Méthodes pour la découverte de combinaisons de SNPs associées avec un phénotype à partir de données génome entier

English title: *Methods for discovering SNP Combinations Associated with a PHEnotype, from genome-wide data*

Summary table: Persons involved in the project

Partner	Name	First name	Position	Involvment (PM)	Role
ARMINES (CBIO)	Azencott	Chloé-Agathe	Permanent Researcher	30	Scientific coordinator
ARMINES (CBIO)	Vert	Jean-Philippe	Professor	6	TT 1
ARMINES (CBIO)	Climente	Héctor	PhD student	12	T 1.1
ARMINES (CBIO)	Slim	Lotfi	PhD student	12	T 1.2
ARMINES (CBIO)			PhD student (to be hired)	36	WP 2
ARMINES (CBIO)			Postdoc 1 (to be hired)	24	WP 1
ARMINES (CBIO)			Postdoc 2 (to be hired)	24	WP 3

1 Context, positioning and objectives

1.1 Context

Differences in disease predisposition or response to treatment can be explained in great part by genomic differences between individuals [64]. In consequence, there is a growing interest for incorporating genomic data into *precision medicine*, and tailoring disease treatment and prevention strategies to the individual genomic characteristics of each patient [3]. Early examples of the usage of genomic information in precision medicine include the breast cancer drug trastuzumab (Herceptin), which dramatically improves the prognosis of patients whose tumor overexpresses the HER-2 gene, or the colon cancer drugs cetuximab (Erbitux) and panitumumab (Vectibix), which have little effect on patients that have a mutation in the KRAS gene. How can we further encourage such discoveries?

To be able to use genetic characteristics in precision medicine, we need to identifying genetic features associated with disease risk, diagnostic, prognosis, or response to treatment. This endeavor hence depends on collecting considerable amounts of molecular data for large numbers of individuals. It is enabled by thriving developments in genome sequencing and other high-throughput experimental technologies, thanks to which it is now possible to accumulate millions of genomic descriptors for thousands of individuals. Unfortunately, we still lack effective mathematical methods to reliably detect, from these data, which of these genomic descriptors (or *features*, or *variables*) determine a phenotype such as disease predisposition or response to treatment [41]. Today, this is a major hurdle on the path to precision medicine.

Genome-wide association studies (GWAS) are one of the most prominent tool for detecting genetic variants correlated with a phenotype. They consist in collecting, for a large cohort of individuals, the alleles they exhibit across of the order of hundreds of thousands to several millions of single nucleotide polymorphisms (SNPs), that is to say, individual locations across the genome where nucleotide variations can occur. The individuals are also phenotyped, meaning that a trait of interest (which can be binary, such as disease status, or continuous, such as age of onset) is recorded for each of them. Statistical tests are

then run to detect associations between the SNPs and the phenotype. These statistical tests can account for the correlation between genetic variants [72], leverage linear mixed models to correct for sample relatedness — which invalidates the assumption of population homogeneity underlying most tests [62], or assess the joint contribution of multiple genetic loci, either additively [83] or multiplicatively for pairwise interactions [31].

While GWAS have provided novel insights into the pathways underpinning many common human diseases [61], a number of frustrating results have also been reported [76]. Indeed, most of the genotype-to-phenotype associations they have detected are weak, too many of their findings have failed to be replicated in other studies, and the genetic variants they uncovered often fall short of explaining all of the phenotypic variation that is known to be inheritable. This last phenomenon is often referred to as the "missing heritability" problem [48].

One key explanation is that few of the established approaches for GWAS account for the *joint epistatic effects* of multiple SNPs [48, 51, 54]. Indeed, several SNPs might act together towards a phenotype, for example by targeting multiple redundant parts of a same pathway.

Moreover, GWAS are *statistically underpowered*, as the number of SNPs investigated is orders of magnitude larger than the sample size: only SNPs with a large effect size can be detected. This also results in a *robustness* issue, particularly for complex models: which SNPs are deemed associated with the phenotype can vary a lot across related datasets [20]. This suggests that current approaches often capture idiosyncrasies rather than truly relevant SNPs.

Thankfully, a wealth of additional biological knowledge can be leveraged to address these issues. This knowledge can for example take the form of functional annotations, or of other molecular measurements (such as gene transcripts abundance or DNA methylation status) on the same samples. Because genes cooperate through their interaction (physical or otherwise), in this project we will mostly focus on biological knowledge thas has been organized in *molecular networks*. Pre-existing network information can help capture the intuition that several SNPs are more likely to act together towards a phenotype if they affect genes that interact physically or along a pathway.

Such networks are often used post-hoc, to facilitate the interpretation of the results of a GWAS [34, 79]. They therefore rely on single-SNP analyses. By contrast, several *machine learning* approaches have been proposed to look for patterns of SNPs that are connected on molecular networks, hence capturing relevant information in a statistically sound fashion [4, 5] [57]. In essence, constraining the search for SNPs of interest reduces the dimensionality of the search space and therefore increases statistical power. However, these methods currently only account for the effect of each SNP either individually or additively. There is therefore a need for **novel machine learning procedures for genome-wide association studies that integrate biological networks and model non-additive epistatic effects between variants.**

1.2 Objectives and scientific hypotheses

In this project, we make the hypothesis that part of the missing heritability can be discovered by combining GWAS data with established biological knowledge, organized as networks. We surmise that this calls for novel data mining procedures, which successfully model non-linear interactions between genetic loci and compensate for the lack of statistical power by incorporating knowledge about pathways or genomic interactions as well as data collected for multiple related phenotypes.

The goal of SCAPHE is to **develop methods that enable the discovery, from data generated by high-throughput genomic technologies, of SNP combinations associated with a given phenotype.** Ultimately, this project aims at generating novel biological hypotheses based on strong statistical evidence. This will be achieved by casting GWAS as feature selection problems, and developing network-guided GWAS, in which the selected SNPs are encouraged to follow a given network structure, through three orthogonal methodological directions, detailed in the following work packages (WP):

— the development of methods for non-additive, multi-locus, network-guided GWAS (**WP 1**);

— the development of biomarker discovery algorithms explicitly designed for robustness (**WP 2**);

— the joint analysis of multiple related phenotypes (**WP 3**).

These work packages will be supported by three tranversal tasks (TT):

— controlling false discovery rate (**TT 1**);

— high-performance computing to deal with the large dimensionality of the data (**TT 2**);

— biological applications, which will guide the methodological developments we propose (**TT 3**).

These methodological developments will take the form of novel algorithms implemented in Open Source software packages.

1.3 Originality and relevance in relation to the state of the art

Collecting SNP microarray data from cohorts of cases and controls has become routine when trying to determine the genetic underpinning of a given disease. But the techniques to analyze these data are lagging behind, and most studies are still limited to searching for single-locus associations, with limited power. Yet the fields of statistics, machine learning, and data mining, which are central to the analysis of genomic data, have dramatically progressed in the last twenty-five years. *Feature selection*, which aims at identifying the most important features in a data set and discarding those that are irrelevant or redundant, is of particular interest for identifying biologically relevant features. However, too few of the recent efforts in this area have been focused on the challenges that genomic data represent: **stable, structured feature selection with few samples in high dimension**.

Indeed, there is a broad gap between the number of features (SNPs) we are able to measure for a given sample (easily reaching tens of millions with current technologies) and the number of samples we can collect (more commonly in the order of thousands). This **high-dimensional, low sample-size** situation drastically limits the power of general-purpose statistical and machine learning approaches [18]. In sharp contrast with the current "big data" vision, we cannot expect this problem to disappear with improvements in technology: the number of individuals with a given condition that we can sequence will never outgrow the millions of features that can be collected about them, in particular for rare diseases.

To alleviate this issue, genomic data can usually be paired with a wealth of prior knowledge that can be exploited to constrain the feature selection procedure. In addition to ensuring that new findings are consistent with other sorts of evidence, this increases statistical power. Such prior knowledge can in particular take the form of **explicit network structure**, such as the organization of genes or proteins in networks of interactions, regulatory relationships between genetic loci, or contact maps defining the 3D structure of the genome. SCAPHE's focus on biological networks is motivated by the assocation of network modules with diseases [8], and the success of network-based heuristics at the gene level [81].

To date, most of the statistical genetics approaches that incorporate structure are limited to pre-defining groups of variables susceptible to "work together". These groups are then evaluated by means of various statistical tests [84]. In contrast, recent endeavors of statistics and machine learning to elegantly incorporate data structure directly in the learning procedure are giving promising results [49] in various applications. The resulting ideas have started to be applied to biomarker discovery from genomic data [57] [4]. Although combining statistical tests designed by statistical geneticists with machine learning approaches can increase the statistical power of GWAS [50], it is fair to say that these methods still have not significantly advanced the field of complex genomic data analysis. Indeed, finding the relevant features is much harder than finding those that give optimal predictivity [55].

One deficiency of these approaches is that, because of the very high dimensionality considered, they have generally been limited to contemplating only additive effects between the variables, although many biological phenomena are **nonlinear**. While a variety of statistical tests have been developed to characterize so-called *epistatic* effects, most of those are limited to quadratic models involving only two SNPs at a

time [54], and there is, to the best of our knowledge, no work that combines them with existing biological networks.

Moreover, a major shortcoming of current approaches is to generally **fail to guarantee the stability of their selection**. The *stability* (or *robustness*) of feature selection procedures, meaning their ability to retain the same features upon minor perturbations of the data, remains a major predicament in the high-dimensional, low sample-size setting. Current algorithms are typically highly unstable, often yielding widely different results for different sets of samples relating to the same question [20]. In practice, ranking features one by one, based on t-test scores, often still yields the most stable selection [37]. This high variability implies that these algorithms capture idiosyncrasies rather than truly relevant features. This casts doubts on the reliability of predictive algorithms built on the selected features and impedes interpreting these features to yield novel biological insights. Recent efforts to make feature selection algorithms more stable, based for instance on multiple repetitions of the procedure on subsamples of the data [65], are yielding encouraging results. However, applications of these methods to GWAS remain rare, and they are seldom used in combination with epistasis detection or network information.

Finally, another type of tools developed by the machine learning community to alleviate shortcomings of analyses of relatively few samples in high dimension is the use of **multitask** approaches. These approaches are driven by the assumption that there are benefits to be gained from jointly learning on related tasks. A number of precision medicine settings lend themselves well to this approach. These include pharmacogenomics [66], in which the phenotypes are responses to different treatments, or eQTL studies, in which the phenotypes are gene expression levels [15]. Various methods for the detection of epistasis have been applied to eQTL studies [33], but existing multi-phenotype methods are typically limited to additive interaction effects [42]. Among those, several propose to incorporate network information. They typically use variants of the Lasso that focus more on predicting risk than detecting associated SNPs [16].

There is therefore a need for novel multitask, stable machine learning procedures that integrate network data to drive the discovery of SNP combinations associated with a phenotype.

1.4 Methodology and risk management

We propose casting GWAS as a feature selection problem, and addressing the objectives of SCAPHE by building on the regularized relevance framework. This framework is close to that of Lasso and its structured sparsity variants. However, its emphasis is on feature relevance rather than predictivity, which allows for building on a large body of work from statistical genetics. Moreover, the optimization is done directly in binary space, yielding formulations that offer better computational efficiency in very high dimension.

WP 1: Non-additive, multi-locus, network-guided GWAS Building on the regularized relevance first proposed in [5] and further developed in [4], we will develop tools for the network-guided discovery of non-additive combinations of multiple SNPs that are associated with a phenotype.

We will first use statistical tests developed for multiplicative effects between pairs of SNPs [54]. To alleviate the computational and statistical burdens of testing of the order of 10^{12} pairs (for a data set containing 10^6 SNPs), we will pair these approaches both with ideas from the domain of testable patterns [73] and selective inference [70] and with parallel implementations on graphic processing units (GPUs) [35].

We will also investigate complex machine learning models. Because SNPs are categorical variables, random forests are a natural candidate, and have been applied to GWAS data before [67, 85] but never in combination with the prior network knowledge that can make them more robust. Short conjunctions or disjunctions [21], being easily interpreted, are a very promising avenue as well.

WP 2: Robustness-driven algorithm design Robustness to slight changes in the data is a key feature of GWAS algorithms. It is therefore of prime importance to integrate this aspect in their design.

We will adopt strategies based on performing selection on multiple random subsamples of the data [65]. Such techniques have been applied to GWAS data before [1], but never when considering interactions

between SNPs nor network data. In particular, we will explore how these strategies can be explicitly implemented in the regularized relevance framework [4, 5]. In connection with randomization, we will explore the use of dropout approaches, which can be interpreted as building machine learning models favoring rare but informative features (i.e. SNPs, here) [77].

We will also increase robustness by employing strategies for weighing samples according to their suitability for the task at hand. As existing variance reduction approaches [27] to determine these weights may be too computationally demanding for genome-wide data, we will attempt to learn them as was for instance done in [71].

WP 3: Multi-phenotype approaches The assumption that there are benefits to be gained from jointly learning on related tasks has long driven the field of multitask learning. A number of precision medicine settings lend themselves well to this approach. For example, in pharmacogenomics – studying the genomic cause of response to different treatments [66] – one could perform GWAS for each treatment individually, but jointly discovering SNPs for all of them reduces the dimensions-to-sample ratio of the data and can increase statistical power. eQTL studies, in which gene expression levels are used as phenotypes, also fall withing this setting [15]. In this context, multi-phenotype approaches can also be a way to integrate gene expression data as intermediate phenotypes in a GWAS.

We will extend our algorithms to their multitask version, by enforcing that similar SNPs are selected across similar tasks [69]. When a measure of similarity between tasks is known, for instance between chemicals [6], we will make use of this information as it can improve both power and robustness [11].

Transversal tasks The three work packages will be supported by three transversal tasks:

TT 1. Controlling false discovery rate, through analyses in the spirit of [13].

TT 2. High-performance computing. To deal with the large dimensionality of the data, we will develop parallel implementations of our algorithms for multi-core architectures, computing clusters, or GPUs.

TT 3. Applications. The methodological developments proposed here will be guided by applications both on public data and on private data within the partnership between CBIO and Institut Curie.

Feasibility and risk assessment Achieving the objectives of SCAPHE requires the combination of machine learning, statistical genetics, and high-performance computing. The experience of the scientific coordinator at the interface of these domains, together with the rich research environment of CBIO make for outstanding conditions in which to successfully carry this project. CBIO is a common research center of MINES ParisTech and ARMINES that is tightly connected with Institut Curie, hence harboring close connections to experts in cancer research, machine learning, and statistics.

In addition, the three work packages are orthogonal, and substantial advances can be achieved in any of them independently, hereby mitigating risk.

We are attempting to tackle a particularly challenging case of high-dimensional, low sample size data analysis. Should our approaches fall short of solving the problems we set out to address, they could still be applied to restricted sets of variants on focused projects, and bring forth useful biological insights. The methodological advances brought forth by SCAPHE can also have applications in other domains, such as the analysis of functional brain imaging data for the identification of regions of the brain associated with a particular function, or remote sensing, where one tries to identify which sensors are really needed.
While SCAPHE falls well within the scientific mission of ARMINES and CBIO, it is unique in two aspects. First, it focuses on SNP data, which differ from other types of genetic data in that they are discrete and exhibit specific patterns of correlation. Second, its focus is on biomarker discovery, rather than on phenotype prediction, and we use and develop different and complementary methodological tools.

2 Organization and resources

2.1 Scientific coordinator and team

Chloé-Agathe Azencott is a junior faculty member at MINES ParisTech. She is an internationally recognized expert on machine learning for genetics and drug discovery applications. References to her publications are underlined in the current document. She joined CBIO in 2013 and has been holding a permanent position since December 2016. She will lead SCAPHE and devote 80% of her time to this project.

The scientific team will be composed of 6 additional members of CBIO:

- 3 PhD students:
  - H. Climente will devote 12 months to T 1.1 and T 1.2;
  - L. Slim will devote 12 months to T 1.3;
  - a PhD student to be hired (PhDS), to work full time (36 months) on WP 2;
- 2 postdoctoral fellows to be hired (PD1 on this grant, and PD2 on other funding), for 24 months each, to work on T 1.4 and WP 3 respectively;
- and 1 senior researcher, J.-P. Vert, involved for a total of 6 months, on TT 1.

ARMINES is a private non-profit research and technological organisation (RTO) funded in 1967, having common research centres with the Ecoles des Mines: Paris (Mines ParisTech), Albi-Carmaux (Mines Albi-Carmaux), Alès (Mines Alès), Douai (Mines Douai), Nantes (Mines Nantes) and Saint-Etienne (Mines Saint-Etienne), gathering public and private personnel and means, to collaborate on an arms lengths basis and perform research contractual activities and academic research training. The Joint Research Units and the collaboration between ARMINES and the Ecole des Mines are organised within the frame of a convention signed with each Ecole des Mines in conformity with the Law dated April 18th, 2006, under the administrative authority of the French Minister of Industry; ARMINES having in addition, the duty to manage research contractual activities and the related intellectual property rights. ARMINES currently shares 48 Joint Research Units (Common Research Centres) with the Ecoles des Mines. With a total turnover of more than EUR 44,72 million (2015), ARMINES is amongst the top of private contract research institutions affiliated to higher education entities. ARMINES is also member of EARTO (European Association of RTOs), EIRMA (European Industrial Research Association Management) and of the Carnot Institute.

CBIO (Centre for Bioinformatics) is a common research center between ARMINES and Mines ParisTech. It is linked with Institut Curie's "Genome and cancer: Bioinformatics, Biostatistics, Epidemiology and Computational Systems" (U900) team, through a partnernship between MINES ParisTech, ARMINES and INSERM.

2.2 Means of achieving the objective

Proposed machine learning framework

We propose addressing the objectives of SCAPHE by building on the *regularized relevance* framework first proposed in [5] and further developed in [4]. Regularized relevance is a machine learning framework in which one performs feature selection by indentifying the subset of features (SNPs) that maximizes the sum of a data-driven *relevance function* and a domain-driven *regularizer*.

The relevance function quantifies the importance of a set of features (SNPs) with respect to the task under study. It can be derived from a statistical test of association between groups of SNPs and a phenotype.

The role of the regularizer is to encourage the selected features to be compatible with a priori constraints on the feature space, that is, prior knowledge about the SNPs. A simple example of regularizer, built on prior belief that a limited number of SNPs should be involved, is the cardinality of the selected set. Another regularizer, which uses the Laplacian of a predefined network over the SNPs, encourages the selected SNPs to be connected on this network. SNP networks can be constructed from gene networks by connecting SNPs that are near the same gene, or near two interacting genes. As those become available,

we will also be able to consider networks encoding trans-eQTL interactions or physical 3D contacts. The specific choice of gene network is application-dependent. Enforcing both sparsity and connectivity simultaneously can be done by combining these two regularizers linearly.

If $\mathcal{V}$ is the set of all p SNPs, W the $p \times p$ adjacency matrix of the network, and $R : \mathcal{V} \to \mathbb{R}$ the relevance function, we hence want to find the set $\mathcal{S}$ of SNPs that is the solution of:

$$\underset{\mathcal{S}\subseteq\mathcal{V}}{\arg\max} \quad \underbrace{R(\mathcal{S})}_{\text{association}} - \lambda \left(\underbrace{|\mathcal{S}|}_{\text{sparsity}} + \alpha \underbrace{\sum_{p\in\mathcal{S}} \sum_{q\notin\mathcal{S}} W_{pq}}_{\text{connectivity}} \right).$$

Here $\lambda \in \mathbb{R}^+$ is a parameter which controls the balance between the relevance and the regularization terms. α controls the respective importance of the sparsity and the connectivity constraints.

This problem can be reformulated as a minimum cut problem, which can be efficiently solved thanks to maximum flow algorithms [26]. The improvements we obtained over state-of-the-art methods with this formulation [5] encourage the choice of the regularized relevance framework for SCAPHE.

This formulation is close to that of Lasso [74] and its structured sparsity variants [32, 49]. However, Lasso-like approaches focus on the minimization of a prediction error, while the regularized relevance shifts the emphasis to the maximization of feature importance with respect to the question under study. By absolving itself from a specific predictive algorithm, this framework is more appropriate for the identification of stable, biologically interpretable sets of SNPs. It also present the advantage of building on a large body of work from statistical genetics, by naturally integrating score tests that account for population structure [82], linkage disequilibrium between SNPs [43], or confounding factors [36]. Moreover, in this framework, optimization is done directly over binary states (selected/non selected) rather than over real-valued coefficients. This presents the conceptual advantage of yielding formulations that can be optimized without resorting to convex relaxation, and offers better computational efficiency in very high dimension.

Neither the regularized relevance framework nor structured sparsity variants of the Lasso currently make it possible to model non-additive effects between SNPs. They are typically less robust than simple statistical tests [28], selecting sets of SNPs that have higher predictive value but are less easily interpreted. Finally, while both these frameworks have multitask versions, alleviating the burden due to the relatively small number of samples by analysing multiple related phenotypes simultaneously, neither exploits the relationship between these phenotypes. We address these three independent aspects through three orthogonal workpackages.

WP 1: Non-additive, multi-locus, network-guided GWAS

This work package aims at proposing tools for the network-guided detection of non-additive effects between SNPs in GWAS data. For this purpose, we will propose variants of the regularized relevance framework, in which the association term will be able to account for, first, pairwise interactions between SNPs (**Task 1.1**) and, later on, more complex models of interaction (**Task 1.4**). To prepare for this later step, we will devote **Task 1.3** to building ways of scoring sets of SNPs within these complex models. To address the increased computational and statistical burdens of non-additive multi-locus models, **Task 1.2** will focus on developing admissible heuristics to avoid investigating all pairs of SNPs in the algorithm developed in T 1.1.

Task 1.1: Network-guided detection of pairwise epistasis *Team members: HC, CAA.*

Objectives The objective of this task will be to develop a combinatorial optimization formulation of network-guided GWAS that accounts for multiplicative effects between pairs of SNPs.

Work programme We will use a relevance function that accounts for pairwise epistatic effects. We will build this term from statistical tests developed for detecting multiplicative effects between pairs of SNPs [54, 80].

Feasibility and risk assessment The SKAT test score with a quadratic kernel [83] can be used to create a relevance function that accounts for pairwise interactions. The resulting optimization problem that can still be rewritten as a graph cut minimization, and hence solved efficiently. Nevertheless, because of the number of pairs of SNPs contemplated, the computational resources required might still be too intensive for genome-wide data runs. Statistical power might also be too low. T 1.2 will address these issues.

Task 1.2: Accelerated network-guided detection of pairwise epistasis *Team members: HC, CAA.*

Objectives Searching exhaustively for pairwise epistasis in data containing a million SNPs requires testing of the order of 10^{12} pairs of SNPs. We hence expect the algorithms developed in T 1.1 to be computationally intensive and to suffer from lack of power. The objective of this task is to reduce this burden.

Work programme We will by develop admissible heuristics to reduce the number of statistical tests to perform. For this purpose, we will combine our network-guided algorithms with previous work specific to epistasis detection [86] or from the domain of testable patterns [44, 73, 52]. Screening rules, which discard irrelevant variables from the optimization in Lasso problems early on [23], can also serve as a starting point.

Feasibility and risk assessment Pushed by increasing needs for large-scale machine learning methods, the fields of testable patterns and screening rules have considerably grown in recent years. There is little doubt that they can provide tools to help cut down on the number of computations to perform.

Task 1.3: Detection of complex epistatic patterns *Team members: LS, JPV, CAA.*

Objectives There is no reason to believe epistasis to be limited to the interaction of only two SNPs. However, we will not be able to test even triplets of SNPs on most data sets, both for computational reasons and because the effects would have to be very strong to be detected on such data. The objective of this task is to develop machine learning models to detect complex patterns of interaction.

Work programme Because SNPs are categorical variables, random forests are a natural candidate, and have been applied to GWAS data before [67, 85, 46]. Indeed, they are able to evaluate the importance of a SNP towards a phenotype while accounting for the effects of the rest of the genotype [2]. However, this importance score does not reveal the interaction pattern of a highly scored SNP. We will modify this score, which is built by examining in how many trees of the forest the SNP is selected, to account for the other SNPs that appear in those trees. Alternatively, we will study whether set covering machines, which gave promising results in other bioinformatics applications [21], can be applied to GWAS.

Feasibility and risk assessment This task is rather exploratory, as there is currently no method that we know of to compute the type of importance score we wish for. However, we will be able to rely on existing literature on feature selection in random forests [24, 45], as well as encouraging preliminary results by LS, to guide our work. In addition, the other tasks of SCAPHE can be successfully accomplished even if this one should fail.

Task 1.4: Network-guided detection of complex epistatic patterns *Team members: PD1, CAA.*

Objectives The goal of this task is to integrate network constraints to algorithms for the detection of complex epistatic patterns.

Work programme We will use relevance functions that stem from complex machine learning models, such as the random forest models developed in T 1.3. Another interesting starting point is covering sets machines, which detect short conjunctions or disjunctions of features that explain an outcome [21], leading to easily interpretable models. These have yet to be applied to large-scale GWAS data and combined with network constraints.

Feasibility and risk assessment It is as of yet unclear whether the integration of relevance scores coming from complex machine learning models to the regularized relevance framework will yield formulations that can still be efficiently solved, for instance by graph cut algorithms. Should the new formulations

prove computationally intractable, we will consider adapting other frameworks to our needs; for example, [58] integrates biological pathways to random forest predictions, and could be a starting point for the integration of biological networks and a focus on feature selection.

WP 2: Robustness-driven algorithm design

GWAS and biomarker discovery face a major issue of robustness: which SNPs are deemed associated with the phenotype can vary a lot across related data sets, even across overlapping subsets of the same data set [20]. Perhaps surprisingly, this question has only recently started to come under investigation [30] and this aspect is often overlooked when proposing novel approaches [56]. Most of the work in that domain has tried to group features together in meta-features, based either the data at hand or on prior knowledge [47], to yield lower-dimensional representations. Unfortunately, these groupings, if done wrongly, can confuse the feature selection procedure even more.

A more promising approach, called *stability selection*, has theoretically been proven to increase robustness in various settings [65]. Rooted in the field of ensemble learning, stability selection is based on combining the results of a large number of runs of a feature selection procedure on bootstrap samples of the data. Conceptually, it can therefore be used on top of any of our algorithms. Stability selection has been applied to GWAS data before [1], but never when considering a network structure to guide the feature selection procedure. In building bootstrap samples on which to run the feature selection procedure, both individuals and SNPs are sampled uniformely at random. In consequence, SNPs will appear in most subsamples without their neighbors, thus breaking the network structure and hindering the application of our network-guided GWAS algorithms.

In this work package, we will assess whether respecting the network structure when creating the aforementioned bootstrap subsamples increases robustness (**T 2.1**), and whether there are benefits to be gained from integrating the idea of stability selection directly in the regularized relevance formulation (**T 2.2**). In addition, we will explore an independent direction for the robustness of our algorithms, based on weighing samples according to their suitability for the procedure, in the spirit of [60] (**T 2.3**).

Task 2.1: Structured stability selection *Team members: PhDS, CAA.*

Objectives The goal of this task is to evaluate whether there are benefits, in terms of robustness, to be gained from respecting the given network structure when creating the subsamples of the data used within the stability selection framework.

Work programme We will implement a variant of the stability selection procedure [65] where SNPs are sampled so as to appear together with most of their neighbors in most subsamples. We will take inspiration from large graph sampling strategies [40], although those might have aims that differs from ours. Because the stability selection procedure was designed for Lasso algorithms [65], we will first focus on network-guided Lasso problems to assess the benefits of our proposed algorithm. Only then will we combine regularized relevance algorithms, starting with the original one from [5] and moving on to those developed in WP 1, with this new stability selection procedures and assess their robustness.

Feasibility and risk assessment Stability selection relies on large numbers of repeatitions of the feature selection procedure, hence increasing their computational burden. Nevertheless, these are trivially parallelizable, and this should not be a major hindrance. A more critical issue is that sampling features not at random might be at odds with key principles of ensemble learning, which draws its strength from combining weak learners. By using the network to sample features, we might build features selectors that are not weak enough from the "widsom of crowd" effect. The main goal of this task will be to assess whether this is the case, and to report on our conclusions.

Task 2.2: Robust regularized relevance *Team members: PhDS, CAA.*

Objectives The goal of this task is to integrate directly the concept of stability selection to the regularized relevance framework, and to assess whether the resulting algorithm improves the stability of our network-guided GWAS algorithms.

Work programme Instead of a single relevance function, we wil compute one relevance function per bootstrapped sample of the data. We will then use a regularizer that enforces small symmetric differences between the features selected for each of these samples. Solving the resulting problem can be done using a maximum flow solver on a meta-network in which each feature is duplicated as many times as bootstrap iterations, in a formulation similar to [69]. To be able to solve it in practice, we will explore ways to reduce the size of this network, and will call on the high-performance strategies from TT. 2.

Feasibility and risk assessment The main risk we incur is that the computational complexity of the approach we propose is too high. We will start by evaluating our algorithms on small sets of data, to assess whether there is a benefit large enough to warrant working on parallelization and other high-performance computing strategies to make them scalable.

Task 2.3: Sample weighting for robustness *Team members: PhDS, CAA.*

Objectives Stability can also be enforced through instance weighting schemes [60]. Intuitively, these weight samples according to their suitability to feature evaluation. The objective of this task is to explore how these ideas can be incorporated in the regularized relevance framework.

Work programme Variable-reduction approaches [27] can be employed to determine the weights. However, they might be too computationally demanding for very high-dimensional data. We will address this by learning the weights in a fashion similar to the one developed in [71].

Feasibility and risk assessment The goal of this task is to evaluate alternatives to stability selection to enforce the robustness of our GWAS algorithms. In this task, we chose to focus on sample weighting schemes as they seem the most straightforward to integrate to the regularized regression framework. If our experiments were to show that such approaches are not suitable, there is a variety of other techniques that we could explore. Those include dropout approaches, which can be interpreted as building machine learning models favoring rare but informative features (i.e. SNPs, here) [77]. Modeling the feature covariance is also kwown to improve the robustness of feature selection algorithms [59].

WP 3: Multi-phenotype approaches

The assumption that there are benefits to be gained from jointly learning on related tasks has long driven the field of multitask learning. A number of precision medicine settings lend themselves well to this approach, starting with the study of similar phenotypes in different populations. Another example is pharmacogenomics – studying the genomic cause of response to different treatments [66] – in which one could perform GWAS for each treatment individually. Searching for SNPs for all treatments jointly reduces the dimensions-to-sample ratio of the data and can increase statistical power.

We have developed a multitask version of the regularized relevance framework to enforce that similar SNPs are selected across similar tasks [69]. It is often possible to define similarities between the phenotypes; for example, in the case of a pharmacogenomics screen, the similarity between phenotypes can be directly derived from the similarity between the drugs that were used [6]. In cancer applications, integrative analyses have shown interesting similarities between cancer types [78]. The formulation in [69], however, does not make use of the similarity bewteen phenotypes. Incorporating phenotype similarity to the multi-phenotype formulation, in essence to enforce that phenotypes that are more similar have more underlying SNPs in common, can improve both the power and the robustness of multitask approaches [11].

The goal of this work package is to propose and evaluate new tools for multi-phenotype network-guided GWAS that make use of a notion of similarity between the phenotypes. We will start from existing additive formulations (**T. 3.1**) before extending those to non-additive formulations (**T. 3.2**).

A multi-phenotype setting of particular interest is that of eQTL studies, in which gene expression levels are used as phenotypes [15]. Indeed, gene expression has emerged as a bridge for explaining the relation between genotype and higher level phenotypes such as diseases, and can be empowered by the addition of interaction networks [63]. However, relatively little work considers epistatic effects in genome-wide eQTL studies [33], or addresses eQTL studies as a multi-phenotype problem. Bayesian frameworks for

multi-trait association studies [42] are a notable exception. They do not, however, make use of interaction networks. We will therefore focus **T. 3.3** on the specific case of eQTL studies.

Task 3.1: Multi-phenotype network-guided GWAS *Team members: PD2, CAA. Objectives* The goal of this task is to propose and evaluate a novel formulation of regularized relevance that incorporates a measure of similarity, or covariance, between the phenotypes.

Work programme The objective function of [69] can easily be modified to include a term that accounts for the covariance between tasks, in a formulation similar to that of [87]. The optimisation problem remains equivalent to a minimum cut and can still be solved with a maximum flow algorithm.

Feasibility and risk assessment Our previous experience [11, 22] shows that using task similarity can benefit multitask approaches, especially within sets of closely related tasks [10]. Nevertheless, integrating phenotype covariance might not bring much benefit compared with the initial formulation. In this case, we would proceed to using this initial formulation in T 3.2.

Task 3.2: Multi-phenotype network-guided detection of epistasis *Team members: PD2, CAA.*

Objectives The goal of this task is to incorporate epistatic effects between SNPs to a multi-phenotype formulation of regularized relevance.

Work programme We will start by using the relevance scores defined in T 1.1 and T 1.2 for the pairwise detection of epistasis within the multi-task formulation of regularized relevance. If incorporating a phenotype covariance matrix was proven successful in T 3.1, we will also seek to integrate these relevance score to this new formulation. Finally, we will also envision building on these models to extend the tools from T 1.4 to the multi-phenotype setting, but this is unlikely to happen during SCAPHE because of the timeline.

Feasibility and risk assessment Conceptually, combining pairwise epistatic detection approaches with the existing multitask formulations of regularized relevance (whether [69] or developed in T 3.1) should be straightforward once T 1.1 has been completed. If needed, we would devote more time and resources to T 1.1 rather than to T 3.2. Using multiple phenotypes should increase statistical power, and the work done in T 1.2, combined with the high-performance computing efforts of TT 2, should help alleviate computational burdens sufficiently for the successful completion of this task.

Task 3.3: Network-guided eQTLs *Team members: PD2, CAA.*
Objectives The goal of this task will be to develop methods for eQTL epistasis that make use of known relationships between the genes which expression we are trying to explain.

Work programme We will apply the methods developed in previous tasks specifically to the case of eQTL studies. Here phenotype covariance matrices can be for example defined based on previously published gene expression data, to try and avoid capturing non-genetic factors. One of the specificities of this setting is that both gene expressions and SNP data can be used to explain a higher-level phenotype, such as disease status. A number of tools, such as Sherlock [29], make use of that idea to identify disease-associated genes.They do not, however, allow for interactions between the SNPs, nor account for non-regulatory influences of a SNP on a gene product (such as functionally damaging alteration of the protein product); moreover, they focus on one gene at a time. We will endeavor to build more general models of genotype-expression-phenotype association, with the help of the tools developed in previous tasks.

Feasibility and risk assessment We are likely to encounter statistical issues due to the large number of tests (all SNPs against all genes) that these studies entail. One way to reduce the number of tests in eQTL studies is filtering loci based on prior knowledge. Rather than testing all possible pairs (or larger groups) of loci for interactions, only loci that are predetermined to be most relevant/promising based on biological knowledge are selected. For example, [39] focused on detecting epistasis between variants in coding regions and those in cis-regulatory regions. Developments achieved on this task will most likely be only preliminary, as eQTL studies challenges and opportunities reach far beyond the scope of SCAPHE.

2.2.1 Transversal tasks

TT 1: Controlling false discovery rate *Team members: JPV, HC, LS, PhDS, PD1, PD2, CAA.*

Objectives The ability to guarantee an upper bound on false discovery rate (FDR), that is, the expected fraction of irrelevant SNPs among all SNPs selected by our algorithms, is very important for clinical applicability. Quantifying, for each selected SNP, the statistical significance of its being selected, is far from straightforward: the feature selection procedure itself changes the distribution of the usual test statistics under the null hypothesis. Correlations between SNPs, due to linkage disequilibrium, and the use of prior knowledge further complicate the matter. This transversal tasks aims at ensuring that this aspect is addressed across all the above work packages.

Work programme A solution has been proposed in the case of the Lasso [25], but not for the type of network-guided algorithms we will develop in SCAPHE. We will systematically conduct empirical evaluations of the FDR of our methods. We will also explore how recent developments such as knock-off filters [13], multi-layers p-filters [7] and selective inference algorithms for high order interaction features [70] can be applied to our algorithms.

Feasibility and risk assessment Empirical evaluations based on various underlying genetic models is conceptually straightforward. The choice of these underlying models can greatly influence the results, but these evaluations will nevertheless give important insights. Developing theoretical control of false discovering rate is more challenging, but the recent emergence of multiple approaches, from knock-off filters to selective inference, gives us a diversity of potential research directions to achieve this goal. The expertise of J.P. Vert will be a great asset towards ensuring the success of this task.

TT 2: High-performance computing *Team members: HC, LS, PhDS, PD1, PD2, CAA.*

Objectives Addressing problems involving tens of millions of features, over multiple (potentially thousands, in the case of eQTLs) data sets, while incorporating various forms of prior knowledge and taking into account non-additive effects will require to focus explicitly on efficient and scalable implementations. The goal of this task is to develop these implementations.

Work programme We will hence develop parallel implementations of our algorithms for multi-core architectures, computing clusters, or graphics processing units (GPUs). We will first use parallel maxflow implementations [17] to speed up solving the problems we will have reformulated as minimum cut problems. GPUs will also allow for the efficient computation of non-modular relevance functions, as for example in the case of quadratic relevances [35]. Because the proposed approaches are typically linear in the number of samples, we do not expect large sample sizes to become a major computational issue.

Feasibility and risk assessment GPU implementations are particularly useful when the task can be broken down in many very simple tasks that can each be executed without requiring much access to memory, meaning that they do not necessarily improve on total runtime. We will make sure to first assess which GPU implementation of the minimum cut algorithm is the most suited to the type of networks we are building, given the global properties of such networks. In any event, several parts of our algorithms, such as the computation of the relevance scores, can be trivially parallelized on either CPU or GPU.

TT 3: Applications *Team members: HC, LS, PhDS, PD1, PD2, CAA.*

Objectives The methodological developments we are proposing in SCAPHE aim at enabling the discovery, from data generated by high-throughput genomic technologies, of SNP combinations associated with a given phenotype. Hence all of our work will be guided and driven by specific GWAS applications, with a focus on phenotypes related to precision medicine.

Work programme NCBI's Database of Genotypes and Phenotypes (dbGap) [75] will provide us with a wealth of GWAS data for various complex phenotypes, including breast cancer applications. Some of the cohorts made available also include gene expression data we can use to conduct eQTL studies in T 3.3. For toxicogenomics applications in WP 3, we can additionally make use of data from the Cancer

Cell Line Encyclopedia (CCLE) [9]. The UK Biobank [68] will also provide access to GWAS data relevant for precision medicine purposes. Because larger sample sizes alleviate statistical power issues, we will initially consider phenotypes such as age at menarche (272 995 samples), which is possibly related to breast cancer, or blood pressure ($\sim$140 000 samples), for which traditional GWAS have already identified 107 SNPs [81]. Previous research [53] hints at the presence of epistatic effects for this phenotype, which motivates applying our methods to this phenotype of importance for human health.
In addition and in parallel, we will collaborate with biologists and clinicians of Institut Curie on focused projects with specific data. The principal investigator has already initiated collaborations with N. Andrieu (PhD) on breast cancer, F. Reyal (MD, PhD) on pharmacogenomics of triple negative breast cancer, and O. Delattre (MD, PhD) on the onset of Ewing's sarcoma.
We will use biological networks derived from public databases such as BioGRID [14]. For cancer-related phenotypes, we will also use cancer-specific networks developed within Institut Curie, such as the ACSN maps [38] (also publicly available).

Feasibility and risk assessment The major risks here are either that the studied phenotypes are not subject to epistatsis, or that our approaches still do not alleviate the statistical burden sufficiently on the data at hand for this task to lead to novel biological insights. Close collaborations with experts in epidemiology, molecular biology and clinics at Institut Curie will allow us, when necessary, to define relevant intermediate phenotypes (such as age at menopause or menarche for breast cancer) to study, individually or jointly with cancer endpoints, or restrict the list of loci to investigate. The diversity of phenotypes we are planning to investigate, from clinical measurements to drug responses, gene expressions and various cancers, also increases the chance that SCAPHE directly results in biologically interesting discoveries.

2.2.2 Deliverables

Each of the tasks detailed in the three workpackages above will result in open source code for the network-guided detection of combination of SNPs associated with a phenotype in GWAS data. When appropriate, code will be released as an update to our contributed Bioconductor package [19]. When possible, parallel CPU and GPU implementations will also be distributed, in accordance with TT 2. Each task will also result in one or more publication(s), to be published in an Open Access venue, presenting the tool, theoreatical guarantees when available and power analysis (see TT 1), and behavior on simulated data, as well as results obtained on real data (see TT 3).

2.3 Resources

2.3.1 Budget

Personnel costs Personnel costs are the major item of the proposal. The Principal Investigator will devote 80% of her time to this project. She will oversee the scientific direction and management of the project, and contribute to all its tasks and aspects. This grant will not fund her salary.

H. Climente, L. Slim and one of the postdoctoral fellows to be hired will not be funded by this grant. We are therefore budgeting the following personnel expenses:

- One **PhD student** (**36PM** for a total of EUR 112 080), to be hired, will work on **WP 2** (robustness-driven algorithm design).
- One **Postdoctoral fellow** (**24PM** for a total of EUR 112 927), to be hired, will focus on **T 1.4** (network-guided detection of complex epistatic patterns).
- Finally, a **senior researcher** (J.-P. Vert) (**6PM** for a total of EUR 55 456) will be involved in **TT 1** (quantifying power gains).

Equipment The central purchase of SCAPHE will be a graphics processing unit (GPU) server for scientific computing, equiped with two Tesla K40M GPUs (EUR 10 000). Leveraging the computing power of GPUs is an integral part of the realization of this proposal and a core component of TT 2. Working on scientific computing GPUs will allow us not only to develop fast algorithms to be run on such a server, but also to

quickly develop single-GPU implementations to be run on consumer-level graphics cards. An additional EUR 10 000 have been budgeted for the purchase of new compute nodes to upgrade the CPU computing cluster maintained by CBIO at ARMINES. We will use this cluster to develop and run parallel CPU implementations of our algorithms. Hence the tools we develop will be usable on a range of equipments, from desktop computers to advanced computing environments including both CPU clusters and GPU nodes.

Other direct costs Additional expenses, budgeted for a total of EUR 10 000, will include:

- access to the UK Biobank data (`http://www.ukbiobank.ac.uk/`), an invaluable source of GWAS data;
- missions for dissemination actions, particularly for the attendance of conferences where member of the SCAPHE team will present our work;
- publication by the PI and her team of scientific articles in Open Access peer-reviewed journals.

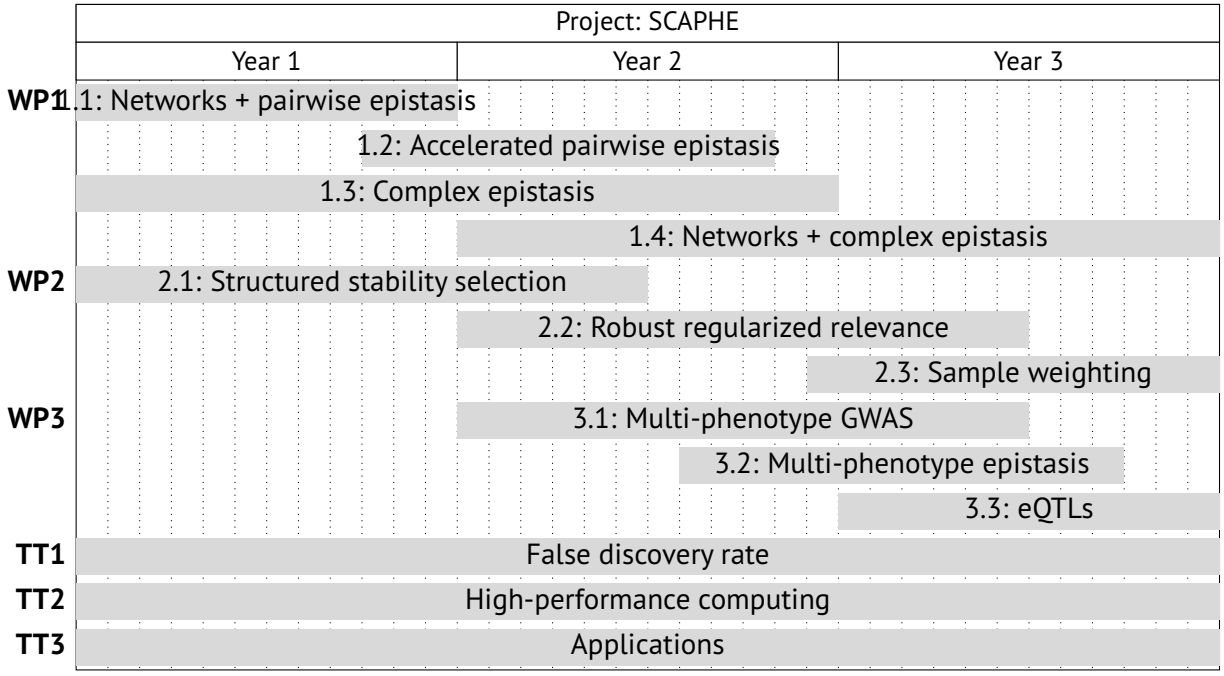
Funding received The Principal Investigator has previously obtained funding to recruit two PhD students, H. Climente and L. Slim, who will be part of the SCAPHE team. H. Climente is funded by the IC-3i international PhD Program of Institut Curie (`training.curie.fr/ic3iphd`), which is co-funded by the European Union's Horizon 2020 research and innovation programme under the Marie Sk lodowska-Curie grant agreement No 666003. L. Slim is funded by a CIFRE PhD fellowship (private sector scholarship, subsidised by ANRT, the French national association for research and technology) in partnership with Sanofi-Aventis R&D.

Costs We are asking for a total of 251 kEUR over 36 months. The following table details the costs associated with SCAPHE.

Cost Category			**Total in Euro**
Direct Costs	**Personnel**	Principal Investigator, otherwise funded (30PM)	(180 579)
		Senior Staff (6PM)	55 456
		Postdoctoral Researcher (24PM)	112 080
		Postdoctoral Researcher, otherwise funded (24PM)	(112 080)
		PhD Student (36PM)	112 927
		PhD Students, otherwise funded (24PM)	(75 285)
	Total Direct Costs for Personnel [A]		(648 407)
	not otherwise funded [B]		**280 463**
	Travel		4 000
	Equipment		20 000
	Others	Publications (including Open Access fees)	4 000
		Other	2 000
	Total Other Direct Costs [C]		**30 000**
Total Direct Costs [D=A+C]			(678 407)
not otherwise funded [E=B+C]			**310 463**
Indirect Costs (Personnel) [F=68%B]			**190 715**
Indirect Costs (Others) [G=7%C]			**2 100**
Total Estimated Costs [D+F+G]			**871 222**
not otherwise funded [H=E+F+G]			**503 278**
Total Requested ANR Contribution [50%H]			**251 639**

2.3.2 Organization of the work

The following Gantt chart describes how the work will be organized.



2.3.3 Ethical concerns

Throughout TT 3, methods used and developed in WP1–WP3 will be applied to the analysis of data measured for precision medicine purposes, which motivates this project. Therefore, the execution of this transversal task involves the processing of genomic data, such as DNA sequences, SNP data, RNA measurements, gene expression measurements, and clinical and demographic data, in compliance with fundamental ethical principles. We will be working on both public data, available from web repositories, according to the terms of use of these repositories, and data previously collected by Institut Curie, that has not been made publicly available. Data access and processing will be realized in accordance with the EU and national legistlation and facilitated by the preexisting partnership agreement between Institut Curie and ARMINES on bioinformatics and systems biology of cancer.

Data that are not publicly available will be accessed by members of the SCAPHE team only, and access will be protected. They will not be made available to third parties. Members of SCAPHE will not attempt to re-identify anonymized data. We remain at the disposal of the ANR for a full Ethics review.

3 Impact and benefits

Response rates of patients to a major drug for their condition are as low as 75% in arthritis, 11% in depression and 5% in asthma [64]. In addition, adverse drug reactions cause an estimated 197 000 deaths every year in Europe [12]. Precision medicine, or the ability to deliver preventive or therapeutic care only to those who do benefit from it, hence holds incomparable social and economic benefits. Achieving it requires discovering the genomic markers that explain the differences between patients, and SCAPHE endeavours to develop robust algorithms for this exact purpose.

Improvements in algorithms designed to reliably extract informative patterns from complex genomics data are imperative to advance therapeutic target discovery and realize the promise of precision medicine. Currently, this is one of the major challenges faced by medical genetics. SCAPHE will contribute to these advances by proposing and applying new tools to identify the SNPs involved in specific disease susceptibilities, prognoses, or responses to treatment.

Down the line, the new tools we will propose will benefit human geneticists and clinicians by providing novel precision medicine insights, potentially resulting in new diagnostic tools or therapeutic targets.

SCAPHE falls well within the scope of Transversal Axis 1 (B.11-Axe 1) defined by ANR, which concerns the elaboration of novel statistical and computational methods for the analysis of omics data towards personalized medicine.

In addition, SCAPHE fits well within the general research topic of CBIO – the development of machine learning tools for bioinformatics. At the same time, the topic of multi-locus GWAS, as well as the proposed framework, are unique within this research center. SCAPHE would therefore promote the development of this topic, and of the scientific coordinator's own team.

Interest for such approaches is not limited to the academic sphere, and our existing parnerships with companies illustrate the potential of the research proposed in SCAPHE for industrial applications. These partnerships will facilitate the rapid uptake of our methods by end-users in epidemiology and the pharmaceutical industry. More specifically, T 1.1 and T 1.2 are a part of the PhD project of H. Climente, which was proposed in collaboration with the SME Pharmatics Limited (Edimburgh, UK), which develops intelligent data analysis tools for applications in multiple areas of precision medicine and health analytics. T 1.3 is part of the PhD project of L. Slim, which is funded for its main part by Sanofi-Aventis R&D, a major multinational pharmaceutical company. All three of these tasks are excluded from the costs of SCAPHE.

The goal of SCAPHE is to develop new machine learning methods for feature selection in high-dimensional data, making use of prior knowledge about the features, available as networks, to alleviate the statistical issues linked with small sample sizes. These issues are not limited to SNP data and indeed span the breadth of the whole of data-driven biology. For example, state-of-the-art approaches on gene expression data yield disconcertingly disparate molecular signatures for the same phenotype [20]. These methodological tools can easily be extended to the analysis of other types of molecular data, such as gene or protein expression levels or methylation markers. Moreover, feature selection methods for high-dimensional data, far from being limited to applications in genomic studies, are of broad interest in a variety of domains ranging from functional brain imaging to quantitative finance and climate science. The work we will conduct on the regularized relevance framework also has the potential to instigate new theoretical and methodological developments in statistics and machine learning.

3.1 Dissemination actions

Our results will be published in Open Access peer-reviewed publications and part of the budget will be devoted to dissemination actions, such as travel to conferences to present our work.

We will put a strong emphasis on developing the code for this project following the Open Source paradigm. All our code, written in Python, R, or C/C++ as well as CUDA or OpenCL for GPU implementations, will be made available through Open Source repositories hosted on GitHub (http://github.com). R packages we develop will also be contributed to Bioconductor, and will hence have to conform to the package guidelines on quality of code and documentation. GPU implementations will be developed in CUDA for efficiency on scientific computing GPUs from NVIDIA, but will be, as much as possible, ported to OpenCL to ensure compatibility with a broader range of graphics cards.

To facilitate usage, we will make available tutorials based on code notebooks such as Jupyter notebooks (http://jupyter.org), and we will insist on user-friendly interfaces and thoroughly document our code.

References

[1] D. H. Alexander, and K. Lange. Stability selection for genome-wide association. *Genetic Epidemiology*, 35(7):722–728, 2011.

[2] A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

[3] S. J. Aronson, and H. L. Rehm. Building the foundation for genomics in precision medicine. *Nature*, 526(7573):336–342, 2015.

[4] C.-A. Azencott. Network-guided biomarker discovery. *Machine Learning for Health Informatics* Lecture Notes in Computer Science 9605. Springer International Publishing, 2016.

[5] C.-A. Azencott, D. Grimm, M. Sugiyama, Y. Kawahara, and K. M. Borgwardt. Efficient network-guided multi-locus association mapping with graph cuts. *Bioinformatics*, 29(13):i171–i179, 2013.

[6] C.-A. Azencott, A. Ksikes, et al. One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical, and biological properties. *Journal of Chemical Information and Modelling*, 47(3):965–974, 2007.

[7] R. F. Barber, and A. Ramdas. The p-filter: multilayer false discovery rate control for grouped hypotheses. *J. R. Stat. Soc. B*, 79(4):1247–1268, 2017.

[8] F. Barrenäs, S. Chavali, et al. Highly interconnected genes in disease-specific networks are enriched for disease-associated polymorphisms. *Genome Biology*, 13:R46, 2012.

[9] J. Barretina, G. Caponigro, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012.

[10] V. Bellón. Personalized adverse effects prediction. PhD thesis, PSL University, 2017.

[11] V. Bellón, V. Stoven, and C.-A. Azencott. Multitask feature selection with task descriptors. *Pacific Symposium on Biocomputing*, volume 21, 261–272, 2016.

[12] J. C. Bouvy, M. L. De Bruin, and M. A. Koopmanschap. Epidemiology of adverse drug reactions in Europe: A review of recent observational studies. *Drug Safety*, 38(5):437–453, 2015.

[13] D. Brzyski, C. B. Peterson, et al. Controlling the rate of GWAS false discoveries. *Genetics*, 205(1):61–75, 2017.

[14] A. Chatr-Aryamontri, R. Oughtred, et al. The BioGRID interaction database: 2017 update. *Nucleic Acids Res.*, 45(D1):D369–D379, 2017.

[15] R. Cheng, J. Borevitz, and R. W. Doerge. Selecting informative traits for multivariate quantitative trait locus mapping helps to gain optimal power. *Genetics*, 195(3):683–691, 2013.

[16] W. Cheng, X. Zhang, Z. Guo, Y. Shi, and W. Wang. Graph-regularized dual Lasso for robust eQTL mapping. *Bioinformatics*, 30(12):i139–i148, 2014.

[17] Y.-K. Choi, and I. K. Park. Efficient GPU-based graph cuts for stereo matching. *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 642–648, 2013.

[18] R. Clarke, H. W. Ressom, et al. The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, 8(1):37–49, 2008.

[19] H. Climente, and C.-A. Azencott. martini: GWAS incorporating networks in R, 2017.

[20] D. Dernoncourt, B. Hanczar, and J.-D. Zucker. Analysis of feature selection stability on high dimension and small sample data. *Computational Statistics & Data Analysis*, 71:681–693, 2014.

[21] A. Drouin, S. Giguère, et al. Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons. *BMC Genomics*, 17:754, 2016.

[22] F. Eduati, L. Mangravite, et al. Opportunities and limitations in the prediction of population responses to toxic compounds assessed through a collaborative competition. *Nature Biotechnology*, 33(9):933–940, 2015.

[23] O. Fercoq, A. Gramfort, and J. Salmon. Mind the duality gap: safer rules for the Lasso. *Proceedings of the 32nd International Conference on Machine Learning*, 333–342. PMLR, 2015.

[24] R. Genuer, J.-M. Poggi, and C. Tuleau-Malot. Variable selection using random forests. *Pattern Recogn. Lett.*, 31(14):2225–2236, 2010.

[25] G. M. Grazier, W. Stefan, C. Alexandra, and T. Robert. Sequential selection procedures and false discovery rate control. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(2):423–444, 2015.

[26] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a posteriori estimation for binary images. *J. R. Stat. Soc.*, 51(2), 1989.

[27] Y. Han, and L. Yu. A variance reduction framework for stable feature selection. *Statistical Analysis and Data Mining*, 5(5):428–445, 2012.

[28] A.-C. Haury, P. Gestraud, and J.-P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 6(12):e28210, 2011.

[29] X. He, C. K. Fuller, et al. Sherlock: Detecting gene-disease associations by matching patterns of expression QTL and GWAS. *American Journal of Human Genetics*, 92(5):667, 2013.

[30] Z. He, and W. Yu. Stable feature selection for biomarker discovery. *Comp. Biol. Chem.*, 34(4):215–225, 2010.

[31] L. A. Hothorn, O. Libiger, and D. Gerhard. Model-specific tests on variance heterogeneity for detection of potentially interacting genetic loci. *BMC Genet.*, 13:59, 2012.

[32] J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. *J. Mach. Learn. Res.*, 12:3371–3412, 2011.

[33] Y. Huang, S. Wuchty, and T. M. Przytycka. eQTL epistasis – challenges and computational approaches. *Frontiers in Genetics*, 4, 2013.

[34] P. Jia, S. Zheng, J. Long, W. Zheng, and Z. Zhao. dmGWAS: dense module searching for genome-wide association studies in protein–protein interaction networks. *Bioinformatics*, 27(1):95–102, 2011.

[35] T. Kam-Thong, C.-A. Azencott, et al. GLIDE: GPU-based linear regression for detection of epistasis. *Human Heredity*, 73(4):220–236, 2012.

[36] H. M. Kang, C. Ye, and E. Eskin. Accurate discovery of expression quantitative trait loci under confounding from spurious and genuine regulatory hotspots. *Genetics*, 180(4):1909–1925, 2008.

[37] L. Kuncheva, C. Smith, Y. Syed, C. Phillips, and K. Lewis. Evaluation of feature ranking ensembles for high-dimensional biomedical data. *ICDM Workshops*, 49–56, 2012.

[38] I. Kuperstein, E. Bonnet, et al. Atlas of Cancer Signalling Network: a systems biology resource for integrative analysis of cancer data with Google Maps. *Oncogenesis*, 4(7):e160, 2015.

[39] T. Lappalainen, S. B. Montgomery, A. C. Nica, and E. T. Dermitzakis. Epistatic selection between coding and regulatory variation in human evolution and disease. *The American Journal of Human Genetics*, 89(3):459–463, 2011.

[40] J. Leskovec, and C. Faloutsos. Sampling from large graphs. *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 631–636, 2006.

[41] C. Li. Personalized medicine – the promised land: are we there yet? *Clinical Genetics*, 79(5):403–412, 2011.

[42] C. Lippert, F. P. Casale, B. Rakitsch, and O. Stegle. LIMIX: genetic analysis of multiple traits. *bioRxiv*, 003905, 2014.

[43] J. Liu, K. Wang, S. Ma, and J. Huang. Accounting for linkage disequilibrium in genome-wide association studies: A penalized regression method. *Statistics and its Interface*, 6(1):99–115, 2013.

[44] F. Llinares-López, M. Sugiyama, L. Papaxanthos, and K. Borgwardt. Fast and memory-efficient sig-

nificant pattern mining via permutation testing. *Knowledge Discovery and Data Mining 21*, 725–734, 2015.

[45] G. Louppe, L. Wehenkel, A. Sutera, and P. Geurts. Understanding variable importances in forests of randomized trees. *Advances in Neural Information Processing Systems 26*, 431–439. Curran Associates, Inc., 2013.

[46] K. L. Lunetta, L. B. Hayward, J. Segal, and P. Van Eerdewegh. Screening large-scale association study data: exploiting interactions using random forests. *BMC Genetics*, 5:32, 2004.

[47] S. Ma, J. Huang, and M. S. Moran. Identification of genes associated with multiple cancers via integrative analysis. *BMC Genomics*, 10:535, 2009.

[48] T. A. Manolio, F. S. Collins, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[49] C. A. Micchelli, J. M. Morales, and M. Pontil. Regularizers for structured sparsity. *Adv. Comput. Math*, 38(3):455–489, 2013.

[50] B. Mieth, M. Kloft, et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Scientific Reports*, 6:36671, 2016.

[51] J. H. Moore. A global view of epistasis. *Nature Genetics*, 37(1):13–14, 2005.

[52] K. Nakagawa, S. Suzumura, M. Karasuyama, K. Tsuda, and I. Takeuchi. Safe pattern pruning: An efficient approach for predictive pattern mining. *arXiv:1602.04548 [stat]*, 2016.

[53] N. C. Ndiaye, E. S. Said, et al. Epistatic study reveals two genetic interactions in blood pressure regulation. *BMC Med. Genet.*, 14:2, 2013.

[54] C. Niel, C. Sinoquet, C. Dina, and G. Rocheleau. A survey about methods dedicated to epistasis detection. *Bioinformatics and Computational Biology*, 285, 2015.

[55] R. Nilsson, J. M. Peña, J. Björkegren, and J. Tegnér. Consistent feature selection for pattern recognition in polynomial time. *Journal of Machine Learning Research*, 8(Mar):589–612, 2007.

[56] S. Nogueira, and G. Brown. Measuring the stability of feature selection. *Machine Learning and Knowledge Discovery in Databases* Lecture Notes in Computer Science 9852, 442–457. Springer International Publishing, 2016.

[57] S. Okser, T. Pahikkala, and T. Aittokallio. Genetic variants and their interactions in disease risk prediction – machine learning and network perspectives. *BioData Mining*, 6(1):5, 2013.

[58] H. Pang, A. Lin, et al. Pathway analysis using random forests classification and regression. *Bioinformatics*, 22(16):2028–2036, 2006.

[59] E. Perthame, C. Friguet, and D. Causeur. Stability of feature selection in classification issues for high-dimensional correlated data. *Statistics and Computing*, 1–14, 2015.

[60] G. Prat, and L. A. Belanche. Improved stability of feature selection by combining instance and feature weighting. *Research and Development in Intelligent Systems XXXI*, 35–49. Springer International Publishing, 2014.

[61] A. L. Price, C. C. A. Spencer, and P. Donnelly. Progress and promise in understanding the genetic basis of common diseases. *Proc. R. Soc. B*, 282(1821):20151684, 2015.

[62] L. Qu, T. Guennel, and S. L. Marshall. Linear score tests for variance components in linear mixed models and applications to genetic association studies. *Biometrics*, 69(4):883–892, 2013.

[63] E. E. Schadt, S. H. Friend, and D. A. Shaywitz. A network view of disease and compound screening. *Nature Reviews Drug Discovery*, 8(4):286–295, 2009.

[64] N. J. Schork. Personalized medicine: Time for one-person trials. *Nature News*, 520(7549):609, 2015.

[65] R. D. Shah, and R. J. Samworth. Variable selection with error control: another look at stability selection. *Journal of the Royal Statistical Society*, 75(1):55–80, 2013.

[66] S. K. Sieberts, F. Zhu, et al. Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis. *Nature Communications*, 7:12460, 2016.

[67] J. Stephan, O. Stegle, and A. Beyer. A random forest approach to capture genetic effects in the presence of population structure. *Nature Communications*, 6:7432, 2015.

[68] C. Sudlow, J. Gallacher, et al. UK Biobank: An Open Access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLOS Medicine*, 12(3):e1001779, 2015.

[69] M. Sugiyama, C. Azencott, D. Grimm, Y. Kawahara, and K. Borgwardt. Multi-task feature selection on multiple networks via maximum flows. *SIAM International Conference on Data Mining*, 199–207. Society for Industrial and Applied Mathematics (SIAM), 2014.

[70] S. Suzumura, K. Nakagawa, Y. Umezu, K. Tsuda, and I. Takeuchi. Selective inference for sparse high-order interaction models. *PMLR*, 3338–3347, 2017.

[71] S.J. Swamidass, C.-A. Azencott, et al. Influence Relevance Voting: An accurate and interpretable virtual high throughput screening method. *Journal of Chemical Information and Modelling*, 49(4):756–766, 2009.

[72] M. A. Taub, H. R. Schwender, S. G. Younkin, T. A. Louis, and I. Ruczinski. On multi-marker tests for association in case-control studies. *Front. Genet.*, 4:252, 2013.

[73] A. Terada, R. Yamada, K. Tsuda, and J. Sese. LAMPLINK: detection of statistically significant SNP combinations from GWAS data. *Bioinformatics*, btw418, 2016.

[74] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc.*, 58:267–288, 1994.

[75] K. A. Tryka, L. Hao, et al. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res*, 42(D1):D975–D979, 2014.

[76] P. M. Visscher, M. A. Brown, M. I. McCarthy, and J. Yang. Five years of GWAS discovery. *Am. J. Hum. Genet.*, 90(1):7–24, 2012.

[77] S. Wager, S. Wang, and P. S. Liang. Dropout training as adaptive regularization. *Advances in Neural Information Processing Systems 26*, 351–359. Curran Associates, Inc., 2013.

[78] B. Wang, A. M. Mezlini, et al. Similarity network fusion for aggregating data types on a genomic scale. *Nature Methods*, 11(3):333–337, 2014.

[79] L. Wang, T. Matsushita, L. Madireddy, P. Mousavi, and S. E. Baranzini. PINBPA: Cytoscape app for network analysis of GWAS data. *Bioinformatics*, 31(2):262–264, 2015.

[80] M. H. Wang, R. Sun, et al. A fast and powerful W-test for pairwise epistasis testing. *Nucleic Acids Research*, 44(12):e115–e115, 2016.

[81] H. R. Warren, E. Evangelou, et al. Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk. *Nat. Genet.*, 49(3):403–415, 2017.

[82] C. Widmer, C. Lippert, et al. Further improvements to linear mixed models for genome-wide association studies. *Sci. Rep.*, 4:6874, 2014.

[83] M. C. Wu, S. Lee, et al. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am. J. Hum. Genet.*, 89(1):82–93, 2011.

[84] B. L. Yaspan, and O. J. Veatch. Strategies for pathway analysis from GWAS data. *Current Protocols in Human Genetics*, Chapter 1, 2011.

[85] M. Yoshida, and A. Koike. SNPInterForest: A new method for detecting epistatic interactions. *BMC Bioinformatics*, 12(1):469, 2011.

[86] X. Zhang, S. Huang, F. Zou, and W. Wang. Tools for efficient epistasis detection in genome-wide association study. *Source Code Biol. Med.*, 6:1, 2011.

[87] Y. Zhang, and D.-Y. Yeung. A convex formulation for learning task relationships in multi-task learning. *UAI*, 733–742, 2009.