M2 Research Internship (2020-2021)

Improving and extending functionality for multi-omic outlier detection software

Andrea Rau, Florence Jaffrézic, and Denis Laloë

Animal Genetics and Integrative Biology (GABI) Research Unit

Address : Domaine de Vilvert 78352 Jouy-en-Josas Cedex

Telephone : 01 60 10 95 35

Email : andrea.rau@inrae.fr ; florence.jaffrezic@inrae.fr; denis.laloe@inrae.fr

Web : https://www6.jouy.inrae.fr/gabi_eng, https://www6.jouy.inrae.fr/gabi_eng/Our-Research/Research-Teams/GiBBS

Research Unit Director: Claire Rogel-Gaillard

Team name: Genomics, Biodiversity, Bioinformatics, Statistics (GiBBS)

Team leader: Denis Laloë

Team research themes: The GiBBS team conducts research in genetics, bioinformatics, and biostatistics, with a particular focus on real problems and applied studies. In particular, our varied research themes include fields associated with genetic diversity and genetic resources for domestic livestock populations, methodological developments for genomic and transcriptomic data (with a special focus on implementation in open-source software packages). The GiBBS team also includes an in-house bioinformatics and biostatistics consulting service to address the analysis needs of biologists and geneticists in the GABI research unit and broader scientific community.

Keywords : Data integration, multi-omics, multivariate methods, R software, Bioconductor

Project context:

Recent developments in high-throughput sequencing technologies have enabled **deep and multifaceted interrogations of the biological variability** in living organisms, including the genome, transcriptome, proteome, and epigenome. In spite of the increasing availability of these multi-level data, our understanding of the function of the genome and its link to phenotypic or physiological characteristics is still incomplete. A major obstacle in the analysis of these large, multi-omic data is the identification of the most suitable way to model them, while accounting for their **high dimensionality**, their **heterogeneous nature**, and the presence of potentially strong redundancies among groups of **highly correlated variables**.

In this context, we recently proposed an unsupervised multivariate method based on a multi-table extension of Principal Component Analysis, the Multiple Factor Analysis (MFA), to characterize individuals with aberrant multi-omic profiles with respect to a reference population[1]. Our approach, which is implemented in the R/Bioconductor package *padma*[2], creates a multi-omic consensus representation for a user-specified pathway of interest, and facilitates a quantification and visualization of gene- and omic-specific deviations from this consensus for each individual in the population.

In this internship, our objective is to **extend and enhance the functionalities of the approach implemented in *padma***:

(1) In our initial work, the reference population was made up of a largely homogeneous group with a small number of outlier individuals. However, in several related applications the reference groups may in fact consist of several well-separated sub-groups; these may be known *a priori* or need to be inferred from the data (i.e., through a clustering of reference population individuals). The R package must be extended to deal with both cases.
(2) Simulations are needed to verify the performance and behavior of *padma* with blocks corresponding to binary data (e.g., somatic mutations), which have not yet been evaluated.
(3) Our implementation of *padma* currently allows for a pathway-centric multi-omic analysis, but there is some interest in generalizing the user-interface to allow for other analysis configurations (e.g., including clinical data, etc).
(4) A Shiny web interface for interactive visualizations of *padma* output is needed for the R/Bioconductor package.

Our initial work focused on the use of *padma* in identifying aberrant breast and lung tumors from populations studied in The Cancer Genome Atlas[3], and the method development in

---

[1] Rau, A., Manansala, R., Flister, M. J., Rui, H., Jaffrézic, F., Laloë, D., and Auer, P. L. (2020) Individualized multi-omic pathway deviation scores using multiple factor analysis. *Biostatistics*, https://doi.org/10.1093/biostatistics/kxaa029.
[2] https://bioconductor.org/packages/padma
[3] https://portal.gdc.cancer.gov

UMR1313 GABI
Animal Genetics and Integrative Biology unit
Domaine de Vilvert
78350 Jouy-en-Josas - France
+33 1 (0) 1 34 65 22 01

this internship will also make use of these data as well as those from in-house agronomic applications.

Required skills:

- Solid knowledge of R programming (including but not limited to S4 classes and methods, unit testing, Shiny applications) and familiarity with version control using Git/GitHub;
- Knowledge of and experience with multivariate statistical methods (principal components analysis and related methods) and Bioconductor would be appreciated but are not required;
- Comfortable reading/writing English (ability to read and understand scientific articles and write software documentation);
- Candidates should have motivation and interest for genomic applications in general, and multi-omics data integration in particular, but experience in these areas is not required.

The research internship will take place in a research environment that brings together bioinformaticians, biostatisticians, and biologists. The research intern will have a choice of working at one of two Inrae research centers: the Ile-de-France–Jouy-en-Josas–Antony Research Center[4] located in Jouy en Josas (78), or the Hauts-de-France Research Center[5] in Estrées-Mons (80). The work will be jointly supervised by Andrea Rau, Florence Jaffrézic and Denis Laloë, with a combination of remote and on-site supervision.

Remuneration:

- Standard research internship salary (approximately 580 € / month)

[4] https://www.inrae.fr/centres/ile-france-jouy-josas-antony
[5] https://www.inrae.fr/en/centres/hauts-de-france

UMR1313 GABI
Animal Genetics and Integrative Biology unit
Domaine de Vilvert
78350 Jouy-en-Josas - France
+33 1 (0) 1 34 65 22 01