# PhD Position – Stable feature selection for multi-locus genome-wide association studies — ANR project SCAPHE

## Context

Differences in how patients experience disease can be explained in great part by their genomic differences. Enabling precision medicine hence requires identifying genomic features associated with disease risk, prognosis or response to treatment. This is often achieved using genome-wide association studies (GWAS), which look for associations between single nucleotide polymorphisms (SNPs) and an observable trait (phenotype). However, for many complex traits, the SNPs they uncover account for little of the known heritable variation, a phenomenon referred to as the "missing heritability" problem.

ANR project SCAPHE ("Methods for discovering SNP Combinations Associated with a PHEnotype") builds on the hypothesis that this is due to the effect of non-additive interactions between SNPs, together with a lack of robustness stemming from the relatively small sample sizes. This last issue can be alleviated by integrating biological networks to GWAS. SCAPHE proposes to develop *novel machine learning algorithms for GWAS*, integrating biological networks and modeling non-additive SNP effects, to robustly detect SNP combinations associated with a phenotype.

This PhD position will be funded as part of SCAPHE ("Methods for discovering SNP Combinations Associated with a PHEnotype") and will start on **January 2, 2019.** The project will take place under the supervision of Chloé-Agathe Azencott (http://cazencott.info).

## Research topic

Among the causes for missing heritability, the failure to account for joint effects between multiple loci (epistasis) has garnered interest in recent years. However, methods for the detection of epistasis in genome-wide data suffer heavily from the statistical difficulties posed by the broadening gap between the number of features that can be measured (easily reaching tens of millions) and that of samples for which they can be collected (more usually of the order of hundreds or thousands) [Dernoncourt et al., 2014].

One way to address this problem is to reduce the dimensionality of the space of solutions by means of structural constraints. Those can in particular be given by biological networks. Several methods have been developed to that end in recent years [Azencott et al., 2013 ; Azencott, 2016]. However, they still lack *stability*, or *robustness*, to slight changes in the input data. The goal of this PhD project will be to propose and develop methods to integrate stability/robustness to the design of multi-locus GWAS algorithms. Possible research directions include, but are not limited to, stability selection [Alexander et al., 2011 ; Shah et al., 2013] or differencial privacy [Bassily et al., 2016].

### Lab

The project will take place in the Centre for Computational Biology (CBIO — http://cbio.ensmp.fr), a joint laboratory between Mines ParisTech, one of the most prominent French engineering schools, and Institut Curie, a major hospital and research facility dedicated to cancer. CBIO benefits from an exceptional scientific environment with immediate access to experts and collaborators in biology and medicine, enabling a stimulating interdisciplinary exchange. The laboratory is located in the centre of Paris, both in Mines ParisTech and in the nearby Institut Curie.

### Prerequisites

- MSc in computer science / applied maths / engineering or equivalent;
- Solid notions of machine learning or statistics;
- Proficiency in at least one programming language;
- Motivation to work on genomics and bioinformatics applications. Prior experience in these domains is welcome but not mandatory.

### How to apply

Send your CV, a cover letter, your MSc grades and 2 contact references by email at chloe-agathe.azencott@mines-paristech.fr

### Relevant reading:

Alexander, David H., and Kenneth Lange. "Stability selection for genome-wide association." *Genetic epidemiology* 35.7 (2011): 722-728.

Azencott, Chloé-Agathe, et al. "Efficient network-guided multi-locus association mapping with graph cuts." *Bioinformatics* 29.13 (2013): i171-i179.

Azencott, Chloé-Agathe. "Network-Guided Biomarker Discovery." *Machine Learning for Health Informatics*. Springer, Cham, 2016. 319-336.

Bassily, Raef, et al. "Algorithmic stability for adaptive data analysis." *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*. ACM, 2016.

Dernoncourt, David, Blaise Hanczar, and Jean-Daniel Zucker. "Analysis of feature selection stability on high dimension and small sample data." *Computational statistics & data analysis* 71 (2014): 681-693.

Shah, Rajen D., and Richard J. Samworth. "Variable selection with error control: another look at stability selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75.1 (2013): 55-80.