Detection of Gene-gene Interactions: Methodological Comparison on Real-World Data and Insights on Synergy between Methods

Hugo Boisaubert and Christine Sinoquet
LS2N, UMR CNRS 6004, University of Nantes, France

Keywords: Gene-gene Interaction, Machine Learning, Markov Blanket, High-dimensional Data, Comparative Analysis.

Abstract: In this paper, we report three contributions in the field of gene-gene interaction (epistasis) detection. Our first contribution is the comparative analysis of five approaches designed to tackle epistasis detection, on real-world datasets. The aim is to help fill the lack of feedback on the behaviors of published methods in real-life epistasis detection. We focus on four state-of-the-art approaches encompassing random forests, Bayesian inference, optimization techniques and Markov blanket learning. Besides, a recently developed approach, SMMB-ACO (Stochastic Multiple Markov Blankets with Ant Colony Optimization) is included in the comparison. Thus, our second contribution addresses assessing the behavior of SMMB-ACO on real-world data, while SMMB-ACO was mainly evaluated so far through small-scale simulations. We used a published case control dataset related to Crohn's disease. Focusing on pairwise interactions, we report a great heterogeneity across the methods in running times, memory occupancies, numbers of interactions output, distributions of p-values and odds ratios characterizing the interactions. Then, our third contribution is a proof-of-concept study in the context of genetic association interaction studies, to foster alternatives to analyses driven by prior biological knowledge. The principle is to cross the results of several machine learning methods whose intrinsic mechanisms greatly differ, to provide a priorized list of interactions to be validated experimentally. Focusing on the interactions identified in common by two methods at least, we obtained a priorized list of 56 interactions, from which we could infer one interaction network of size 7, four networks of size 4 and six of size 3.

1 INTRODUCTION

Over the past twenty years, automated high-throughput genotyping technologies have allowed a shift from candidate-gene analyses to genome-wide association studies (GWASs). The primary objective of GWASs is to detect associations (*i.e.*, statistical dependences) between genetic variants and a phenotype of interest, in a population under study. Aiming to better understand the biology of diseases, GWASs are expected to foster prevention and improve drug treatment, to usher the era of personalized medicine. In the latter, prevention and drug treatment are designed depending on the genetic profile of the patient.

Typically, in GWASs, between a few thousand to ten thousand subjects are genotyped, which provides the measure of DNA variation at characterized *loci* called genetic markers, spread over the genome. Single nucleotide polymorphisms (SNPs) are widely-used genetic markers. Depending on the microarray used, the number of SNPs ranges from a few hundred thousands to a few millions. From now on, we will consider SNP-based association studies.

New biological insights were gleaned by exploring GWAS hits for diseases such as inflammatory bowel disease, type 2 diabetes, cardiovascular diseases, bipolar disorder, as well as some cancers, to cite a few. However, most of the inherited risk remains to be explained for most phenotypes investigated so far, a situation named *missing heritability*. Therefore, complementary lines of investigation have started to explore alternative heritable components of complex phenotypes, encompassing rare variants, structural variants, epigenetics, and genetic interactions. This paper focuses on computational approaches designed to identify genetic interactions, also named epistatic interactions. From a statistical point of view, epistasis defines the deviation from the model in which the cumulative effects of multiple SNPs linearly determine the phenotype. A persuasive piece of evidence supports the role of genetic interactions to explain where part of the missing heritability hides: biomolecular interactions are ubiquitous in gene regulation and biochemical and metabolic systems (Furlong, 2013; Gilbert-Diamond and Moore, 2011). Biological evidence of epistasis has been put forward in several publicati-

Boisaubert, H. and Sinoquet, C.
Detection of Gene-gene Interactions: Methodological Comparison on Real-World Data and Insights on Synergy between Methods.
DOI: 10.5220/0007374400300042
In *Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019)*, pages 30-42
ISBN: 978-989-758-353-7

ons (Gao et al., 2010; Nicodemus et al., 2013; Gibert et al., 2017). The gap between the plausible high number of epistatic interactions existing in genomes and the limited number of results published may be explained by the computational challenge posed by epistasis detection. Moreover, loss-of-function mutations that occur *de novo* or persist within populations at low frequencies are known to significantly alter epistatic interactions (Mullis et al., 2018).

In the remainder of this article, a combination of SNPs that interact to determine a phenotype is called an interaction. A k-way interaction is a combination of k interacting SNPs. A 2-way interaction will also be called a gene-gene interaction (with SNPs either in exons or introns).

A key motivation for the large-scale comparative study reported in this paper lies in the following observation: we miss feedback about the respective behaviors of methods designed to implement GWASs on *real-world data*. This observation extends to Genetic Association Interaction Studies (GAISs), and *a fortiori* to genome-wide AIS (GWAISs). This paper contributes to fill this lack. Another strong motivation for our work was to analyze how SMMB-ACO (Sinoquet and Niel, 2018), a method proposed most recently, compares with other approaches, on *real* GWAIS data. The remainder of the paper is organized as follows. Section 2 presents a succinct overwiew of the recent state-of-the-art. Section 3 provides the motivations for our study. Section 4 depicts the five methods involved in our study, in a broad-brush way for the four reference methods chosen, and in more details for the recently developed SMMB-ACO. Section 5 focuses on the experimental protocol, the real-world datasets analyzed, the implementation and parameter adjustement of the five methods. The experimental results, discussion and feedback gained are presented in the last section.

2 RELATED WORK

Performing a GAIS is challenging. In the category of statistical approaches, **multivariable multiplicative linear regression** (MMLR) offers a framework to model the relationship between a continuous variable of interest (outcome) y and multiple interacting predictors $x_1, x_2, ..., x_q$ (continuous or categorical), such as in $y \sim \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \beta_{12}\ x_1 x_2$, with $q = 2$. In MMLR, interaction terms allow to escape from the pure linear scheme ($y \sim \beta_0 + \beta_1\ x_1 + \beta_2\ x_2$). One step further, **linear generalized regression** (LGR) provides a way to model an outcome that is not linearly determined by predictors: for this purpose, a *link* function f is used to transform the outcome y, to match the real distribution of y. Besides, similarly as for MMLR, interaction coefficients may be specified for LGR. For example, the LGR model to adjust in the case of two interacting predictors writes: $f(y) \sim \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \beta_{12}\ x_1 x_2$. Obviously, MMLR and LGR cannot be used to analyze data on a genome scale, because of the combinatorial issue posed by the enumeration of combinations of q predictors, with $2 \leq q \leq r$, and upper bound r arbitrarily set by the user. Moreover, identifying an appropriate link function f in LGR may not be trivial. To note, **logistic regression** (LR) is a specific case of LGR where the link function is known, and allows to model a binary outcome. Typically, in case control association studies, with p representing the probability to be affected by the pathology of interest, a LR model with two interacting predictors writes: $logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \beta_{12}\ x_1 x_2$. We will further specify to which aim and how LR is used in the comparative study reported here. Approaches in the line of **multifactor-dimensionality reduction** (MDR) are also compelled to test all combinations of q SNPs, and also fail to handle GWAS data unless GPU calculation is used (Gola et al., 2016).

The high dimensionality of GWAS data advocates the design of (supervised) machine learning and data mining approaches to tackle the problem of epistasis detection. A direct way to reduce the search space is to decrease the data dimensionality. **Relief-based approaches** (RBAs), **random forests** and **penalized regression** are three major feature selection techniques used in epistasis detection.

Algorithms in the line of Relief first compute (genetical) similarities between individuals; a nearest neighbor-based technique then allows to assess importances for SNPs with regard to the phenotype of interest (see (Urbanowicz et al., 2018) for a recent review). Not to speak of the bias induced by the pre-selection of SNPs marginally associated with the phenotype, the computation of pairwise similarities is prohibitively expensive.

A random forest (RF) is a set of decision trees grown from bootstrap samples of observations. At each node, a random subset of K predictors is used to determine the optimal split. The optimal split is the one that decreases the most node impurity for a classification tree (respectively the sum of squared errors for a regression tree) after the split. RFs applied to GWAS data produce a ranking of the markers, by decreasing importance measures (Schwarz et al., 2010; Yoshida and Koike, 2011). Until the 2010s, RF learning was computationally and memory inefficient in high-dimensional settings. The fast implementation

of random forests for high-dimensional data provided *via* ranger is one of the software programs included in our comparative study.

RBAs and RFs both allow to rank and select top scoring SNPs. In the epistasis detection context, a procedure is required downstream RBAs and RFs, to generate gene-gene interactions. This procedure may consist in statistical tests such as regression, or may be a specific approach dedicated to epistasis detection (*e.g.*, (Lin et al., 2012)).

In contrast, penalized regression (PR) such as regularized feature selection through Lasso, Ridge or Elastic Net regression can be used to directly output gene-gene interactions (Ayers and Cordell, 2010). A prominent downfall of these methods is the prohibitive running time. The method reported in (Chang et al., 2018) copes with high data dimensionality through a two-stage procedure: 2-way interactions are first detected within genes using Randomized Lasso and penalized Logistic Regression (RLLR); then, considering the list of SNPs obtained from the latter 2-way interactions, any combination of such two SNPs is tested using again RLLR, to identify cross-gene epistasis. The biological motivation for this kind of dimensionality reduction is questionable since the 2-way interactions within the genes are not kept.

In the panel of standard supervised machine learning and data mining techniques, **support vector machines** (SVMs) and **artificial neural networks** (ANNs) can be used directly for epistasis detection. To this aim, SVMs separate interacting and non-interacting combinations of SNPs using a hyperplane in multi-dimensional space (*e.g.*, (Shen et al., 2012)). ANNs allow to model non-linear feature interactions through network connections. The recent revolution in training feedforward networks with many hidden layers through advanced stochastic gradient descent open a path for deep neural networks (DNNs). However, the DNN used in (Uppu et al., 2016) was learned from small datasets (no more than 1,600 individuals, a few tens of SNPs). In the still more recent work reported in (Fergus et al., 2018), logistic regression was employed to pre-select around 5,000 SNPs to fit a deep learning model (1,500 subjects). Bayesian neural networks (BNNs) merge an ANN with a probabilistic model. Notably, this allows the quantification of variable influence with incertainty measures. In (Beam et al., 2014), BNNS were used to detect epistatic interactions on a relatively limited scale (around a hundred individuals described by 60,000 SNPs).

Bayesian networks (BNs) allow to model dependences between variables in an uncertain context. Therefore BNs offer an appealing framework for gene-gene interaction detection, to discover the best scoring graph structure connecting SNPs to the variable of interest. The branch and bound heuristic used in (Han and Chen, 2011) allowed to process a relatively limited dataset, a published AMD (Age Macular Degenerated) dataset, which describes around 150 individuals for about 110,000 SNPs. In (Jiang et al., 2010), a greedy search performing a forward phase (edge addition) followed by a backward phase (edge removal) is applied. However, for tractability reasons, the process starts including one pair of interacting SNPs exerting a marginal effect on the phenotype, thus addressing a specific case of epistasis called *embedded* epistasis. In the BN framework, feature subset selection stated as **Markov blanket** learning is another line of investigation. In a BN built over the variables of a dataset V, the Markov Blanket (MB) of a target variable T, $MB(T)$, is defined as a minimal set of variables that renders any variable outside $MB(T)$ probabistically independent of T, conditional on $MB(T)$. Typically, a MB of the phenotype is a set of interacting SNPs able to determine the phenotype. Thus, instead of learning a whole BN as abovementioned, algorithms were designed to learn the Markov blanket of a given phenotype of interest. However, in high-dimensional settings, the complexity of MB learning remains challenging. FEPI-MB (Fast epistatic interactions detection using Markov blanket) (Han et al., 2011) and DASSO-MB (Detection of ASSOciations using Markov Blanket) (Han et al., 2010) were able to process the AMD dataset previously mentioned.

Bayesian inference is employed by the popular BEAM algorithm (Bayesian Epistasis Association Mapping) (Zhang and Liu, 2007). BEAM partitions SNPs into three groups: SNPs with a marginal effect on the phenotype, SNPs that jointly contribute to the phenotype, and background SNPs. A Markov chain Monte Carlo (MCMC) process exploits Bayes theory to partition the SNPs into these three groups. Datasets with half a million of SNPs could be processed by BEAM, at the cost of high running times (up to a week and even more).

Other approaches, derived from the **optimization** field, have been proposed to search the space of combinations of SNPs. The method reported in (Aflakparast et al., 2014) combines Bayesian scoring with an **evolutionary-based** heuristic approach; it allowed to process around 1,400 subjets and 300,000 SNPs. **Ant colony optimization** (ACO) was exploited by several proposals. AntEpiSeeker, a widely cited reference, relies on the straightforward adaptation of classical ACO to epistasis detection (Wang et al., 2010), and is tractable on the genome scale. The objective to max-

imize in AntEpiSeeker is the χ^2 statistics of the test of dependence between a group of SNPs and the phenotype. The method described in (Sun et al., 2017) seeks to optimize an objective combining the mutual information measure with a BN-based score, and was able to process the AMD dataset abovecited. **Multi-objective ACO optimization** was also proposed, in which the Akaike information criterion (AIC) score and a BN-derived score must be optimized (Jing and Shen, 2015). The computational burden limited the analysis to separate chromosome datasets.

3 MOTIVATIONS FOR THE STUDY

In light of the literature published on the subject, we can draw a number of remarks about the evaluation and comparison of methods developed to cope with epistasis detection. First, to evaluate the performance of a method, multiple datasets (for instance 100) must be generated under some controlled (*i.e.*, simulated) condition. To generate statistics for this condition, a performance measure (*e.g.*, power, F-measure, for example) must be computed for each dataset. As this performance measure is a function of true positives, false positives and false negatives, its computation requires that multiple executions (for instance, 100) of the same stochastic method be achieved for each simulated dataset. Thus, for tractability reasons, simulations on the genome scale are only accessible to centres with outstanding intensive computing and storage resources (*e.g.*, (Chatelain et al., 2018)), and the overwhelming majority of studies still compare methods on simulated datasets describing 100 SNPs (and a few thousand observations). It follows that the methods are compared, and subsequent conclusions drawn, in conditions that in no way reflect the real-world situation of GWAS analyses.

Second, for the same tractability reasons, publications analyze the method they propose on a single GWAS dataset but never perform comparisons with other methods on GWAS datasets. On the one hand, comparing several methods requires authors to adjust parameters for methods they did not develop and do not always know well. For those approaches that rely on supervised machine learning, parameter adjustment means running 10 times the method (in a 10-fold cross-validation scheme) under each instantiation (from a grid of instantiations) of the set of parameters. The computational burden is therefore prohibitive. On the other hand, the same parameter adjustment issue arises for optimization-based approaches. Besides, the latter methods generally output several solutions and this number of such solutions may vary across the grid of instantiations. Thus, it is not straightforward to assign a score to such outputs, which impedes the ability to rank instantiations.

Third, in publications focused on a novel method, running times are but exceptionally reported for executions on simulated datasets, as well as for executions on real GWAS datasets. Moreover, when a publication compares a novel method with other methods, the comparison of running times across methods is practically always missing.

Fourth, so far, no extensive AIS analysis was designed to provide interactions jointly identified by several approaches, with the aim of generating a short list for further biological validation.

The works reported in this paper were designed with the previous four points in mind. Finally, another strong motivation for our work was to analyze how SMMB-ACO (Sinoquet and Niel, 2018), a method proposed most recently, compares with state-of-the-art approaches, on *real* data.

4 THE FIVE APPROACHES COMPARED

The four state-of-the-art approaches selected fall into various categories: random forest, Bayesian inference, ant colony optimization, Markov blanket learning. The newly developed SMMB-ACO method combines Markov blanket learning with ant colony optimization.

So far, any novel method proposed was generally compared to Random Jungle (Schwarz et al., 2010). We therefore selected ranger, the successor of Random Jungle, which is a fast implementation of random forests to handle high-dimensional data (Wright and Ziegler, 2017).

Also a reference for epistasis detection, BEAM3 (Zhang, 2012), the successor of BEAM (Zhang and Liu, 2007), was incorporated in our study. Similarly as BEAM, BEAM3 employs a MCMC search technique to probabilistically assign SNPs to three groups (background, marginal dependence with the phenotype, involvment in an interaction). Moreover, the MCMC simulation allows to assign a statistical significance to each SNP, thus avoiding costly permutation-based tests. A major difference with BEAM is that BEAM3 detects flexible interaction structures using *disease* graphs. Besides, BEAM3 dynamically accounts for the unknown linkage disequilibrium (LD) among SNPs. LD is defined as the network of dependences that exists among genetic data, as the result of evolutionary events. The aim is to fil-

ter out the secondary associations due to LD: secondary associations will not be admitted in the disease graph when better candidates are already present. On the one hand, the complexity of the disease graph is reduced. On the other hand, it is expected that the primary disease associations are reported with improved resolution. BEAM3 was able to process around 4,700 subjects described by about 400,000 SNPs.

AntEpiSeeker (Wang et al., 2010), a reference in epistatis detection, is the third method considered in our study. In each iteration of AntEpiSeeker, ants each select a SNP set of user-defined size from the initial dataset, according to a probability distribution $\mathbb{P}$, and calculate a χ^2 statistics to assess the dependence strength with the phenotype. Feedback on the learning process is memorized through the so-called pheromone levels, based on the χ^2 statistics. The SNP sets with the highest χ^2 statistics are recorded. The probability distribution $\mathbb{P}$ is updated based on the pheromone levels following the standard ACO scheme. At the end of the iterations, a user-specified number of best SNP sets is available, together with a list $\mathcal{L}$ of top SNPs ranked by decreasing pheromone levels. Finally, in a post-processing phase, each best set S is examined: if the size of the interactions to be uncovered is q, the subsets of S of size q whose SNPs are all in $\mathcal{L}$ are kept as solutions. To note, the false positive issue is addressed as follows: if two interactions overlap, the one with the smaller p-value is kept.

FEPI-MB (Han et al., 2011) and DASSO-MB (Han et al., 2010) are two deterministic algorithms that tackle epistasis detection through feature subset selection based on Markov blanket learning. The key ingredients in these two algorithms are the forward and backward phases. In a forward step, a SNP is added to the growing MB provided it is the candidate SNP most dependent with the phenotype, conditional on the MB, and that this dependence is statistically significant. Conversely, a backward phase successively examines all SNPs belonging to the current MB; each such SNP is removed from the MB based on (statistically significant) conditional independence. We chose DASSO-MB, whose backward phase is more elaborate than in FEPI-MB: FEPI-MB removes a SNP if it is shown significantly independent conditional on the current MB; in contrast, DASSO-MB discards a SNP as soon as it shown independent with the phenotype, conditional of a *subset* of the current MB. Indeed, such a subset could be the MB to be discovered.

Finally, at the crossroads of machine learning and optimization, the fifth method retained in our study is SMMB-ACO (Stochastic Multiple Markov Blankets with Ant Colony Optimization) (Sinoquet and Niel, 2018). SMMB-ACO is an hybrid approach that combines Markov blanket construction with stochastic and ensemble features. To address the issue of scalability in high-dimensional settings, SMMB-ACO relies on a heuristic designed to search promising areas of the search space.

In each iteration of SMMB-ACO, several ants each learn a suboptimal Markov blanket from a subset of SNPs sampled from the initial set. The MB learning performed by each ant runs a forward phase intertwined with backward phases. In this respect, MB learning in SMMB-ACO is similar to that in DASSO-MB. However, a genuine difference in the SMMB-ACO and DASSO-MB forward steps is the following: SMMB-ACO stochastically adds a group of SNPs associated with the phenotype, whereas DASSO-MB incorporates the SNP most associated with the phenotype. The two MB learning algorithms are described and commented in Figures 1 (a) and (b). The stochastic feature of SMMB-ACO relies on SNP sampling, following a probability distribution $\mathbb{P}$ updated based on pheromone levels. It is possible to specify a specific operating mode for SMMB-ACO, to cope with high-dimensional data: a two-pass process is then triggered. Figures 1 (c) and (d) outline this process.

5 EXPERIMENTAL SETTING

We first present the experimental protocol. Then, the real-world datasets used are briefly described. Third, we focus on implementation aspects. This section ends with considerations about the parameter adjustment of the approaches compared.

5.1 Experimental Protocol

We consider SNPs coded on 0, 1 and 2 to respectively denote major homozygous, heterozygous and minor homozygous, where the allele with minor frequency is the disease susceptibility allele. We call **interaction of interest** (IoI) any 2-way interaction for which logistic regression ($y \sim \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \beta_{12}\ x_1 x_2$) provides a significant p-value for the interaction coefficient β_{12}, given some specific significance threshold. As highlighted previously, the RF-based approach ranger can only tackle feature selection. Downstream ranger's execution, we thus generated C_{20}^2 2-way interactions from the selection of 20 SNPs with the highest importance measures. Then we selected the IoIs at significance threshold 5×10^{-4}. To put all approaches on the same footing for the comparison, we filtered out the outputs of BEAM3, AntEpiSeeker and DASSO-MB and adap-

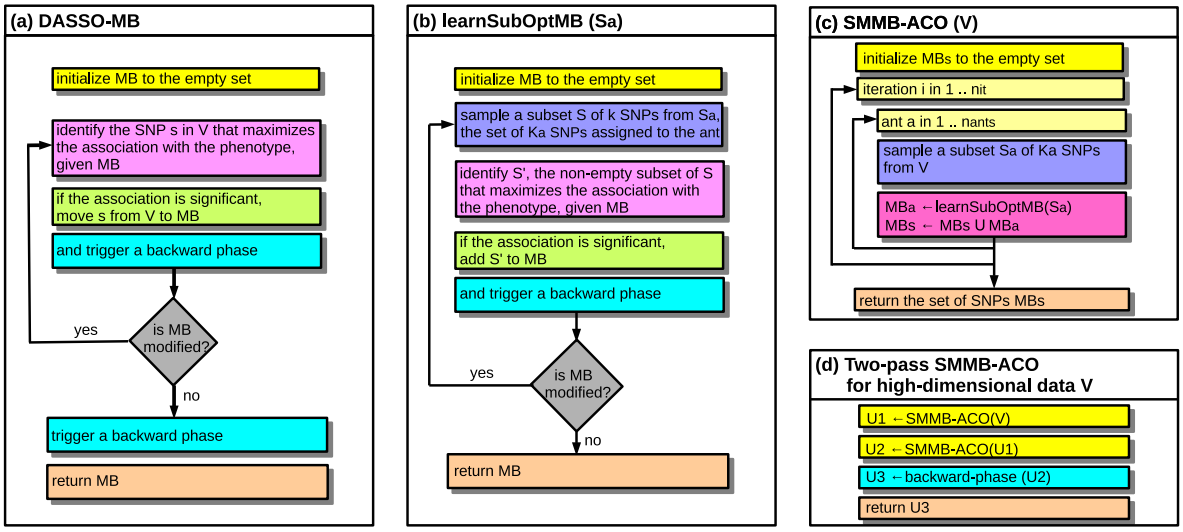


Figure 1: Sketches of the DASSO-MB and SMMB-ACO algorithms. (a) DASSO-MB. (b) SMMB-ACO stochastic procedure to learn a suboptimal Markov blanket. (c) and (d) Two-pass SMMB-ACO algorithm adapted to high-dimensional data. MB: Markov blanket. V is the initial set of SNPs. (a) Incorporating SNPs one at a time as in DASSO-MB hampers the epistasis detection process: since the independence test achieved at first iteration is conditioned on the empty Markov blanket (MB), a SNP marginally dependent with the phenotype is incorporated from the outset, which biases the whole MB learning. (b) Instead, SMMB-ACO addresses this issue by including groups of SNPs. To this aim, each forward step starts with the sampling of a set S of k SNPs, from the subset S_a of size K_a that is assigned to the ant that is driving the learning of the suboptimal MB. For each non-empty subset S' of S, a score is computed, that assesses the association strength between S' and the phenotype, conditional on the growing Markov blanket MB. The subset S' with the highest association score is incorporated into MB if the association is statistically significant. (c) After all iterations are completed, the set of SNPs obtained as the union of all suboptimal MBs is returned. This set of SNPs is the set U_1 returned by the first pass of SMMB-ACO (see (d)). (d) A second pass of SMMB-ACO is performed with U_1 as the input. This time, the resulting set U_2 is submitted to a backward phase, to yield U_3, a set of SNPs. To include SMMB-ACO in our experimental protocol, we suppressed the post-processing phase of the native algorithm (Sinoquet and Niel, 2018), which outputs as an interaction any suboptimal MB generated in (b) provided it is contained in the set U_3 obtained in (d). In our protocol, for reasons detailed in subsections 5.1 and 5.4 (last paragraph), the post-processing phase of SMMB-ACO consisted in the generation of interactions of interest (IoIs), as defined in subsection 5.1, from set U_3.

ted the post-processing in SMMB-ACO to keep only IoIs at significance threshold 5×10^{-2}. The thorough justification for the use of two significance thresholds will be provided in subsection 5.4.

In this comparative analysis, DASSO-MB is the only deterministic approach. Each other (stochastic) method was run 10 times on each dataset.

To note, generating all 2-way interactions from a set of t SNPs and assessing their dependence with the phenotype through logistic regression may be computationally expensive (*e.g.*, 30 hours if $t = 20$ SNPs). This result shows the necessity to use advanced methods for datasets scaling in tens of thousands of SNPs, as is the case in this study.

5.2 Real-World Datasets

We used the genome-wide data related to Crohn's disease (CD) provided by the Wellcome Trust Case Control Consortium (WTCCC, https://www.wtccc.org.uk/). Major pathways involved in Crohn's disease have emerged from standard single-SNP GWASs (Graham and Xavier, 2013). This background motivated our choice to analyze the WTCCC dataset related to Crohn's disease. Using the cohort of cases affected by CD and two cohorts of unaffected (controls) provided by the WTCCC, we generated 23 datasets related to the 23 human chromosomes. We applied the quality control procedure specified by the WTCCC to each dataset. In particular, this procedure dismisses SNPs having more than 1% of missing data and subjects having more than 5% of missing data, and checks for the so-called Hardy-Weinberg equilibrium at 5.7×10^{-7} threshold. After quality control, the size of the population of cases and controls is 4,686 (1,748 affected; 2,938 unaffected). The statistics about the number of SNPs per dataset are as follows: the average is 20,236; the minimum and maximum are 5,707 and 38,730, respectively. Finally, we imputed data using a k-nearest neighbor procedure, in which the missing variant of subject s is assigned the variant most frequent in the nearest neighbors of s.

Table 1: Implementations for the five software programs used in the comparative study.

ranger	http://dx.doi.org/10.18637/jss.v077.i01
BEAM3	http://www.mybiosoftware.com/beam-3-disease-association-mapping.html
AntEpiSeeker	http://nce.ads.uga.edu/ romdhane/AntEpiSeeker/index.html
DASSO-MB	not distributed by its authors, reimplemented
SMMB-ACO	https://ls2n.fr/listelogicielsequipe/DUKe/130/SMMB-ACO

5.3 Implementation of the Comparative Study

Except for DASSO-MB, all approaches are available on the Internet (Table 1); they are coded in C++. We recoded DASSO-MB in C++. The extensiveness of our comparative study required intensive computing resources from the Tier 2 CCIPL data centre (Intensive Computing Centre of the Pays de la Loire Region) (Intel 2630v4, 2×10 cores 2,2 Ghz, 20×6 GB). We exploited the OpenMP intrinsical parallelization of the C++ implementations of ranger, BEAM3 and SMMB-ACO. We also exploited data-driven parallelization to run each stochastic method 10 times on each dataset. Because of the heterogeneity of the running times across the methods and of memory shortage events, we had to balance the workload distribution between (i) sequentially processing 23 chromosome datasets for one method on one node (process_23Chrs_1) and repeating this job 9 times (on other nodes), and (ii) processing a single chromosome dataset 10 times for one method on one node (process_1Chr_10) and repeating this job for the remaining chromosomes (on other nodes). In the case of the parallelized software programs ranger, BEAM3 and SMMB-ACO, the 20 cores of a given node were employed in parallel. We managed the workload using the three following modalities: short, medium and long, for expected calculation durations respectively below 1, 5 and 30 days. When a timeout occurred in a node, depending on the degree of completion of the job, we either switched to a modality with higher time limit (process_23Chrs_1) or to a chromosome by chromosome management (process_1Chr_10). In total, we performed 1,035 chromosome-wide analyses.

5.4 Parameter Adjustment

Most machine learning methods require the tuning of a number of parameters. Table 6 in Appendix recapitulates the main parameters of the software programs used in our study.

The software program ranger was specifically designed by its authors to handle high-dimensional data. Through a complementary study (results not shown), we tried various values of mtry between $\sqrt{n}$ and n, the total number of SNPs. On the datasets concerned, the optimal value is shown to be $\frac{5}{8}n$.

To set the value of the product *number of iterations* $\times$ *number of ants* in AntEpiSeeker while attempting to diminish the large number of interactions output by this method, we conducted a preliminary study. In this preliminary study, the number of iterations was kept to AntEpiSeeker default value (450). We varied the number of ants between 500 and 5,000 (step 500). We observed that using 1,000 ants, we could control the total number of interactions reported to less than 15,000. while still guaranteeing a coverage of 10 for each SNP in the largest chromosome-wide dataset.

We set the numbers of iterations of the burn-in and stationary phases of BEAM3, following the recommendation of its author.

DASSO-MB's unique parameter is a type I error threshold, and its adjustment is straightforward.

For a fair comparison, in theory, one would set the product $n_{it} \times n_{ants}$ (number of ACO iterations $\times$ number of ants) in SMMB-ACO to the value chosen for AntEpiSeeker. However, two points must be taken into account. First, AntEpiSeeker software program is not parallelized, whereas SMMB-ACO is: during each of the n_{it} SMMB-ACO iterations, n_{ants} Markov blankets are learned in parallel. Second, the complexities of an iteration in AntEpiSeeker and of an iteration in SMMB-ACO are not comparable: in AntEpiSeeker, each ant draws a set of SNPs and computes the corresponding χ^2 statistic; in SMMB-ACO, each ant grows a Markov blanket *via* a forward phase intertwined with full backward phases. We adjusted SMMB-ACO parameters n_{it}, n_{ants} and K_a (number of SNPs drawn by each ant), in order to guarantee in theory that each SNP of the initial dataset would be drawn a sufficient number of times in the scope of a single run. With the parameter setting $(n_{it}, n_{ants}, K_a) = (360, 20, 160)$, we expect a coverage of 30 for the largest datasets, in a single run. We recall that 10 runs are performed for each stochastic method.

A type I error threshold is used for the independence tests in AntEpiSeeker, and for the conditional independence tests in DASSO-MB and SMMB-ACO. The common value choosen was 5×10^{-4}. It is common to the threshold fixed for the logistic regression used downstream ranger execution. A less stringent threshold of 5×10^{-2} was used for the logistic regressions performed in the filtering stages downstream BEAM3, AntEpiSeeker and DASSO-MB executions as well as in the post-processing stage in

Table 2: Orders of magnitude of the running times and memory occupancies for the five software programs used. Otherwise stated, the average running time indicated is computed from the 23 chromosome datasets (ten executions for each dataset).

Method	Average running time	Memory occupancy
ranger	feature selection: 14 ± 7 *mn* post-processing: 60 ± 20 *mn*	2 ± 0.6 GB
BEAM3	extremely volatile across the chromosome datasets Chr7 to Chr23: 54 ± 66 *s* Chr6: 22.4 ± 1.5 *h* Chr1 to Chr5: above 8 days	79 ± 46 GB
AntEpiSeeker	16 ± 3 *mn*	0.5 ± 0.2 GB
DASSO-MB	82 ± 22 *s*	1.5 ± 0.7 GB
SMMB-ACO	extremely volatile across the chromosome datasets 30 ± 17 *mn* (average on shortest executions) otherwise, up to 3 days, with large variations	43 ± 17 GB, extremely volatile even across the 10 executions on a given chromosome dataset; many execution abortions due to memory limitation (120 GB)

SMMB-ACO. A recapitulation is provided in Figure 2.

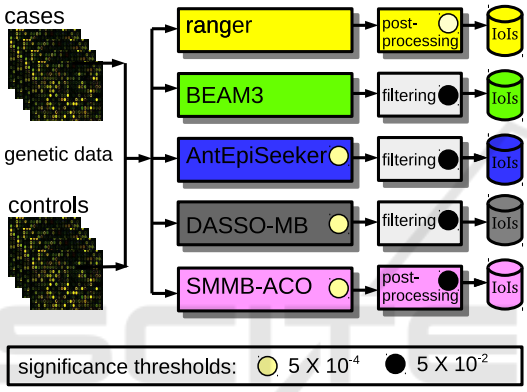


Figure 2: Flow diagram for the comparative analysis.

6 RESULTS AND DISCUSSION

We first compare running times and memory occupancies across the five approaches. Then we compare the numbers of interactions of interest (IoIs) identified by the five methods and analyze the distributions of p-values and odds ratios obtained. Third, we focus on the IoIs jointly identified by several methods. This section ends with a discussion.

6.1 Running Times and Memory Occupancies

There are salient features to draw from Table 2.

DASSO-MB is the software program both much faster and far less greedy in memory than its competitors. AntEpiSeeker is remarkable in that it shows a low running time across all chromosomes. The quickness of ranger is further impeded by the exhaustive test of 2-way interactions performed downstream. The behaviors of BEAM3 and SMMB-ACO are both extremely volatile across the datasets, for various reasons. In SMMB-ACO, the stochastic feature translates in the great heterogeneity of memory occupancy, possibly up to memory shortage, even for short chromosomes. In around the third of the datasets, it was necessary to launch additional runs (up to 5), to obtain the 10 runs required by our protocol. Nevertheless, the processing of all chromosomes remains feasible within 5 days, on 10 nodes. In contrast, we experienced timeouts with BEAM3, for the 5 largest chromosomes. In these cases, we were compelled to specify large timeouts (30 days), with the consequence of longer waiting times, to guarantee that executions demanding more than 8 days could be completed. Despite these prohibitive running times, BEAM3, the program most greedy in memory on average for the datasets considered, never ran out of memory.

6.2 Interactions of Interest Identified

Numbers of Interactions of Interest. Table 3 highlights contrasts between the methods. First, with only 18 interactions, it was nearly expected that DASSO-MB would not detect IoIs. In the remainder of this article, we will not mention DASSO-MB anymore. Second, a salient feature is the great heterogeneity in the numbers of IoIs detected by the four other methods. These numbers scale in a ten thousands, a thousand, a hundred and a few tens for AntEpiSeeker, SMMB-ACO, BEAM3 and ranger respectively.

Figure 3 focuses on the distribution of IoIs across the chromosomes. A first conclusion is that a sharp contrast exists between AntEpiSeeker and SMMB-ACO, whose IoIs are abundantly present in nearly all chromosomes, and BEAM3 and ranger, whose IoIs are confined to 10 and 5 chromosomes respectively. Besides, the number of IoIs in BEAM3, around four times higher than in ranger, is circumbscribed to a number of chromosomes that is two times less than for ranger.

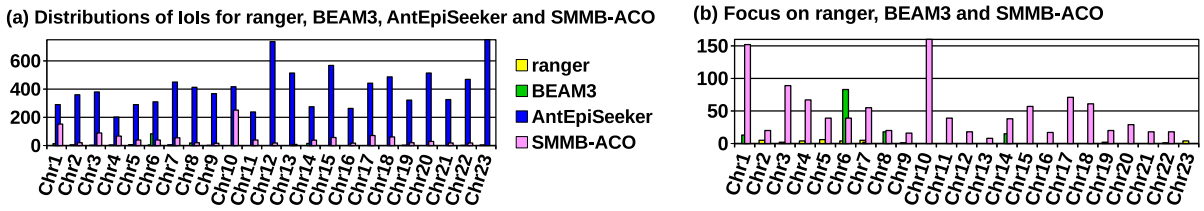


Figure 3: Comparison of the distributions of *interactions of interest* detected with ranger, BEAM3, AntEpiSeeker and SMMB-ACO. AntEpiSeeker detected 13,062 IoIs which are spread over the 23 chromosomes (smallest number of IoIs for a chromosome: 202; median number: 380). Moreover, IoIs are overly abundant in chromosome X, whose presence is not known to bias Crohn's disease onset (4,427 IoIs representing 34.9% of AntEpiSeeker's IoIs; the corresponding bar is truncated in subfigure (a)). These observations comfort the hypothesis of a high rate of false positives. SMMB-ACO identified 1,142 IoIs distributed across all chromosomes except chromosome X (smallest number of IoIs for a chromosome: 8; median number 38; largest number: 251; the corresponding bar (Chr10) is truncated in subfigure (b)). In constrast, the 131 IoIs detected by BEAM3 are located within 5 chromosomes only, whereas the 34 IoIs identified by ranger are distributed across 10 chromosomes. As regards BEAM3, Chr1, Chr6, Chr7, Chr8 and Chr14 respectively harbour 13, 83, 2, 18 and 15 IoIs. The IoIs detected by ranger are located on Chr2 to Chr7, Chr9, Chr19, Chr22 and Chr23 (minimum number of IoIs for these 10 chromosomes: 1; maximum number: 6).

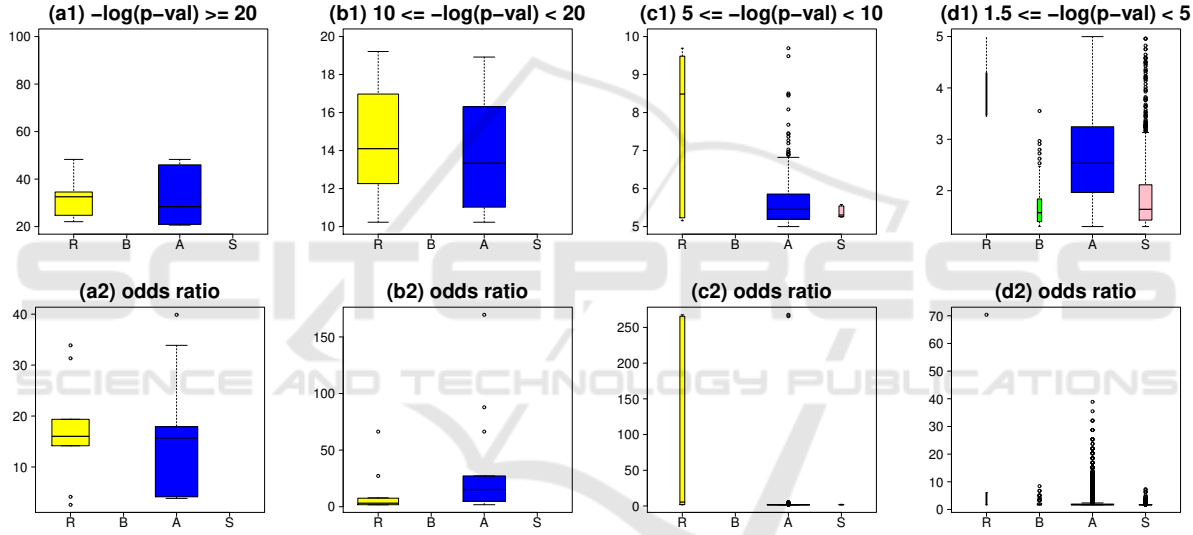


Figure 4: Distributions of p-values and odds ratios for the *interactions of interest* detected with ranger, BEAM3, AntEpiSeeker and SMMB-ACO. IoIs: interactions of interest. R: ranger. B: BEAM3. A: AntEpiSeeker. S: SMMB-ACO. Each subfigure (x2) shows the distribution of odds ratios for the IoIs whose p-values fall into subfigure (x1). $-log_{10}(5 \times 10^{-2}) = 1.5$.

Table 3: Comparison of the numbers of interactions detected with the five approaches.

	Number of interactions identified	Number of interactions of interest (IoIs)
ranger	34	**(34)** (100%)
BEAM3	1,082	**131** (12.1%)
AntEpiSeeker	14,670	**13,062** (89.0%)
DASSO-MB	18	0
SMMB-ACO	6,346	**1,142** (18.0%)

Distributions of P-values and Odds Ratios. The subfigures 4 (a1) to (d1) and Table 4 allow to compare the distributions of p-values observed for the IoIs. We consider four intervals for the p-values. Again, a great heterogeneity is observed across the methods. A first remark is that AntEpiSeeker and ranger are the only two methods to show p-values within the two first intervals (*i.e.*, below 10^{-10}) (even down to 10^{-50} for some outliers in both methods). A second observation is that ranger and AnEpiSeeker's p-values spread over whole third interval $]10^{-10}, 10^{-5}]$, whereas SMMB-ACO's lowest p-values range in $[10^{-5.5}, 10^{-5}]$. In contrast, BEAM3 is the only method whose 131 p-values are all contained in the fourth interval (and are even confined to $[10^{-3.5}, 5 \times 10^{-2}]$). Besides, another discrepancy is evidenced: we have seen that ranger and AntEpiSeeker's IoIs are distributed in all four intervals; however, a sharp contrast exists between these methods. Two thirds of the 34 ranger p-

Table 4: Distributions of p-values for the *interactions of interest* detected with ranger, BEAM3, AntEpiSeeker and SMMB-ACO. Four significance intervals are shown for $-log_{10}$(p-value). $-log_{10}(5 \times 10^{-2}) = 1.5$.

	≥ 20	[10, 20[	[5, 10[	[1.5, 5[
ranger	10	12	6	6
BEAM3	0	0	0	131
AntEpiSeeker	13	13	458	12,578
SMMB-ACO	0	0	6	1,136

values are concentrated in the two first intervals, whereas an overwhelming majority of the 13,062 AntEpiSeeker p-values are distributed in the third and fourth intervals, with a balance of 3.6% / 94.4% between these intervals, respectively. In addition, it is now clear that the statistical threshold of 5×10^{-4} in the post-processing downstream ranger is not a functional equivalent for the same statistical threshold used by AntEpiSeeker, DASSO-MB and SMMB-ACO during their learning processes.

In parallel, subfigures 4 (a2) to (d2) show the distributions of odds ratios for the interaction coefficients. In standard GWASs, SNPs with frequencies between 1% and 5% exert effects of moderate size, with odds ratios below 1.5, most often in range [1.1,1.3], and up to 2.1 (Stadler et al., 2010). A few publications document odds ratios for GWIASs. For example, odds ratios in range [1.3,1.4] are reported in (Li et al., 2018). Nevertheless, much higher odds ratios may be reported, such as values between 2 and 7, and even up to around 12 in (Grange et al., 2015). If we consider odds ratios above 20 as outliers, ranger and AntEpiSeeker are the two only methods that provide such outliers. The higher number of outliers observed for AntEpiSeeker can be explained by a much higher number of IoIs. In contrast, the odds ratios observed for BEAM3 and SMMB-ACO are all below 9 (see Figure 5). Consistently with (Grange et al., 2015), we observe that outliers for odds ratios do not necessarily coincide with outliers for p-values.

Interactions of Interest Jointly Identified by at Least Two Approaches. Beyond the methodological comparison on a real dataset, we wish to examine whether IoIs were jointly output by at least two approaches. The fact that two methods whose core mechanisms greatly differ identify common IoIs suggests that the corresponding short list of IoIs could be tested in priority by the biologists (Ritchie and Van Steen, 2018). None of the 131 IoIs identified by BEAM3 is detected by another method. On the contrary, 32 of the 34 IoIs detected by ranger were also detected by AntEpiSeeker. AntEpiSeeker and SMMB-ACO detected 16 common IoIs. SMMB-ACO and ranger have only 3 IoIs in common. One IoI was jointly

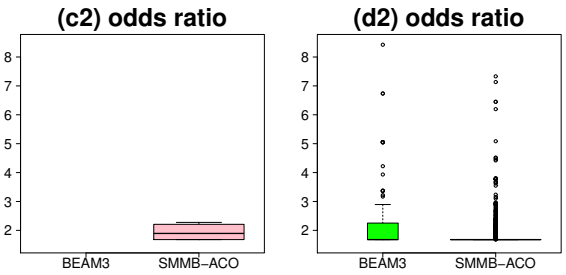


Figure 5: Focus on subfigures 4 (c2) and (d2) showing the distributions of odds ratios for BEAM3 and SMMB-ACO.

identified by AntEpiSeeker, ranger and SMMB-ACO. Given the number of interactions output by AntEpiSeeker, an overlap was expected between AntEpiSeeker and some other method. On the other hand, our study indicates that the mechanisms behind AntEpiSeeker and SMMB-ACO, which both use ACO, explore different sets of solutions.

Biological Insights. The 56 IoIs jointly selected by two or three methods are related to 25 known genes. From these 56 IoIs, we could infer 11 interaction networks: six of size 3 in Chr2, Chr5, Chr6, Chr7 and Chr19; four of size 4 in Chr3, Chr14, Chr16 and Chr17; and one of size 7 within Chr10. Besides a number of standard single-SNP GWASs, the few AISs devoted to CD focus on genes or pathways already known to contribute to the disease onset. It is not a surprise that our study highlights six genes already known to impact CD onset: NOD on Chr16, CCNY and NKX2-3 on Chr10, LGALS9 and STAT3 on Chr17, and SBNO2 on Chr19 (McGovern et al., 2015; Khor et al., 2011). It was also expected that our protocol designed for AIS investigation without prior biological knowledge would detect novel interaction candidates, which it does.

The network of size 7 is presented in Table 5. It is related to 5 known genes. It is beyond the scope of this study focused on methodological and computational aspects, to bring deeper biological insights on the potential mechanisms involved in the networks and IoIs.

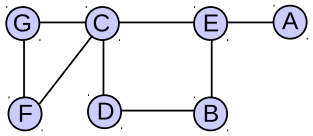
6.3 Discussion

The presence of p-value and odds ratio outliers is not an issue in AntEpiSeeker, in regard to the large number of AntEpiSeeker IoIs. However, this presence in the few ten IoIs output by ranger comforts the necessity to lower the significance threshold used for ranger down to the threshold used in filtering and post-processing, for the other methods (in our case

Table 5: Network of 7 SNPs identified in chromosome 10. iv: intron variant; nctv: nc transcript variant; uv: upstream variant. LINC01475 (long intergenic non-protein coding RNA 1475), related to nodes D, E and F in the interaction network, is expressed in 7 tissues including colon, small intestine, duodenum and appendix. NKX2-3, the other gene related to node F, is a member of the NKX family of homeodomain-containing transcription factors; the latter are involved in many aspects of cell type specification and maintenance of differentiated tissue functions. Node A is related to CREM which encodes a transcription factor that binds to the cAMP responsive element found in many cellular promoters. Alternative promoter and translation initiation site usage enables CREM to exert spatial and temporal specificity in cAMP-mediated signal transduction. This gene is broadly expressed (36 tissues including colon, small intestine and appendix). Node B is related to CCNY. As all cyclins, CCNY controls cell division cycles, regulates cyclin-dependent kinases; it is ubiquitous (27 tissues). Node G, CPXM2, a protein of the carboxypeptidase X, M14 family member 2, is broadly expressed in 21 tissues.

	SNP identifier	location (bp)	related gene	function
A	rs2505639	35185493	CREM	iv
B	rs16935948	35260820	CCNY	iv
	rs3936503	35260329	CCNY	iv
C	rs10761659	62685804	—	—
	rs10995271	62678726	—	—
D	rs10883365	99528007	LINC01475	nctv
	rs10883367	99528233	LINC01475	iv
E	rs1548964	99529896	LINC01475	iv
	rs1548962	99529978	LINC01475	iv
	rs6584283	99530544	LINC01475	iv
F	rs10883371	99532698	LINC01475 NKX2-3	uv 2kb
G	rs7067790	123917521	CPXM2	iv
	rs17680424	123917559	CPXM2	iv

	IoI	p-value	IoI	p-value
	AE	0.00358	CD	0.04767
	BD	0.00112	CF	0.04386
	BE	0.00115	CG	0.00624
	CE	0.04267		



5×10^{-2}). We would expect a larger number of IoIs for ranger, and thus a lower proportion of IoIs with extreme odds ratios or p-values.

On the CD dataset, DASSO-MB is of no help. The verbose AntEpiSeeker provided a wealth of results in both a wide spectrum of p-values and of odds ratios. SMMB-ACO neither provided outliers for p-values or odds ratios, but generated plausible odds ratios (up to 9) and showed lowest p-values in the order of $10^{-5.5}$. The widely cited software BEAM3 could not pinpoint IoIs with p-values lower than $10^{-3.5}$. In this respect, SMMB-ACO seems more promising than the renowned BEAM3, on the CD dataset.

On the CD dataset, 56 IoIs were detected by two methods at least. A first experimental confirmation is that SMMB-ACO and AntEpiSeeker, which both use ACO, are nevertheless intrinsically different since their overlap is only 16. Second, the IoI overlaps between ranger and AntEpiSeeker, and between ranger and SMMB-ACO, tend to feature ranger as a revelatory tool of duplicate IoIs. This remark advocates the relaxation of the significance threshold used for ranger in this feasibility study, to emphasize the potential revelatory role of ranger.

7 CONCLUSION AND FUTURE WORK

In the GWAS field, small-scale simulations reveal nothing about the effectiveness of methods on large datasets. In particular, the ratio between the number of SNPs and number of subjects observed is not comparable between simulated and real datasets.

This paper focuses on four state-of-the-art approaches designed to detect epistasis, together with the recent proposal SMMB-ACO. Our work departs from the standard framework as it reports the extensive comparative analysis of these five approaches on large-scale real data. We described an experimental protocol conceived to output comparable sets of (2-way) interactions across the approaches. We considered 23 chromosome-wide case control datasets related to Crohn's disease. We achieved 1,035 genetic analyses and observed a great heterogeneity across methods in all aspects: running times and memory requirements, numbers of interactions of interest (IoIs) output, p-value and odds ratio ranges.

This work served as a feasibility study to further extend the comparative analysis to six other real-world datasets. At this scale (10,441 chromosome-wide analyses on 161 datasets), we will be able to confirm or infirm the trends observed for the CD dataset. A still more comprehensive study would also extend the analysis to various genetic models.

Beyond the enlightening methodological comparison on real datasets, the present work allowed to cross the IoIs of several machine learning methods whose intrinsic mechanisms greatly differ. Priorizing the interactions jointly identified by at least two such methods is a defensible option to obtain a short list, when it is not affordable to test experimentally all the interactions generated. Indeed, the 56 IoIs obtained from the CD dataset allowed to infer six, four and one

networks of respective sizes 3, 4 and 7, and six of the genes involved in these networks are already known to contribute to the disease onset. Applying the revised protocol to six other genome-wide datasets will allow us to confirm whether ranger can be considered as a revelatory tool of duplicate IoIs.

REFERENCES

Aflakparast, M., Salimi, H., Gerami, A., Dubé, M.-P., Visweswaran, S., et al. (2014). Cuckoo search epistasis: a new method for exploring significant genetic interactions. *Heredity*, 112:666–764.

Ayers, K. and Cordell, H. (2010). SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genetic Epidemiology*, 34(8):879–891.

Beam, A., Motsinger-Reif, A., and Doyle, J. (2014). Bayesian neural networks for detecting epistasis in genetic association studies. *BMC Bioinformatics*, 15(1):368.

Chang, Y.-C., Wu, J.-T., Hong, M.-Y., Tung, Y.-A., Hsieh, P.-H., et al. (2018). GenEpi: gene-based epistasis discovery using machine learning. bioRXiv, doi: https://doi.org/10.1101/421719.

Chatelain, C., Durand, G., Thuillier, V., and Augé, F. (2018). Performance of epistasis detection methods in semi-simulated GWAS. *BMC bioinformatics*, 19(1):231.

Fergus, P., Montanez, C., Abdulaimma, B., Lisboa, P., and Chalmers, C. (2018). Utilising deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American women. arXiv preprint, arXiv:1801.02977.

Furlong, L. (2013). Human diseases through the lens of network biology. *Trends in Genetics*, 29:150–159.

Gao, H., Granka, J., and Feldman, M. (2010). On the classification of epistatic interactions. *Genetics*, 184(3):827–837.

Gibert, J.-M., Blanco, J., Dolezal, M., Nolte, V., Peronnet, F., and Schlötterer, C. (2017). Strong epistatic and additive effects of linked candidate SNPs for Drosophila pigmentation have implications for analysis of genome-wide association studies results. *Genome Biology*, 18:126.

Gilbert-Diamond, D. and Moore, J. (2011). Analysis of gene-gene interactions. *Current Protocols in Human Genetics*, 0 1: Unit1.14.

Gola, D., Mahachie John, J., van Steen, K., and König, I. (2016). A roadmap to multifactor dimensionality reduction methods. *Briefings in Bioinformatics*, 17(2):293–308.

Graham, D. and Xavier, R. (2013). From genetics of inflammatory bowel disease towards mechanistic insights. *Trends in Immunology*, 34:371–378.

Grange, L., Bureau, J.-F., Nikolayeva, I., Paul, R., Van Steen, K., et al. (2015). Filter-free exhaustive odds ratio-based genome-wide interaction approach pinpoints evidence for interaction in the HLA region in psoriasis. *BMC Genetics*, 16:11.

Han, B. and Chen, X.-W. (2011). bNEAT: a Bayesian network method for detecting epistatic interactions in genome-wide association studies. *BMC Genomics*, 12(Suppl.2):S9.

Han, B., Chen, X.-W., and Talebizadeh, Z. (2011). FEPI-MB: identifying SNPs-disease association using a Markov blanket-based approach. *BMC Bioinformatics*, 12(Suppl.12):S3.

Han, B., Park, M., and Chen, X.-W. (2010). A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinformatics*, 11(Suppl.3):S5.

Jiang, X., Neapolitan, R., Barmada, M., Visweswaran, S., and Cooper, G. (2010). A fast algorithm for learning epistatic genomic relationships. In *Proceedings of the Annual American Medical Informatics Association Symposium (AMIA2010)*, pages 341–345.

Jing, P. and Shen, H. (2015). MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. *Bioinformatics*, 31(5):634–641.

Khor, B., Gardet, A., and Ramnik, J. (2011). Genetics and pathogenesis of inflammatory bowel disease. *Nature*, 474(7351):307–317.

Li, Y., Xiao, X., Han, Y., Gorlova, O., Qian, D., et al. (2018). Genome-wide interaction study of smoking behavior and non-small cell lung cancer risk in Caucasian population. *Carcinogenesis*, 39(3):336–346.

Lin, H., Chen, Y., Tsai, Y., Qu, X., Tseng, T., and Park, J. (2012). TRM: a powerful two-stage machine learning approach for identifying SNP-SNP interactions. *Annals of Human Genetics*, 76(1):53–62.

McGovern, D., Kugathasan, S., and Cho, J. (2015). Genetics of inflammatory bowel diseases. *Gastroenterology*, 149(5):1163–1176.

Mullis, M., Matsui, T., Schell, R., Foree, R., and Ehrenreich, I. (2018). The complex underpinnings of genetic background effects. *Nature communications*, 9(1):3548.

Nicodemus, K., Law, A., Radulescu, E., Luna, A., Kolachana, B., et al. (2013). Biological validation of increased schizophrenia risk with NRG1, ERBB4, and AKT1 epistasis via functional neuroimaging in healthy controls. *Archives of General Psychiatry*, 67(10):991–1001.

Ritchie, M. and Van Steen, K. (2018). The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. *Annals of Translational Medicine*, 6(8):157.

Schwarz, D., König, I., and Ziegler, A. (2010). On safari to random jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics*, 26(14):1752–1758.

Shen, Y., Liu, Z., and Ott, J. (2012). Support vector machines with L1 penalty for detecting genegene interactions. *International Journal of Data Mining and Bioinformatics*, 6:463–470.

Sinoquet, C. and Niel, C. (2018). Enhancement of a stochastic Markov blanket framework with ant colony optimization, to uncover epistasis in genetic as-

sociation studies. In *Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN2018)*, pages 673–678.

Stadler, Z., Thom, P., Robson, M., Weitzel, J., Kauff, N., et al. (2010). Genome-wide association studies of cancer. *Journal of Clinical Oncology*, 28(27):4255–4267.

Sun, Y., Shang, J., Liu, J.-X., Li, S., and Zheng, C.-H. (2017). epiACO - a method for identifying epistasis based on ant colony optimization algorithm. *BioData Mining*, 10:23.

Uppu, S., Krishna, A., and Gopalan, R. (2016). Towards deep learning in genome-wide association interaction studies. In *Proceedings of the 20th Pacific Asia Conference on Information Systems (PACIS2016)*, page 20.

Urbanowicz, R., Meeker, M., LaCava, W., Olson, R., and Moore, J. (2018). Relief-based feature selection: introduction and review. *Journal of Biomedical Informatics*, 85:189–203.

Wang, Y., Liu, X., Robbins, K., and Rekaya, R. (2010). AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Research Notes*, 3:117.

Wright, M. and Ziegler, A. (2017). ranger: a fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

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

Zhang, Y. (2012). A novel Bayesian graphical model for genome-wide multi-SNP association mapping. *Genetic Epidemiology*, 36(1):36–47.

Zhang, Y. and Liu, J. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nature Genetics*, 39:1167–1173.

APPENDIX

Table 6: Parameter adjustment for the five methods.

Software	Parameter description	Value
ranger	**num.trees** number of trees	500
	mtry number of variables to possibly split at in each node, with n, the total number of variables	5/8 n
	impmeasure type of importance measure	Gini index
BEAM3	**itburn** number of iterations in burn-in phase	50
	itstat number of iterations in stationary phase	50
Ant- -EpiSeeker	**iAntCount** number of ants	1000
	iItCountLarge number of iterations for the large haplotypes	150
	iItCountSmall number of iterations for the small haplotypes	300
	iEpiModel number of SNPs in an epistatic interaction	2
	pvalue p-value threshold (after Bonferroni correction)	5×10^{-4}
	alpha weight given to pheromone deposited by ants	1
	phe initial pheromone rate for each variable	100
	rou evaporation rate in ant colony optimization	0.05
DASSO-MB	**alpha** global type I error threshold	5×10^{-4}
SMMB-ACO	$\mathbf{n_{it}}$ number of ACO iterations	360
	$\mathbf{n_{ants}}$ number of ants	20
	$\mathbf{K_a}$ size of the subset of variables sampled by each ant	160
	k size of a combination of variables sampled amongst the K above variables ($k < K$)	3
	$\boldsymbol{\alpha'}$ global type I error threshold	5×10^{-4}
	$\boldsymbol{\tau_0}$ constant to initiate pheromone rates	100
	$\boldsymbol{\rho}$ and $\boldsymbol{\lambda}$ two constants used to update pheromone rates	0.05 0.1
	$\boldsymbol{\eta}$ vector of weights, to account for prior knowledge on the variables	1
	$\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ two constants used to adjust the relative importance between pheromone rate and prior knowledge on the variables	1 1