Rapport pour CSI n°2

Florian Privé

1 CV

1.1 Publications

- Privé, Florian, Hugues Aschard, and Michael GB Blum. "Efficient implementation of penalized regression for genetic risk prediction." bioRxiv (2018): 403337. (Preprint: https://doi.org/10.1101/403337)
- Privé, Florian, et al. "Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr." Bioinformatics (2018). (Open Access: https://doi.org/10.1093/bioinformatics/bty185)

1.2 Responsabilités

- Créateur et co-organisateur de "R in Grenoble", un groupe d'utilisateurs de R à Grenoble, se réunissant chaque mois (https://r-in-grenoble.github.io/).
- Participation à l'organisation de 3 Data Challenge (https://goo.gl/Cjvemi, https://goo.gl/rUyBqZ, https://goo.gl/KfpKKS)

1.3 Enseignements

- Mathématiques (cours + TDs) à des L1 scientifiques (128h)
- Formation R avancé (https://privefl.github.io/advr38book/) à des doctorants (30h + 30h cette année)
- Principes et méthodes statistiques à des 1A de l'ENSIMAG (18h de TDs cette année)
- Instructeur Software Carpentry pour un cours d'introduction à R (1 journée)

1.4 Conférences

- Rencontres R 2018: The R package bigstatsr: Memory- and Computation-Efficient Statistical Tools for Big Matrices.
- eRum 2018: An R package for statistical tools with big matrices stored on disk.
- Recomb-Genetics 2018: Predicting complex diseases: performance and robustness.
- LIFE 2018: Predicting complex diseases: performance and robustness.
- hackseq 2017: Developing advanced R tutorials for genomic data analysis.
- useR!2017: The R package bigstatsr: Memory- and Computation-Efficient Tools for Big Matrices.
- ALT'2016: Goodness-of-fit tests for the Weibull distribution with censored data.

2 Formation suivies

Récapitulatif de participation aux Formations
Florian PRIVE

Doctorat : MBS - Modèles, méthodes et algorithmes en biologie, santé et environnement
Ecole Doctorale : Ingénierie pour la santé la Cognition et l'Environnement
Etablissement : Communauté Université Grenoble Alpes
Date de la 1ere inscription en thèse : 7 octobre 2016 (3 A en 2018)
Directeur de thèse : Michael BLUM (EDISCE)
Sujet de thèse : Score de risque génétique utilisant de l'apprentissage statistique.

Formations suivies

Catégorie : Activité du chercheur et pratique du métier

- JEUDIS DE LA SECURITE : Le travail sur écran (15 décembre 2016) PLURIEL - amphi Nord - 701 rue de la Piscine - Domaine universitaire
1.5 heures

Total du nombre d'heures pour la catégorie Activité du chercheur et pratique du métier : *1.5 h*

Catégorie : Formations scientifiques

- Présentation d'un poster à la journée de l'ED (31 mai 2018)
12 heures enregistrées par : Ingénierie pour la santé la Cognition et l'Environnement.

Total du nombre d'heures pour la catégorie Formations scientifiques : *12 h*

Catégorie : Insertion professionnelle

- EQUIVALENCE ATELIER-PROJET
35 heures

Total du nombre d'heures pour la catégorie Insertion professionnelle : *35 h*

Catégorie : Formations du Label RES

- AUT-16 : Introduction au métier d'enseignant-chercheur (07 décembre 2016) Centre l'Escandille
21 heures

Total du nombre d'heures pour la catégorie Formations du Label RES : *21 h*

Catégorie : Langues (Anglais - Français langues étrangères F.L.E.)

- [Anglais] - Rédaction d'articles scientifiques # session 3 (17 janvier 2018) LANSAD - Maison des Langues et des Cultures -1141 Av. Centrale - 38400 Saint-Martin-d'Hères
24 heures

Total du nombre d'heures pour la catégorie Langues (Anglais - Français langues étrangères F.L.E.) : *24 h*

Catégorie : Outils numériques et méthodologiques pour la recherche

- LOGICIEL R AVANCE (28 février 2017) IMAG, 700 avenue Centrale, DU
30 heures

Total du nombre d'heures pour la catégorie Outils numériques et méthodologiques pour la recherche *: 30 h*

Total participation : 123.5 heures / 6 modules

3 Calendrier prévisionnel de fin de thèse

3.1 Fin 2018

- Publication du deuxième papier
- Analyse exploratoire pour le 3ème papier
- Rédaction des premières parties de la thèse

3.2 Janvier - Mai 2019

- Enseignements (R avancé + Cours de statistique à l'Ensimag)
- Analyse finale + rédaction du second papier

3.3 Mai - Juillet 2019

- Soumission du 3ème papier
- Fin de rédaction de la thèse
- Conférence useR!2019 à Toulouse

3.4 Juillet - Septembre 2019

- Répondre aux reviews du 3ème papier
- Vacances
- Soleil
- Champagne

3.5 Fin septembre 2019

- Soutenance
- Champagne

4 Plan de thèse

4.1 Introduction

4.1.1 Genotype data

GWAS SNP data larger and larger.

4.1.2 Polygenic Risk Scores

Combine many SNP into a single score.

4.1.3 What can PRS be used for?

Epidemiology + Prediction

4.1.4 Motivations of the thesis

Using statistical learning methods to improve PRS.

4.2 Methods for deriving PRS (SotA)

4.2.1 Tools for SNP data analysis

4.2.1.1 Imputation

4.2.1.2 Data formats

4.2.1.3 PCA

4.2.1.4 GWAS

4.2.1.5 Paper 1

4.2.2 Deriving PRS

4.2.2.1 Using summary statistics

PRS + LDpred (+ lassosum)

4.2.2.2 Using the whole genome

LMM + Gad + paper 2

4.2.3 Predict more and perspective

4.2.3.1 Generalize to different populations

4.2.3.2 Use other data (environmental, clinical)

4.2.3.3 Ethical aspects?

4.3 Appendix: computational aspects

4.3.1 GWAS tricks

4.3.2 Scaling tricks

4.3.3 memory-mapping / parallel / examples?