CNRS UMR 8199
Génomique Intégrative et Modélisation des Maladies Métaboliques
Directeur : Pr. Philippe FROGUEL
CNRS, Université de Lille, Institut Pasteur et Institut de Biologie de Lille
Fédération de Recherche 3508, Labex EGID

Développement d'une Chaîne de Traitement de Données de RNA-seq

Stage en Biostatistique (Master 2)

L'unité CNRS UMR 8199

Qui sommes-nous ?

L'unité **CNRS UMR 8199** (*Génomique Intégrative et Modélisation des Maladies Métaboliques*) est un laboratoire de recherche regroupant *60 personnes* dirigées par le Professeur Philippe FROGUEL.
Elle fait partie de l'*Institut Européen de Génomique du Diabète* (**EGID**) et a été lauréate en 2011 (renouvelé jusqu'en 2025) des appels à projets « *Laboratoire d'Excellence* » (**LABEX**) et « *Equipement d'Excellence* » (**EQUIPEX LIGAN MP**).

Que faisons-nous ?

Nos activités de recherche portent sur la caractérisation de variations génétiques associées à des maladies métaboliques telles le diabète et l'obésité et utilisent les approches modernes de génomique, bioinformatique, biostatistique, biologie moléculaire et modèles animaux.

Description du stage

Au sein de l'équipe de biostatistique, le/la candidat-e retenu-e devra dans un premier temps se familiariser avec les données de transcriptomique issues du séquençage de l'ARN (RNA-seq) et veillera à sélectionner les méthodes les plus appropriées à l'analyse de ces données.

Dans un second temps, il/elle s'appropriera les méthodes d'analyse statistique sélectionnées (basées sur le modèle linéaire généralisé et les approches de classification) ainsi que les outils pour réaliser ces analyses.
Il/elle adaptera le cas échéant les méthodes existantes en vue de leurs intégration dans une chaîne de traitement, *c.-à-d.*, contrôle-qualité, analyse, annotation et visualisation.

Enfin, l'étudiant-e implémentera une chaîne de traitement des données de RNA-seq, en incorporant les méthodes identifiées, ainsi que les visualisations adéquates pour rendre compte de la nature des données et des résultats obtenus.
L'étudiant-e appliquera ces méthodes et plus généralement la chaîne de traitement sur des données disponibles au sein du laboratoire et/ou sur des données qui pourraient être généré durant la période de stage.

L'implémentation de cette chaîne de traitement des données de RNA-seq se fera avec le langage R, au travers de l'interface de développement Rstudio sur une infrastructure Linux Debian.
L'étudiant-e pourra être amené-e à développer une extension R et à utiliser les extensions R : rmarkdown, shiny et ggplot2.

N° ordre (Philippe FROGUEL) : 590784922
CNRS UMR8199 / EGID – Faculté de Médecine – Pôle Recherche
1 Place de Verdun – Aile Ouest – 1er étage – 59045 LILLE CEDEX
Tél. : 33-(0)3-74-00-81-01 (ou) 81-00 (secrétariat)

Profil recherché

- Formation supérieure (Master 2 ou ingénieur) en (bio)statistique.
- Bonnes connaissances théoriques en statistiques, notamment du modèle linéaire (et ses extensions) et des méthodes de classification.
- Maîtrise du langage R.
- Maîtrise écrite et orale de l'anglais (lecture d'articles scientifiques).
- Connaissances dans le domaine de la transcriptomique serait un atout.

Informations administratives

Durée de stage : Entre 5 et 6 mois pendant l'année académique 2019-2020.
Gratification de stage : Selon les taux en vigueur.
Lieu du stage : Au sein de l'équipe Biostatistique de l'UMR 8199.

L'équipe de biostatistique

Qui sommes-nous ?

L'équipe de biostatistique est en charge des analyses statistiques et contribue à l'élaboration du design des études, notamment dans le cadre des demandes de financements.
L'équipe comporte trois membres et est supervisée par Mickaël Canouil, Ph.D..

Que faisons-nous ?

L'objectif principal de notre équipe consiste à apporter un soutien méthodologique fort aux différentes stratégies proposées par le laboratoire. Cela inclut, des développements méthodologiques et le développement d'outils de visualisation et d'analyse, notamment pour les données issues des technologies de puce et de séquençage à haut débit (*p. ex.*, Methyl-seq, RNA-seq, etc.).

Nos réalisations

Notre expertise dans l'utilisation du programme R et des extensions qui lui sont associées, nous a permis de développer des applications shiny (extension R) pour la visualisation et l'analyse de données de différentes natures et en particulier les données de transcriptomiques (*p. ex.*, NanoString, qPCR, etc.), ainsi que des outils dynamiques de calcul de puissance statistique, dans un rôle de soutien à l'activité de recherche de l'unité.

Notre équipe a également développé des extension R :

- CARoT (« **C**entralised and **A**utomated **R**ep**o**rting **T**ools »), permettant notamment l'estimation des composantes génétiques ethniques, ainsi que le contrôle-qualité des données issues de puces de génotypes et de méthylation ;
- NACHO (« **NA**nostring quality **C**ontrol das**H**b**O**ard »), permettant le contrôle-qualité des données issues de la technologie NanoString, en particulier au moyen d'une application shiny ;
- snpEnrichment, « *SNPs Enrichment Analysis* » ;
- clere, « *Simultaneous Variables Clustering and Regression* ».

Contact

Envoi d'un CV et d'une lettre de motivation à :
Mickaël CANOUIL, Ph.D. (Tél. +33 (0) 374 00 81 29 ; mickael.canouil@cnrs.fr).

N° ordre (Philippe FROGUEL) : 590784922
CNRS UMR8199 / EGID – Faculté de Médecine – Pôle Recherche
1 Place de Verdun – Aile Ouest – 1er étage – 59045 LILLE CEDEX
Tél. : 33-(0)3-74-00-81-01 (ou) 81-00 (secrétariat)

CNRS UMR 8199
Génomique Intégrative et Modélisation des Maladies Métaboliques
Directeur : Pr. Philippe FROGUEL
CNRS, Université de Lille, Institut Pasteur et Institut de Biologie de Lille
Fédération de Recherche 3508, Labex EGID

Development of an RNA-seq Analysis Pipeline

Biostatistics Internship (Master 2)

The CNRS UMR 8199 unit

Who are we?

The **CNRS UMR 8199** (*Integrated Genomics and Metabolic Diseases Modelling*) unit is a research laboratory comprising *60 people* led by Professor Philippe FROGUEL.
It is part of the *European Institute of Diabetes Genomics* (**EGID**) and has been awarded in 2011 (renewed until 2025) for grant "*Laboratory of Excellence*" (**LABEX**) and "*Equipment of Excellence*" (**EQUIPEX LIGAN MP**).

What do we do?

Our research activities focus on the characterisation of genetic variations associated with metabolic diseases such as diabetes and obesity and use modern approaches to genomics, bioinformatics, biostatistics, molecular biology and animal models.

Internship description

Firstly, within the biostatistics team, the candidate will need to familiarise himself with the transcriptomic data from RNA sequencing (RNA-seq) and to select the most appropriate methods for analysing those data.

Secondly, he/she will get to grips with the selected statistical analysis methods (*i.e.*, generalised linear model-based approach and classification methods), as well as the tools implementing those methods.
He/she will adapt, if necessary, the existing methods for their integration into an data processing/analysis pipeline, *i.e.*, quality control, analysis, annotation and visualisation.

Lastly, the student will have to implement an RNA-seq data processing pipeline, incorporating the identified methods, as well as the relevant visualisations to account for the nature of the data and the results.
The student will apply these methods and more generally the pipeline on data available within the laboratory and/or on data that could be generated during the internship.

The implementation of this RNA-seq pipeline will be done in R, through the Rstudio development interface on a Linux Debian infrastructure.
The student may be led to develop an R package and use the R packages: rmarkdown, shiny et ggplot2.

N° ordre (Philippe FROGUEL) : 590784922
CNRS UMR8199 / EGID – Faculté de Médecine – Pôle Recherche
1 Place de Verdun – Aile Ouest – 1er étage – 59045 LILLE CEDEX
Tél. : 33-(0)3-74-00-81-01 (ou) 81-00 (secrétariat)

Intern Requirements

- Master 2 (or engineer) in (bio)statistics
- Good knowledge in statistical analysis, in particular (generalised) linear regressions (and its extensions) and clustering methods.
- R language proficiency.
- English proficiency (scientific articles).
- Knowledge in the field of transcriptomic would be an asset.

Internship information

Internship duration: From 5 to 6 months during the academic year 2019-2020.
Internship salary: Depending on current rates.
Location of the internship: With the biostatistics team in the CNRS UMR 8199 unit.

The Biostatistics team

Who are we?

The biostatistics team is in charge of the statistical analyses and supports the researchers in the design of studies, especially for grant applications.
The team currently counts three members and is supervised by Mickaël Canouil, Ph.D..

What do we do?

The main objective of our team consists in bringing a strong methodological support to the different strategies tackled by the unit.
These strategies go from methodological developments to the development of visualisation tools for data generated through arrays and high-throughput sequencing platforms (*e.g.*, Methyl-seq, RNA-seq, etc.).

Our achievements

Our expertise in the R environment and in the associated packages allowed us to develop several web applications (shiny), used to analyse and to navigate through our data, especially transcriptomic data (*e.g.*, NanoString, qPCR, etc.).

Our team also developed R packages:

- CARoT ("**C**entralised and **A**utomated **R**ep**o**rting **T**ools"), allows to estimate genetic components (*i.e.*, "ethnicity") and quality-control of arrays (genotyping and methylation);
- NACHO ("**NA**nostring quality **C**ontrol das**H**b**O**ard"), allows quality-control of NanoString data, especially through an interactive web application (shiny);
- snpEnrichment, "*SNPs Enrichment Analysis*" ;
- clere, "*Simultaneous Variables Clustering and Regression*".

Contact

Send a CV and cover letter to:
Mickaël CANOUIL, Ph.D. (Tél. +33 (0) 374 00 81 29 ; mickael.canouil@cnrs.fr).