

PEPR digital health
Axis Methods and Models for Multimodal Data Integration (M4DI)

PhD position

Extracting genotype-phenotype associations from network integration of deep and multiscale phenotypic data

Keywords

Multilayer network, clustering, genomic data, electronic health records (EHR)

IRP Summary

- Scientific summary of the project

Background: Deep phenotypic data are available in health databases through medical questionnaires and electronic health records (EHR), which can include drug prescriptions, lab results, information extracted from notes using natural language processing, or billing codes. Informative phenotypes are however often not directly available. There are two categories of methods to define disease phenotypes from health databases. First, methods relying on a predefined validated algorithm specifically created for a given phenotype to identify cases and controls (expert approach using metadata). Second, data-driven methods relying on automated approaches. Network-based approaches belong to this second category and rely on similarity measures between individuals to identify homogeneous subgroups (phenotypic clusters).

Objectives: We recently developed an unsupervised network approach to analyze drug prescriptions from the EGB, a French medico-administrative database (~660k individuals) and identified clinically relevant patient subgroups [1], [2]. We aim to extend this work by integrating a broader range of multiscale phenotypic and genomic data. We aim to extract clinically meaningful information, to identify patient subgroups and to correlate these subgroups with genomic variations.

Methods: We will first extend our initial approach to multiscale phenotypes by building a multiplex network integrating the different levels of phenotypic information (drugs, diagnoses, lab tests...). Each layer will contain nodes as individuals and edges corresponding to the similarity between individuals for this level of information. We will then explore the multiplex networks to identify clusters of phenotypically homogeneous

individuals. Finally, we will identify the associations with genomic data using a supervised learning approach such as model-based clustering algorithms [3].

Data/Application: We are involved in different national programs (INSERM cross-cutting program in genomics/France Genomic Medicine 2025) with access to both phenotype and genotype data from the Constances cohort and POPGEN: 10k individuals from France. We further have access to data from the UK Biobank: 450k individuals.

Expected results: We expect to develop new methods and software that would help to i) identify homogeneous phenotypic subgroups of patients and ii) decipher the genetic bases of these subgroups.

- The PhD students will be part of the doctoral network of the project M4DI (<https://m4di.univ-amu.fr/>), one of the axes of the PEPR digital health (<https://pepr-santenum.fr/en/>). As such, they will have a host lab together with funding for a research stay of about 4 months in a secondment lab. The PhD student will further have the opportunity to participate in data challenges and collaborate with other axes of the PEPR digital health.

Host lab: INSERM NeuroDiderot & INSERM-INRIA HeKa, Paris

NeuroDiderot is an interdisciplinary Inserm institute dedicated to neuroscience research. **Anne-Louise Leutenegger** has a consolidated expertise in the development of statistical methods and software for genotype-phenotype association. She has led several projects to identify the molecular bases of human traits for both rare and common disorders. She is currently co-leading the method developments of the Inserm cross-cutting program on Genomic Variability GOLD (14 Inserm teams).

HeKa (Centre de Recherche des Cordeliers) is an Inserm-Inria team whose objective is to develop methodologies and tools, and apply them in clinics towards a learning health system.

Anne-Sophie Jannot is the PI of the DROMOS project. She has long-standing experience in modelling healthcare data and more specifically data from the SNDS and a long-standing experience in developing methods integrating expert knowledge and longitudinal data.

Secondment lab: INSERM MMG, Marseille

The **Systems Biomedicine team at the Marseille Medical Genetics Institute (MMG)** develops methods to exploit large and complex biological data and apply those methods to study genetic diseases. The team is highly interdisciplinary and fosters collaborations with mathematicians and computer scientists along with biologists and clinicians. The team is associated with the CENTURI convergence and the French Bioinformatics institutes. **Anaïs Baudot** is the head of the Systems Biomedicine team. She is an expert in computational methods for biomedicine, in particular in the development of data integration approaches and network algorithms.

Expected profile

Master degree or equivalent in a quantitative field (e.g., biostatistics, applied mathematics, data science-related, bioinformatics, computational biology). An understanding of genetics/genomics would be a plus. Good background in programming (e.g., Python).

Flexibility and adaptability to work in a multidisciplinary environment on multiple projects and in collaboration. Good communication and reporting skills. Fluency in English.

Application

- Please include your CV, a cover letter and 2 references
- Use [M4DI IRP6] in the title of the application e-mail
- Contact Anne-Louise Leutenegger, anne-louise.leutenegger@inserm.fr, Anne-Sophie Jannot, annesophie.jannot@aphp.fr & Anaïs Baudot, anais.baudot@univ-amu.fr

References

- [1] J. Lambert, A. Leutenegger, A. Jannot, and A. Baudot, "Tracking clusters of patients over time enables extracting information from medico-administrative databases," *Journal of Biomedical Informatics*, vol. 139, Mar. 2023, doi: 10.1016/j.jbi.2023.104309.
- [2] J. Lambert, A.-L. Leutenegger, A. Baudot, and A.-S. Jannot, "Improving patient clustering by incorporating structured label relationships in similarity measures." *medRxiv*, p. 2023.06.06.23291031, Jun. 10, 2023. doi: 10.1101/2023.06.06.23291031.
- [3] S. Bussy, A. Guilloux, S. Gaïffas, and A.-S. Jannot, "C-mix: A high-dimensional mixture model for censored durations, with applications to genetic data," *Stat Methods Med Res*, vol. 28, no. 5, pp. 1523–1539, May 2019, doi: 10.1177/0962280218766389.