Postdoctoral Position in Computational Biology

Topic: Network-guided analysis of biobank-scale whole genome sequencing data for therapeutic research

Context: This postdoctoral position is opened at the Centre for Computational Biology (CBIO, [https://cbio.ensmp.fr/](https://cbio.ensmp.fr/)) of Mines ParisTech & Institut Curie (PSL Research Institute, Paris, France) in the context of a collaboration with Janssen R&D. The researcher will be based at CBIO under the supervision of Chloé-Agathe Azencott ([http://cazencott.info](http://cazencott.info)), and work closely with the Population Analytics team of Janssen in Spring House, Pennsylvania (USA).

CBIO is a research center dedicated to the development of methods and tools in the fields of machine learning, statistics and computer vision in order to analyze massive data generated in life sciences and medicine.

The Population Analytics team belongs to the Computational Sciences Department of Janssen R&D, which partners with scientists across therapeutic and functional areas in order to identify novel targets, predict target-associated liabilities, identify opportunities for drug repurposing, and assess patient stratification to better inform these analyses as well as patient selection for clinical trials.

Scientific context: Discovering safe and effective therapies for complex diseases requires identifying genomic regions associated with disease risk. To that effect, much effort has been devoted to collecting large data sets made of genomics data from individuals with and without disease. Analyzing these data sets is challenging statistically (orders of magnitude more variants than samples), computationally (large number of variants and, for some data sets, samples), and biologically (interpretation of statistical results).

Biological networks, in which nodes are genes or other biological entities and links are functional or physical relationships between these entities, are often used to encode established biological knowledge. Several approaches have been proposed to guide discovery in genome-wide genomics data sets, under the hypothesis that genomic regions associated with disease risk are likely to be connected on a given biological network. These approaches encourage discoveries that are consistent with established knowledge, which increases discovery power and facilitates interpretation. However, they differ in their mathematical modeling and assumptions and hence give different solutions on the same problem. In Climente-González et al. (2021), we proposed a consensus network approach, based on several existing tools, to identify networks of
genetic loci involved in familial breast cancer susceptibility. The tools are applied to genome-wide association study (GWAS) data for about 2,500 samples and 200,000 variants (after quality control).

The goal of this project will be to scale up and apply this approach to the analysis of whole-genome sequencing (WGS) data (~800M variants after quality control) generated for up to 500,000 participants of the UK Biobank (Bycroft et al., 2018).

**Role:** Your role in this project will be to
- Participate in the selection of a relevant case study phenotype.
- Participate in the choice of a relevant case study phenotype, how to map genetic variants to genes, and a relevant gene-gene interaction network to use as prior knowledge.
- Compare and benchmark tools to compute an association score for a gene from a list of gene variants mapped to this gene and the data so as to select the most appropriate for the study.
- Scale-up the network-guided GWAS tools used in Climente-González et al. (2021) to data sets containing up to 500,000 samples and hundreds of millions of variants. See [https://github.com/hclimente/gwas-tools](https://github.com/hclimente/gwas-tools) for the current implementation of those tools, applicable to datasets of a few thousand samples and a few hundreds of thousands of variants.


**Qualifications:** A PhD in computer science, computational biology, statistical genetics, or equivalent is required. Experience with genome-wide association studies, whole-genome sequencing data, or biological networks is strongly preferred.

**Salary:** ~ 2430 euros / month (gross) depending on experience.

**Location:** CBIO Mines ParisTech & Institut Curie (Paris, France)

**Duration:** 24 months.

**Start date:** As soon as possible.

**Applying:** Please email chloe-agathe.azencott@mines-paristech.fr with your CV and a motivation letter in English.