cadre des données RNA-seq ne peuvent pas être directement utilisés pour l'étude des données scRNAseq. Le nombre de publications et de logiciels dédiés au traitement des données scRNAseq croît exponentiellement depuis quatre ans. Les méthodes proposées couvrent toutes les étapes usuelles d'analyse : normalisation des données (afin de corriger les biais techniques et rendre comparable les expressions des différentes cellules) ; classification non supervisée des cellules ; détection des gènes marqueurs pour chaque classe de cellules ; analyse différentielle globale ; inférence de trajectoires (classement pseudotempore)...



Exemple d'une classification de cellules de souris en 16 classes, obtenue à partir des log-comptages normalisés. A gauche, représentation t-SNE des cellules selon leur log-comptages en 16 classes. Violin plot de l'expression en log-comptage dans les 16 classes (au centre) et représentation du niveau d'expression sur le t-SNE de deux gènes marqueurs pour la classe 15 (par rapport à toutes les autres classes).

Le projet TTIL SingleCell, porté par deux animatrices de la plateforme biostat, Sandrine Laguerre (TBI) et Cathy Maugis-Rabusseau (IMT), a eu pour objectif de fédérer les différents acteurs (biologistes, bio-informaticiens et statisticiens) sur le pôle toulousain autour de l'analyse des données scRNAseq. Il a permis une collaboration entre le laboratoire STROMALab (qui travaille sur la régénération des tissus et des organes) et l'Institut de Mathématiques de Toulouse. Celle-ci a permis le développement d'une application R/shiny qui permet aux biologistes l'exploitation des résultats de classification et de tester la stabilité des gènes marqueurs identifiés. Grâce au projet TTIL, deux journées thématiques autour du traitement des données scRNAseq ont été organisées en 2018-2019 à Toulouse, avec des exposés variés couvrant toutes les étapes d'analyse. Elles ont permis de favoriser les échanges autour du traitement des données scRNAseq à Toulouse afin de partager nos expériences, faire émerger de nouvelles collaborations entre biologistes, bio-informaticiens et statisticiens, mener une veille active sur les publications et les logiciels dédiés à cette thématique.