Postdoctoral position in machine learning & bioinformatics

Topic: Statistical machine learning for the integration of transposable elements variability and epivariability in genotype-to-phenotype studies

Supervision: Chloé-Agathe Azencott https://cazencott.info in collaboration with Vincent Colot https://www.ibens.bio.ens.psl.eu/spip.php?rubrique37 and Pierre Baduel https://pbaduel.com/

Location: CBIO Mines Paris & Institut Curie (Paris) https://cbio.ensmp.fr/

Funding: through ANR project STEVE (*Advancing genotype to phenotype Studies by considering Transposable Elements Variability and Epivariability*).

Duration: up to 36 months.

Start date: as soon as possible.

Qualifications: A PhD in bioinformatics, statistics, machine learning or equivalent. Prior experience with population genetics or genomics data is not necessary, but motivation for working hands-on with sequencing data is required.

Scientific context: Most recent efforts to identify genetic features linked with heritable differences in complex traits have focused on looking for associations between single nucleotide polymorphisms (SNPs) and a phenotype in genome-wide association studies (GWAS). However, there is increasing evidence that one of the shortcomings of GWAS is that they ignore transposable element (TE) sequences, which make up a large part of the genome of many species, including humans and many plants.

Indeed mobilization of TEs can generate large effect mutations and mobile TEs as well as non-mobile TE derivatives have been shown to affect the expression of genes near the site of their insertion; furthermore, this can be modulated by epigenetic mechanisms, such as DNA methylation of the TE sequence. However, until recently, the repetitive nature of TEs has made it difficult to call their variants, explaining why they remain under-used in genotype-to-phenotype studies.

Our goal is to provide knowledge and tools that will make it possible to incorporate TE variants (presence/absence, sequence variability) and epivariants (DNA methylation status) in GWAS. To this end, ANR-funded project STEVE focuses on model plant organism *Arabidopsis thaliana*, and aims at (1) comprehensively assessing and

understanding TE epivariability in *A. thaliana*; and (2) building a new methodological framework to incorporate TE variants and epivariants in genotype-to-phenotype association studies. Statistical challenges include that (1) these variants tend to have low frequency, and (2) they can vary not only in presence/absence, but also in their sequence and in methylation status. Our goal is therefore to develop tests for biologically meaningful groups of multivariate variants.

In the first phase of this project, we have gathered 720 *A. thaliana* strains for which DNA methylation data have been obtained using Illumina short-read sequencing, and identified ~1,000 reference TE sequences with DNA methylation variation as well as ~20,000 non-reference TE sequences. In addition, we have used ONT long-read sequencing to determine, for ~100 of the 720 strains, the full sequence and DNA methylation state of all the non-reference TE sequences they contain.

Your role will be to:
– extract from the pre-processed ONT and Illumina sequencing data the DNA methylation states of reference and non-reference TE sequences as well as of their adjacent regions;
– organize the data mentioned above in a coherent and easily accessible database that will enable cross-strains comparisons;
– investigate the genomic and environmental determinants of TE epivariability, using flexible multivariate models such as random forests and the associated measures of feature importance;
– develop statistical tests of association between TE-containing alleles, their epigenetic state, and a phenotype that is either the transcriptional levels of adjacent genes across the 720 strains or the ~400 quantitative traits collected for diverse subsets of strains (Arapheno database).

Applying: Please email chloe-agathe.azencott@minesparis.psl.eu with your CV and a motivation letter (in either English or French).