Position in Computational Biology

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

Context: This position is opened at the Centre for Computational Biology (CBIO, https://cbio.ensmp.fr/) of Mines ParisTech & Institut Curie (PSL Research Institute, Paris, France) under the supervision of Chloé-Agathe Azencott (http://cazencott.info) and Florian Massip (https://flomass.github.io/).

7 months of funding are guaranteed in the context of a collaboration with Janssen R&D, during which the researcher will work at CBIO, closely with the Population Analytics team of Janssen in Spring House, Pennsylvania (USA).

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 (WES) data (~600K variants after quality control) generated for up to 200,000 participants of the UK Biobank (Bycroft et al., 2018).

We have already made progress in that direction, identified a case study phenotype, prepared the data, and run the pipeline for four of the six tools used in Climente-González et al. (2021, 2023).

References:
Bycroft, C. et al. (2018). The UK Biobank resource with deep phenotyping and genomic data. *Nature* 562, 203–209
Climente-González, H. et al. (2021). Boosting GWAS using biological networks: A study on susceptibility to familial breast cancer. *PLoS Comput Biol* 17(3): e1008819.
Climente-González, H. et al. (2023). A network-guided protocol to discover susceptibility genes in genome-wide association studies using stability selection. *STAR Protocol* 4(1): 101998.

Role: Your role in this project will be to

- Perform a sensitivity analysis, evaluating the impact of randomly removing a fraction of the samples;
- Run a similar analysis on summary statistics from two other biobanks;
- Evaluate the impact of several choices, such as that of the tool used to summarize SNP p-values into gene p-values or that of the biological network used.