Machine Learning and Combinatorial Optimization to Detect Gene-gene Interactions in Genome-wide Real Data: Looking Through the Prism of Four Methods and Two Protocols

Hugo Boisaubert and Christine Sinoquet(✉)

Nantes University, LS2N, UMR CNRS 6004, 2 rue de la Houssinière, Nantes, France
hugo.boisaubert@etu.univ-nantes.fr, christine.sinoquet@univ-nantes.fr

Abstract. For most genetic diseases, a wide gap exists between the heritability estimated from familial data and the heritability explained through standard genome-wide association studies. One of the incentive lines of research is epistasis - or gene-gene interaction -. However, epistasis detection poses computational challenges. This paper presents three contributions. Our first contribution aims at filling the lack of feedback on the behaviors of published methods dedicated to epistasis, when applied on real-world genetic data. We designed experiments to compare four published approaches encompassing random forests, Bayesian inference, optimization techniques and Markov blanket learning. We included in the comparison the recently developed approach SMMB-ACO (Stochastic Multiple Markov Blankets with Ant Colony Optimization). We used a published dataset related to Crohn's disease. We compared the methods in all aspects: running times and memory requirements, numbers of interactions of interest (statistically significant 2-way interactions), p-value distributions, numbers of interaction networks and structure of these networks. Our second contribution assesses whether there is an impact of feature selection, performed upstream epistasis detection, on the previous statistics and distributions. Our third contribution consists in the characterization of SMMB-ACO's behavior on large-scale real data. We report a great heterogeneity across methods, in all aspects, and highlight weaknesses and strengths for these approaches. Moreover, we conclude that in the case of the Crohn's disease dataset, feature selection implemented through a random forest-based technique does not allow to increase the proportion of interactions of interest in the outputs.

Keywords: Epistasis detection · Machine learning · Markov blanket · Markov chain Monte Carlo · Random forest · Ant colony optimization · Feature selection · Data-dimension reduction · Extensive comparative analysis

HB and CS are the two joint first co-authors.

A. Roque et al. (Eds.): BIOSTEC 2019, CCIS 1211, pp. 139–169, 2020.
https://doi.org/10.1007/978-3-030-46970-2_8

1 Introduction

Within two decades, genome-wide association studies (GWASs) have introduced a new paradigm in the field of complex disease genetics. GWASs' purpose is to detect statistical dependences, also called associations, that exist between genetic variants and some phenotype of interest, in a population under study. For example, in case-control studies, the phenotype of interest is the affected/unaffected status. Typically, a GWAS considers between a few thousand to ten thousand individuals in a population, for which high-throughput technologies allow to measure DNA variation at characterized *loci* distributed over the whole genome. Single nucleotide polymorphism (SNP) is a type of DNA variation widely-used in GWASs. Hereafter, we will only consider SNP-based association studies. Depending on the genotyping microarray used, GWASs analyze between a few hundred thousand to a few million SNPs. Standard GWASs test each of the SNPs one at a time, to identify a difference between case and control cohorts.

GWASs have allowed a greater understanding of the genetic architecture underlying complex phenotypes [38]. By fostering prevention and design of more efficient drug therapies depending on the genetic profiles of patients, GWASs have contributed to pave the way to personalized medicine. However, despite GWASs' successes, a wide gap exists between the heritability estimated from familial data and the heritability explained by genetic variants *via* standard GWASs, for most phenotypes investigated so far. To close this so-called 'missing heritability' gap [37], complementary venues of research actively investigate alternative heritable components of complex phenotypes. These alternatives encompass additivity of small effects from myriads of common variants, rare variants, structural variants, epigenetics, gene-environment interactions and genetic interactions [44].

This paper focuses on computational approaches designed to detect genetic interactions, also named *epistasis*. Nowadays, the term "epistasis" is widely used to refer to the situation in which genes interact together to determine some phenotype, whereas each of them alone is not influential on this phenotype: the contribution of one gene to a phenotype depends on the genetic profile of the organism under study. To note, the latter phenotype is not directly observed in case-control studies, in which a physiological quantitative phenotype underlies the unaffected/affected phenotypic status expressed. Epistasis can be seen as the result of physical interactions among biomolecules involved in biochemical pathways and gene regulatory networks, in an organism [29].

To illuminate where part of the missing heritability lies, the role of gene interactions is substanciated by a persuasive body of evidence: biomolecular interactions are omnipresent in gene regulation, signal transduction, biochemical networks and physiological pathways [9,12]. These interactions play a key role in transcriptional and post-translational regulations, interplay between proteins as well as intercellular signaling. Biological evidence for epistasis has been documented in the literature (*e.g.*, [10,11,18,27]). In regard of the ubiquitous character of gene-gene interactions, the relatively limited number of findings published is arguably explained by the computational issue raised by epistasis detection.

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 AISs (GWAISs). This paper contributes to fill this lack. Besides, we recently extended our work to assess the impact of feature selection, when applied upstream epistasis detection. Another strong motivation for our work was to analyze how SMMB-ACO [32], 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 overview of the recent state-of-the-art of the domain. Section 3 provides the motivations for our study and sketches our main contributions. 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 two experimental protocols involved, the real-world datasets analyzed, the implementation and parameter adjustment of the five methods. The experimental results, discussion and feedback gained are presented in the last section.

2 A Brief State-of-the-Art in the Computational Landscape of Gene-gene Interactions

This section provides an overview of the various categories of methods designed to address epistasis detection issues.

2.1 Exhaustive Approaches

The detection of gene-gene interactions is no easy task, especially for large datasets. High level interactions, which involve more than two *loci*, pose a formidable computational challenge. For instance, the number of potential pairwise interactions in a dataset of 500,000 SNPs amounts to 12.5×10^{11}; in the same dataset, the number of potential 3-way interactions rises to 2.08×10^{16}, Hereafter, we will describe the main classes of methods and provide an illustration for each, with a highlight on the scalabilities of the methods cited as illustrations.

In the class of **statistical approaches**, **linear generalized regression** (LGR) offers a framework to model the relationship between an outcome variable y and multiple interacting predictors $x_1, x_2, ..., x_q$ (continuous or categorical), such as in $f(y) \sim \beta_0 + \beta_1\ x_1 + \beta_2\ x_2 + \beta_{12}\ x_1 x_2$, with $q = 2$. In this framework, two ingredients allow to escape from the pure linear scheme ($y \sim \beta_0 + \beta_1\ x_1 + \beta_2\ x_2$), in the case of two predictors). On the one hand, interaction terms β_{ij} capture potential interplay between predictors. On the other hand, the *link* function f is used to transform the outcome y, to match the real distribution of y. Obviously, LGR cannot be used straightforwardly to analyze data on the genome scale: the exhaustive enumeration and test of potential q-way interactions is prohibitive, and this task should be performed for q comprised between 2 and r, where r is an upper bound arbitrarily set by the user. Furthermore, identifying an appropriate link function f may not be trivial. Nevertheless, **logistic regression** (LR) is a widely-used specific case of LGR in which the link function is known, to model a binary outcome: in case control studies, with p representing the probability to be affected by the pathology of interest, the 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.

Penalized regression (PR) implemented through Lasso, Ridge or Elastic Net regression can be used for the purpose of epistasis detection [2]. The computational burden of these methods is particularly heavy. The approach described in [5] attempts to palliate this issue through a two-stage procedure. First, pairwise interactions are searched for within each gene, using Randomized Lasso and penalized Logistic Regression (RLLR). Second, pairwise interactions across genes are assessed considering the SNPs obtained in the first stage. RLLR is again used in this second stage. In [33], interactions are searched for each pair of genes. A Group Lasso approach is employed, in which groups comprise either the SNPs of a given gene, or interaction terms relative to a given pair of genes. Though such approaches seem appealing to capture cross-gene epistasis, they each feature a major drawback. In [5], the biological motivation for the data dimension reduction performed *via* the first stage is questionable since 2-way interactions within genes are not necessarily connected to cross-gene interactions. On the other hand, the approach in [33] could only be run on a pre-selected set of a few dozen genes, for each of three real GWAS datasets. These genes were pre-selected based on an univariate analysis, which introduces a bias.

A step further, model-free data mining methods in the line of **multifactor-dimensionality reduction** (MDR) categorize the observed genotypes into high-risk and low-risk groups, for each q-way potential interaction [13]. Since enumerating all potential q-way interactions is required, MDR-based approaches fail to handle large-scale data. An exception is the case when GPU calculation is used [43].

2.2 Dimension Reduction Upstream of Epistasis Detection

A direct way to reduce the search space is to decrease the dataset size. **Filtering based on extrinsic biological knowledge** is expected to yield meaningful and biologically relevant results. However, exploiting additional knowledge such as protein-protein interaction networks or pathways is questionable. Online databases are incomplete and our understanding of biological pathways is limited. Therefore, relying on such knowledge for data dimension purpose would result in a biased analysis in the majority of cases.

In the category of machine learning and data mining approaches, feature selection techniques rely on properties intrinsic to the data, to select SNPs potentially relevant for epistasis detection.

A number of variants were proposed around Relief [36]. The first step in **Relief-based approaches** (RBAs) is to compute pairwise (genetical) similarities between subjects. A nearest neighbor technique is further applied, to assess importances for SNPs with respect to the phenotype of interest. Basically, the method identifies SNPs not sharing the same values between a subject and its nearest neighbors. If this situation arises when the subject and its nearest neighbors neither share the same phenotype, the SNPs' importances are increased; otherwise, the importances are decreased. This step is only repeated over a user-defined number of subjects, which nonetheless requires the costly computation of pairwise similarities. Moreover, Relief-based approaches are prone to pre-select SNPs marginally associated with the phenotype.

Random forest (RF) approaches implement high-dimensional non-parametric predictive models relying on ensemble features. In RFs, bootstrap aggregating [4] allows to convert a collection of weak learners (decision trees in this case) into a strong learner. The decision trees (classification trees for a categorical outcome, regression trees for a continuous outcome) are grown recursively from bootstrap samples of observations. At each node in each tree, the observations (*e.g.*, individuals) that have percolated down this node are splitted relying on an optimal cut-point. A cut-point is a pair involving one of the available variables (*e.g.*, SNPs) and a value in the variable's domain. Over all available variables, the optimal cut-point best discriminates the observations with respect to the outcome of interest (*e.g.*, phenotype). In RFs, the optimal cut-point is determined using a random subset of the initial variables. RFs produce a ranking of the variables, by decreasing importance measure. This measure quantifies the impact of a variable in predicting the outcome and thus potentially reflects a causal effect. RF-based approaches were shown efficient in ranking simulated disease-associated SNPs, to detect gene-gene interactions [24,25]. Computational cost and memory inefficiency were severe impediments to RF learning in high-dimensional settings. In this respect, the advances reported in [30] and [40] render RF-based feature selection practicable for epistasis detection at large scale.

In association studies, feature selection yields a ranking for the SNPs in the initial available dataset. A procedure is required downstream such methods as Relief-based approaches and Random Forests, to generate gene-gene interactions from the top ranking SNPs. Such procedure may boil down to assessing potential interactions through statistical tests. In contrast, a specific approach designed to detect epistasis may be used. We explored both modalities in the work reported here.

2.3 Sampling from Probability Distributions

The popular BEAM algorithm (Bayesian Epistasis Association Mapping) [42] relies on a **Markov Chain Monte Carlo** (MCMC) process to test iteratively each SNP, conditional on the current status of other SNPs. For each SNP, the algorithm outputs its posterior probability of association with disease. BEAM then partitions the SNPs into three categories. One category contains SNPs with no impact on the disease. A second category contains SNPs that contribute independently to the disease. The third category highlights SNPs assumed to jointly influence the disease given particular variant combinations of some other SNPs. BEAM was reported to handle datasets with half a million of SNPs, at the cost of high running times (up to a week and even more).

2.4 Machine Learning Techniques

To detect epistasis, machine learning approaches represent appealing alternatives to parametric statistical methods. Such approaches build non-parametric models to compile information further used for gene-gene interaction detection.

Standard supervised machine learning and data mining techniques can be employed directly for the purpose of epistasis detection. **Support vector machines** (SVMs) separate interacting and non-interacting groups of SNPs using a hyperplane in multi-dimensional space. The work in [31] reports an SVM-based study of 2-way interactions conducted at the genome scale. On the other hand, **artificial neural networks** (ANNs) allow to model non-linear feature interactions. To this aim, non-linear activation functions are used, in conjunction with a sufficient number of hidden layers. Advanced stochastic gradient descent techniques brought a remarkable breakthrough in training feedforward networks with many hidden layers, thereby paving the way to deep neural networks (DNNs). However, so far, DNNs were confined to process small datasets. In [35], a DNN was learned from small datasets (no more than 1,600 subjects, a few dozen SNPs). The DNN used in [8] was learned from around 1,500 subjects and 5,000 SNPs, downstream a filtering stage consisting in logistic regression.

Bayesian Networks (BNs) allow to model patterns of probabilistic dependences between variables represented as nodes in a directed acyclic graph. In the context of epistasis detection, BNs offer an incentive framework to discover the best scoring graph structure connecting SNPs to the disease variable. In [15], a branch and bound heuristic allowed to handle a relatively limited dataset, a published AMD (Age Macular Degenerated) dataset (150 individuals, around 110,000 SNPs). In [19], a greedy search implements a forward phase consisting in edge addition followed by a backward phase orchestrating edge removal. The tractability issue is addressed by starting the greedy search with one pair of interacting SNPs which are each influential on the disease status. This approach is therefore limited to the detection of a specific case of epistasis, named *embedded* epistasis.

In Bayesian networks, the concept of Markov blanket [28] offers an appealing line of investigation for epistasis detection. Given 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 $MB(T)$ probabilistically independent of T, conditional on $MB(T)$. Otherwise stated, $MB(T)$ is theoretically the optimal set of variables to predict the value of T [22]. In the GAIS context, the purpose is to build a MB for the variable representing the affected/unaffected status. Feature subset selection stated as **Markov blanket** learning was thus explored and produced FEPI-MB (Fast epistatic interactions detection using Markov blanket) [16] and DASSO-MB (Detection of ASSOciations using Markov Blanket) [17]. Both approaches were able to process the above mentioned AMD dataset.

2.5 Combinatorial Optimization Approaches

In the **optimization** field, techniques dedicated to AISs browse through the search space of solutions (combinations of potentially interacting SNPs). Various heuristics were proposed, to identify the more relevant combinations of SNPs. In the line of genetic algorithms, the approach described in [1] relies on an **evolutionary-based** heuristic. This method allowed to process around 1,400 subjects and 300,000 SNPs. **Ant colony optimization** (ACO) was exploited by several proposals such as AntEpiSeeker [39], MACOED [20] and EpiACO [34]. The widely cited reference AntEpiSeeker consists in the straightforward adaptation of classical ACO to epistasis detection and is tractable on the genome scale. MACOED, a multi-objective approach employing the Akaike information criterion (AIC) score and a BN-based score, was able to process 1,411 individuals described by 312,316 SNPs (late-onset Alzheimer's disease (LOAD) dataset). Since MACOED needs unaffordable running times to obtain results, the analysis focused on separate chromosome datasets. The unique objective function used in EpiACO combines a mutual information measure with a BN-based score. EpiACO was able to handle the above cited AMD dataset.

3 Motivations and Contributions of Our Study

The critical analysis of the specialized literature led us to draw several remarks about the evaluation and comparison of computational approaches dedicated to epistasis detection.

First, evaluating the effectiveness of a method requires the generation of multiple synthetic datasets, for instance 100, under some controlled assumption. The points to control are the number of interacting SNPs and the strength of the joint effect of simulated influential SNPs on the disease status. When evaluating a non-deterministic method, we have to compute a performance for each synthetic dataset, which is a function of the numbers of true positives, false positives and false negatives recorded through multiple runs of the same method on the same dataset (*e.g.,* power, F-measure). Unfortunately, tractability reasons compell methods' authors to generate simulated datasets whose size remains compatible with the computing and storage resources available to these authors.

Thus, an overwhelming majority of publications rely on synthetic datasets that describe 100 SNPs, a few thousand SNPs at best, for a few thousand subjects. A notable exception is the study reported in [6]. This practical limitation renders questionable the significance of the evaluation and comparison of methods on such small simulated datasets: in no way do such experimentations reflect real-world GWAIS analyses.

Besides, as regards real-world GWAIS analyses, the overwhelming trend in publications is to submit a unique genome-wide dataset to the proposed method. No comparison is performed with other methods. Again, the reason lies in tractability. Comparing several methods at this scale requires authors to adjust a list of parameters for each method. Ideally, adjusting parameters for any approach resorting to supervised machine learning would need running ten times a GWAIS (in a 10-fold cross-validation procedure) in each of the parameter instantiations of a parameter grid. Optimizing the parameters for any heuristics in the field of combinatorial optimization also requires unaffordable running times in general.

A third remark is that running times are but exceptionally reported in publications describing a novel method. If so, they are only reported for small simulated datasets. Instead, assessing orders of magnitude for running times across methods would be far more informative (*e.g.,* for practitioners), if the methods were applied on datasets of realistic scale.

A fourth remark arises from the observed lack of comparative studies focused on epistasis detection methods applied at the genome scale: it is questionable whether the lists of gene-gene interactions output by these methods overlap, and if so, by which overlapping rate.

The works presented in this paper were designed with the four previous points in mind and with the five following related objectives: (i) perform an unprecedented comparative analysis on real-world GWAS data, for a selection of methods dedicated to GWAIS, (ii) assess the respective requirements of these methods, in terms of running times and memory resources, (iii) characterize the solutions respectively output by these methods, (iv) examine the pairwise intersections of solutions output by these methods and possible intersections between more than two methods. A first presentation of these works was published in [3], of which the present paper is an extended version. In addition to the extensive comparative study reported in [3], we have started afresh a novel comparative study. This time, we have run each method compared, after a common feature selection procedure was applied on each of the chromosome-wide datasets considered. The additional three contributions highlighted in this extended version are the following: (v) include additional criteria to characterize the solutions output by the methods compared, namely around the connectivity between SNPs involved in multiple interactions, (vi) characterize the solutions respectively output by the methods in previous and novel experimental conditions, that is without and with the filtering stage, (vii) provide an illustration focused on a network of interactions, and give corresponding biological insights.

4 The Approaches Selected for the Extensive Comparison

To analyze the epistasis detection task in real-world conditions, we selected five approaches illustrating various techniques.

Data dimension reduction upstream of epistasis detection is illustrated through ranger [40] coupled with logistic regression. The reference ranger software is a fast implementation for random forest-based feature selection; it was specifically designed to cope with high-dimensional data. A further argument for including ranger in our comparative study is that, so far, any novel method proposed was generally compared to Random Jungle [30], the precursor of ranger.

BEAM3 [41], the successor of the reference software BEAM [42], implements Bayesian inference *via* the sampling of probability distributions. An MCMC simulation allows to assign a statistical significance to each SNP, thus avoiding expensive permutation-based tests.

In the field of machine learning, feature subset selection stated as **Markov blanket** learning is implemented through DASSO-MB [17]. For didactical reasons, the sketch of DASSO-MB will be provided together with that of the fifth approach selected.

The reference method AntEpiSeeker [39] was incorporated in our study to represent combinatorial optimization heuristics. The ant colony optimization (ACO) technique behind AntEpiSeeker is as follows: in each iteration, the ants each sample a SNP set of user-defined size from the initial dataset, based on a probability distribution $\mathbb{P}$; each ant then assesses the dependence of the SNP set S sampled with the affected/unaffected status through a statistical test (χ^2). The SNP sets showing the highest dependence scores are kept. The pheromone level of each SNP s thus highlighted is computed based on the dependence of the set S that contains s. A standard ACO scheme uses the pheromone levels to update the probability distribution $\mathbb{P}$ of each SNP. At the end of a user-defined number of ACO iterations, a pre-specified number of best SNP sets is available, together with a list $\mathcal{L}$ of SNPs characterized by the highest pheromone levels. The final step of AntEpiSeeker then examines each best SNP set S as follows: given q, the size of the epistatic interactions to be uncovered, each subset of S of size q is kept as an epistatic interaction, provided all its SNPs belong to $\mathcal{L}$. It may happen that two interactions overlap, in which case the one with the smallest p-value is kept.

The selection of DASSO-MB and AntEpiSeeker was no innocent choice. Indeed, the fifth approach included in our comparative study, SMMB-ACO (Stochastic Multiple Markov Blankets with Ant Colony Optimization) [32], is a recent method that borrows from Markov blanket learning and ant colony optimization techniques. In the previous paragraph, we have explained the ACO mechanism behind AntEpiSeeker.

We now highlight the differences between the deterministic DASSO-MB approach and the stochastic and ensemble feature-based SMMB-ACO approach. DASSO-MB chains a forward phase and a backward phase. Starting from an empty Markov blanket (MB), the forward phase adds a SNP to the growing MB based on two conditions: (i) the dependence between this SNP and the target variable (affected/unaffected status in our case) is the highest, conditional on the MB, when compared to all other SNPs; (ii) this dependence is statistically significant. The backward phase successively examines the SNPs belonging to the current MB; a SNP is removed from the MB based on (statistically significant) conditional independence. To discard false positives as soon as possible, DASSO-MB triggers a full backward phase after a SNP has been added during the forward phase. In each iteration of SMMB-ACO, the ants each learn a suboptimal Markov blanket from a subset of SNPs sampled from the initial set. The MB learning scheme in SMMB-ACO relies on a forward phase intertwined with backward phases. In this respect, MB learning in SMMB-ACO is quite similar to that in DASSO-MB. However, SMMB-ACO and DASSO-MB's forward steps fundamentally differ: DASSO-MB attempts to add the SNP showing the strongest dependence with the target variable; in contrast, SMMB-ACO seeks to stochastically add a group of SNPs highly dependent with the target variable. The stochastic feature of SMMB-ACO is implemented through the sampling of groups of SNPs, and relies on a probability distribution $\mathbb{P}$ updated based on pheromone levels. To note, a specific operating mode may be specified for SMMB-ACO, to handle high-dimensional data: a two-pass procedure is then triggered. DASSO-MB and SMMB-ACO are sketched and commented in Fig. 1.

5 Extensive Experimentation Framework

This section starts with the presentation of the experimental road map designed. Second, the real-world datasets used are briefly depicted. Finally, the section focuses on implementation aspects, including parameter adjustment of the approaches compared.

5.1 Experimental Road Map

In the so-called additive model, the SNPs are coded with 0, 1 and 2, which respectively denote major homozygous, heterozygous and minor homozygous. The allele with minor frequency is the disease susceptibility allele. The notion of **interaction of interest** (IoI) is central to our study. An IoI is a 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, whereas no significant p-value is obtained for the regression of the target variable on each individual SNP. Our experimental protocol was two-fold.

To use the random forest-based approach ranger for epistatic detection purpose, we generated 2-way interactions from the top most important SNPs returned by ranger. For tractability reasons, the 2-way interaction candidates were generated from the 20 most important SNPs selected. Thus, C_{20}^{2} 2-way interactions were submitted to logistic regression. To characterize IoIs, the significance threshold 5×10^{-4} was chosen. In our comparative study, we put all approaches on an equitable basis. Therefore, we filtered the interactions obtained from BEAM3, AntEpiSeeker, and DASSO-MB as well as the results obtained

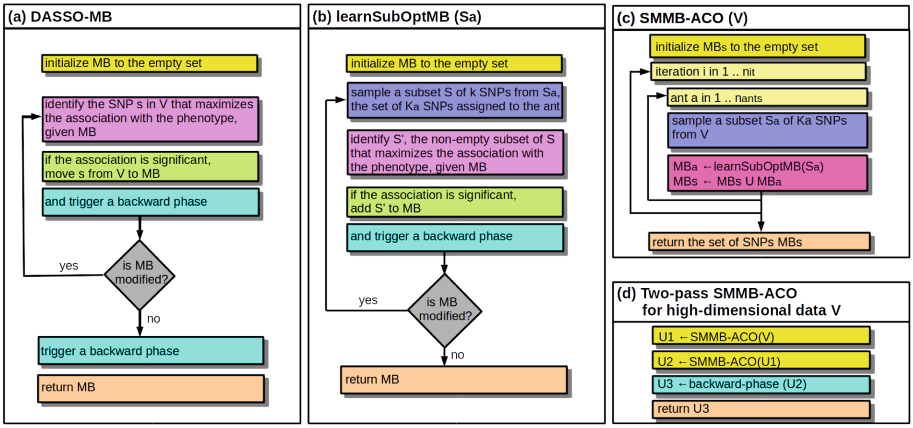

(a) DASSO-MB
initialize MB to the empty set
identify the SNP s in V that maximizes the association with the phenotype, given MB
if the association is significant, move s from V to MB
and trigger a backward phase
is MB modified?
yes
no
trigger a backward phase
return MB
(b) learnSubOptMB (Sa)
initialize MB to the empty set
sample a subset S of k SNPs from Sa, the set of Ka SNPs assigned to the ant
identify S', the non-empty subset of S that maximizes the association with the phenotype, given MB
if the association is significant, add S' to MB
and trigger a backward phase
is MB modified?
yes
no
return MB
(c) SMMB-ACO (V)
initialize MBs to the empty set
iteration i in 1 .. nit
ant a in 1 .. nants
sample a subset Sa of Ka SNPs from V
MBa ←learnSubOptMB(Sa)
MBs ← MBs U MBa
return the set of SNPs MBs
(d) Two-pass SMMB-ACO for high-dimensional data V
U1 ←SMMB-ACO(V)
U2 ←SMMB-ACO(U1)
U3 ←backward-phase (U2)
return U3

Fig. 1. Sketches of the DASSO-MB and SMMB-ACO algorithms. (Figure published in [3]). (a) DASSO-MB. (b) SMMB-ACO stochastic procedure to learn a suboptimal Markov blanket. (c) SMMB-ACO top-level algorithm. (d) Two-pass SMMB-ACO procedure adapted to high-dimensional data. MB: Markov blanket. V is the initial set of SNPs. (a) DASSO-MB adds SNPs one at a time, which hinders the epistasis detection task: since the dependence test achieved at first iteration is conditioned on the empty Markov blanket, this test is indeed a marginal test of dependence; therefore, a SNP marginally dependent with the target variable is added at the outset, which skews the whole MB learning. (b) SMMB-ACO addresses this issue by adding groups of SNPs. For this purpose, each forward step starts with the sampling of a set S of k SNPs, from the subset S_a of size K_a that was assigned to the ant in charge of the suboptimal MB learning. For each non-empty subset S' of S, a score is computed, which measures the association strength between S' and the target variable, conditional on the MB under construction. The subset S' with the highest association score is added to the MB if the association is statistically significant. (c) SMMB-ACO returns the set of SNPs obtained as the union of all suboptimal MBs generated throughout all iterations. (d) In the two-pass procedure adapted to high-dimensional data, SMMB-ACO is first applied on the initial set of SNPs V, which produces the set of SNPs U_1. In the second pass, SMMB-ACO is applied on U_1. This time, the resulting set U_2 is submitted to a backward phase, to yield U_3, a set of SNPs.

from the modified post-processing phase of SMMB-ACO (Details about this modification will be provided in Sect. 5.3). This filtering stage kept the IoIs with significance threshold 5×10^{-2}. The use of two significance thresholds will be substantially justified further (see Subsect. 5.3). For now, the reader needs only keep in mind that AntEpiSeeker, DASSO-MB and SMMB-AC0 *already intrinsically* rely on a significance threshold.

In addition to the experimental protocol just described, we designed afresh novel experimentations. This time, a feature selection procedure was first run on the datasets considered. The second experimental protocol started with feature selection carried out through ranger. For each of the 50 runs of ranger on a given chromosome dataset, the 5,000 SNPs with the highest importances were memorized. The set of 25,000 SNPs thus obtained was then processed to discard duplicate SNPs. The first protocol described in previous paragraph was then applied on the reduced set of SNPs. Lessons learned from our first experimentations [3] motivated the modification of the significance threshold used to identify IoIs with ranger (see Subsect. 5.3). From now on, we will use the symbol "*" to refer to the protocol with feature selection. For instance, the use of BEAM3 in the two frameworks will be referred to by BEAM3 and BEAM3*. A recapitulation is provided in Fig. 2. Given the poor results of DASSO-MB obtained when applying the first protocol, we discarded this method from the second protocol.

To be clear, in the second protocol, ranger* stands for the following process: (i) off-line feature selection by ranger applied 50 times on a chromosome-wide dataset, to provide 50 × 5,000 SNPs from which the resulting set of n_{fs} SNPs with no duplicates is kept, (ii) run of ranger on the reduced dataset of n_{fs} SNPs thus obtained, (iii) generation of $C^2_{n_r}$ 2-way interactions from the n_r SNPs with highest importances output through the second run of ranger, (iv) identification of the IoIs in the $C^2_{n_r}$ 2-way interactions using logistic regression. We highlight here that we set n_r to 20, for consistency with the first protocol.

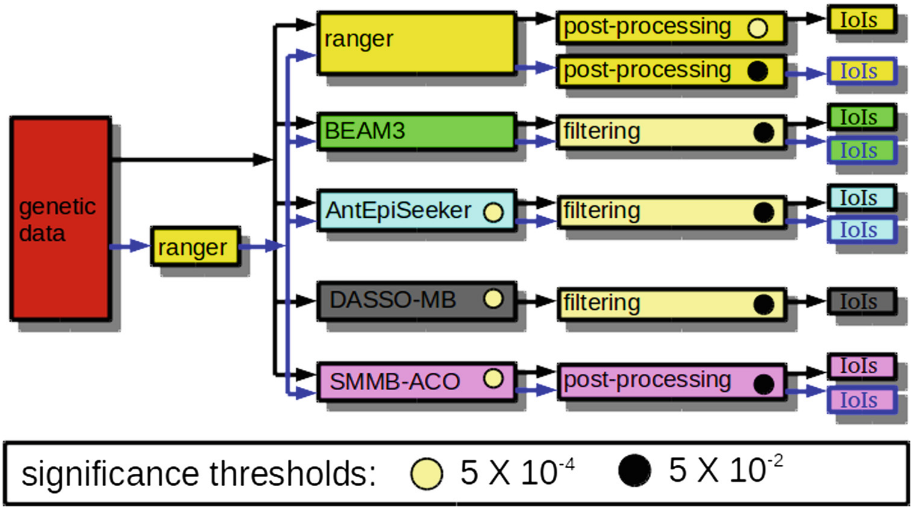


Fig. 2. Flow diagram for the two extensive comparative analyses performed. The data flows relative to first and second experimental protocols respectively appear in black and blue arrows. (Color figure online)

Table 1. Implementations for the five software programs used in the two comparative studies. (Table published in [3]).

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

5.2 Real-World Datasets

We applied the two experimental protocols above described to a Crohn's disease (CD) dataset. This data was made available by the Wellcome Trust Case Control Consortium (WTCCC, https://www.wtccc.org.uk/). The choice of this dataset was motivated by the insights generated by advancements in human genetics into the mechanisms driving inflammatory conditions of the colon and small intestine. Notably, major pathways involved in Crohn's disease and ulcerative colitis have emerged from standard single-SNP GWASs [14]. We relied on the cohort of cases affected by CD and two cohorts of unaffected (controls) provided by the WTCCC, to generate 23 datasets related to the 23 human chromosomes. We followed the quality control protocol specified by the WTCCC. In particular, we excluded subjects having more than 5% of missing data together with SNPs having more than 1% of missing data and excessive Hardy-Weinberg disequilibrium (5.7×10^{-7} threshold). After quality control, we obtained a population of $4,686$ subjects composed of $1,748$ affected and $2,938$ unaffected. 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. The average number of SNPs per chromosome is $20,236$; the minimum and maximum numbers are $5,707$ and $38,730$, respectively.

5.3 Implementation of the Two Comparative Analyses

This subsection first focuses on the intensive computing aspects. Then it describes parameter adjustment for the five methods involved in the experimentations.

Intensive Computing. Except for DASSO-MB, all software programs are available on the Internet (Table 1); they are coded in C++. We recoded DASSO-MB in C++. As mentioned in Sect. 5.1, to include SMMB-ACO in our experimental protocol, we modified the post-processing phase of the native SMMB-ACO algorithm [32]. The native algorithm outputs as an interaction any suboptimal Markov blanket generated (*via* procedure learnSubOPtMB, see Fig. 1 (b)) if all its SNPs belong to the set U_3 (see Fig. 1 (d)) obtained as the final result of the two-pass modality. The adapted post-processing phase of SMMB-ACO consists in the generation of interactions of interest (IoIs), as defined in Subsect. 5.1, from the set U_3.

DASSO-MB is the only deterministic approach of our selection of methods. Each other (stochastic) method was run several times on each dataset. In the first experimental protocol, this number was set to 10 for tractability reasons. For a fair comparison, this number was kept to 10 in the second protocol involving data dimension reduction.

The extensiveness of our two comparative studies required intensive computing resources from a Tier 2 data centre (Intel 2630v4, 2 × 10 cores 2,2 Ghz, 20 × 6 GB). On the one hand, we benefitted from the OpenMP intrinsical parallelization of the C++ implementations of ranger, BEAM3 and SMMB-ACO. In addition, we exploited data-driven parallelization to run each stochastic method 10 times on each dataset. To cope with running time heterogeneity across the methods, together with the occurrence of memory shortages, we had to balance the workload distribution between two strategies. One strategy was to sequentially process the 23 chromosome datasets for one method on one node, and to repeat this task 9 times on other nodes. The alternative strategy was to process a single chromosome dataset 10 times for one method on one node, and to repeat this task for the remaining chromosomes (on other nodes). 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 or shortage event occurred in a node, depending on the degree of completion of the task, we either switched to the first strategy with higher time limit or to a chromosome by chromosome management.

The first batch of experimentations involved 943 chromosome-wide association studies. The second batch involved as many analyses, together with the prior feature selection performed chromosome by chromosome. This pre-processing step involved 50 runs of ranger for each of the 23 chromosome datasets. In total, we run 3,036 chromosome-wide analyses.

Finally, it is important to note that generating all 2-way interactions from a set of t SNPs (as is done when ranger is used for epistasis detection) may be computationally expensive. For example, the exhaustive generation and assessment of dependence with the disease status through logistic regression takes around 30 h for $t = 20$ SNPs.

Parameter Adjustment. In the machine learning and combinatorial optimization fields, adjusting methods' parameters is a recurring issue. Table 7 in Appendix recapitulates the main parameters of the software programs used in the two batches of experimentations.

The software program ranger is a fast implementation of the random forest technique, to cope with high-dimensional data. We therefore left unchanged the default value of 500 for the number of trees in the forest. In a preliminary study (results not shown), we tried various values of mtry between $\sqrt{n}$ and n, the total number of SNPs. On the datasets considered, the optimal value was shown to be $\frac{5}{8}n$. This setting was adopted for ranger used in the first protocol, ranger employed for feature selection in the second protocol as well as for ranger run downstream feature selection in the second protocol. Importantly, a greater computational effort was devoted to the feature selection task carried out by ranger in the second protocol: the number of trees was set to $1,000$ instead of 500.

To attempt to diminish the large number of interactions output by AntEpiSeeker, we conducted a preliminary study. The product "number of ACO iterations × number of ants" impacts this number of interactions. In the preliminary study, the number of ACO iterations was kept to AntEpiSeeker's default value (450); the number of ants was varied between 500 and 5,000 (step 500).

Using 1,000 ants still allows to control the number of interactions output below 15,000, while still guaranteeing a coverage of 10 for each SNP in the largest chromosome-wide dataset. For a fair comparison, this parameter setting was kept in the second experimental protocol.

To set the numbers of iterations of the burn-in and stationary phases of BEAM3, we followed the recommendation of its author. This setting was made in adequacy with the dimension of the data handled in the first protocol. For an unbiased comparison, this setting was also applied to the second protocol.

The first reflex would be to set the product $n_{it} \times n_{ants}$ (number of ACO iterations × number of ants) in SMMB-ACO to the value chosen for AntEpiSeeker. However, two points must be emphasized. On the one hand, 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. On the other hand, an iteration in AntEpiSeeker is far less complex than an iteration in SMMB-ACO: in AntEpiSeeker, each ant samples 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), to expect that each SNP of the initial dataset would be drawn a sufficient number of times during a single run of SMMB-ACO. The parameter setting adopted $(n_{it}, n_{ants}, K_a) = (360, 20, 160)$ guarantees in theory 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 required for the statistical independence tests triggered by AntEpiSeeker and the statistical conditional independence tests run in DASSO-MB and SMMB-ACO. We set a common threshold value of 5×10^{-4} for these three methods. Consistently, in the first protocol implemented in [3], we fixed the same threshold for the logistic regression used downstream ranger execution. On the other hand, we recall that logistic regression is used to identify IoIs from the results output by BEAM3, DASSO-MB, AntEpiSeeker and SMMB-ACO. A less stringent threshold of 5×10^{-2} was chosen for this purpose. Importantly, one of the conclusions of the work reported in [3] highlighted the necessity to relax the threshold of 5×10^{-4} used downstream a run of ranger, to identify more IoIs. Therefore, in the second protocol, the second pass of ranger is followed by IoI identification at 5×10^{-2} significance threshold. This information about the various thresholds used in the two protocols is provided in Fig. 2.

6 Results and Discussion

We first compare the five approaches with respect to running times and memory occupancies. Second, we compare the numbers of interactions of interest (IoIs) identified by these approaches, and we thoroughly analyze the distributions of p-values obtained. Third, we provide insights regarding whether some IoIs were jointly detected by several approaches. A fourth subsection is devoted to the analysis of the networks of IoIs that could be identified by the methods. A fifth subsection provides an illustration focused on a network of 19 IoIs detected *via* SMMB-ACO, and gives corresponding biological insights. This section ends with a discussion.

6.1 Computational Running Times and Memory Occupancies

Table 2 highlights a great heterogeneity between the methods compared.

We first comment the complexities observed when chromosome-wide datasets are input to the methods (first protocol). DASSO-MB is both much faster and far less greedy in memory than its competitors. A salient feature of AntEpiSeeker is that it shows low running times across all chromosomes. The software program ranger is fast (around 14 mn for the 10 runs on a chromosome). However, this quickness is hindered by the exhaustive generation of C_{20}^2 2-way interactions further submitted to logistic regression (between 40 and 80 mn cumulated over the 10 runs on a chromosome). The trends observed for BEAM3 and SMMB-ACO are respectively extremely disparate across the datasets. When processing the largest chromosomes with BEAM3, we first experienced timeouts. We therefore specified the highest timeouts possible (30 days), with the consequence of longer waiting times in job queues. Indeed, in BEAM3, the cumulative running time may exceed 8 days for the largest chromosomes, which is over the "medium" timeout of 5 days. Besides prohibitive running times, BEAM3 si also the approach most greedy in memory on average for the datasets considered. Nonetheless, BEAM3 never ran out of memory.

As regards SMMB-ACO, a great heterogeneity in running times was also observed across the chromosome-wide datasets. SMMB-ACO is faster than BEAM3. However, the stochastic feature of SMMB-ACO translates into an extreme heterogeneity of memory occupancies across the chromosomes, even across the 10 executions on a given chromosome. In particular, we observed memory shortages, even for short chromosomes (for a limitation of 120 GB per node). Because of these shortages, for around the third of the datasets, we had to launch additional runs (up to 5), to obtain the 10 runs required by our protocol. Nevertheless, the processing of all chromosomes by SMMB-ACO remains feasible within 5 days, on 10 nodes.

We recall that a crucial step in the second protocol is the dimensionality reduction task: an off-line feature selection driven by ranger is applied 50 times on each chromosome-wide dataset, to provide $50 \times 5,000$ SNPs from which the resulting set of n_{fs} SNPs with no duplicates is kept. We emphasize that for each chromosome, the $5,000$ top ranked SNPs were remarkably well conserved throughout the 50 executions of ranger. Namely, n_{fs} varied between 5,000 and 5,150.

Table 2. Trends observed for the running times and memory occupancies for the methods compared, in the two protocols. Otherwise stated, the average running time indicated is computed over the 23 chromosome-wide datasets (it measures the average for the running time cumulated over 10 executions, for a chromosome-wide dataset). FS: off-line feature selection driven by ranger.

Method	First protocol		Second protocol	
	Average running time	Memory occupancy	Average running time	Memory occupancy
ranger	1.3 h ± 27 mn	2 GB ± 0.6	FS: 11 h ± 5.5 19.6 mn ± 7	4 GB 120 MB ± 20
BEAM3	Chr7 to Chr23: 54 s ± 66 Chr6: 22.4 h Chr1 to Chr5: above 8 days	79 GB ± 46	Chr7 to Chr23: 24.1 h ± 0.1 Chr6: 38.9 h Chr1 to Chr5: below 5 days	800 MB ± 200
AntEpiSeeker	16 mn ± 3	0.5 GB ± 0.2	15 mn ± 4	70 MB ± 5
DASSO-MB	82 s ± 22	1.5 GB ± 0.7	—	—
SMMB-ACO	Chr7 to Chr23: 30 mn ± 17 Chr1 to Chr6: up to 3 days	43 GB ± 17 Many execution abortions	17.3 h ± 3.3	700 MB ± 150

The running times obtained in the second protocol, in which reduced sets of n_{fs} SNPs are processed, give rise to several comments.

As expected, the cumulative running time for ranger diminishes with the size of the dataset (at most one hour an a half for a chromosome-wide dataset *versus* half an hour at most for a reduced dataset). The renown scalability of ranger is confirmed by our study [40].

In the MCMC-based software BEAM3, the cumulative running time shows an unexpected trend, on average, for chromosomes 7 to 23: the average was around one minute on a chromosome-wide dataset; it is around one day after dimension reduction. For chromosome 6, for instance, we still observe a higher running time for the reduced dataset than for the chromosome-wide dataset (38.9 h *versus* 22.4 h), but this time, the orders of magnitude are quite similar. In contrast, the expected ratio is observed for chromosomes 1 to 5 (above 8 days *versus* 5 days). For instance, the cumulative running time over 10 BEAM3 runs took 2 days and 18 h for chromosome 4, in the second protocol, whereas this cumulative time was above a week in the first protocol. To explain the unexpected high running times likely to be observed on the reduced datasets, we contacted BEAM3's author. The slowdown observed can be explained by BEAM3 having to deal with a lot of dependences, in order to find independent signals. This becomes particularly acute when SNP selection is applied. In contrast, if the SNP pool is *relatively* large, the program may converge faster because many more SNPs are independent.

For the same parameterization of AntEpiSeeker, the respective running times relative to two datasets of different sizes are not expected to differ. This fact is confirmed in Table 2.

In the first protocol applied to SMMB-ACO, we highlighted a great heterogeneity across chromosomes, with cumulative running times frequently reaching 3 days. In contrast, on the reduced datasets, a series of 10 SMMB-ACO runs can be processed at ease within a day.

Finally, BEAM3 remained the most greedy algorithm in the second protocol, closely followed by SMMB-ACO.

6.2 Interactions of Interest

Number of Interactions of Interest and Spatial Distribution over the Chromosomes. Table 3 highlights contrasts between the approaches. First, with only 18 interactions obtained *via* the first protocol, DASSO-MB was not expected to output IoIs, which is confirmed. We therefore excluded DASSO-MB from the second protocol. In the remainder of this article, we will not mention this method anymore. Second, a salient feature is the great heterogeneity in the numbers of IoIs detected by the four other methods. In the first protocol, these numbers scale in a ten thousands, a thousand, a hundred and a few tens for AntEpiSeeker, SMMB-ACO, BEAM3 and ranger respectively.

Table 3. Comparison of the numbers of interactions detected by ranger, BEAM3, AntEpiSeeker and SMMB-ACO, in the two protocols. N_t: total number of interactions identified by a method; N_{IoIs}: number of interactions of interest identified from the N_t previous interactions.

	First protocol		Second protocol	
	N_t	N_{IoIs}	N_t	N_{IoIs}
ranger	34	(**34**) (100%)	180	**180** (100%)
BEAM3	1,082	**131** (12.1%)	130	**11** (8.5%)
AntEpiSeeker	14,670	**13,062** (89.0%)	8,647	**7,633** (88.7%)
DASSO-MB	18	0	—	—
SMMB-ACO	6,346	**1,142** (18.0%)	498	**88** (17.7%)

Table 3 shows an impact of dimension reduction in the decrease of the total number of interactions output, N_t: the ratio of N_t measured for the first protocol to N_t observed for the second protocol is around 8 for BEAM3, close to 1.5 for AntEpiSeeker, and nearly 13 for SMMB-ACO. In the second protocol, AntEpiSeeker still outputs over 7,000 IoIs, whereas BEAM3 and SMMB-ACO respectively generate a dozen and less than a hundred IoIs. The situation of ranger

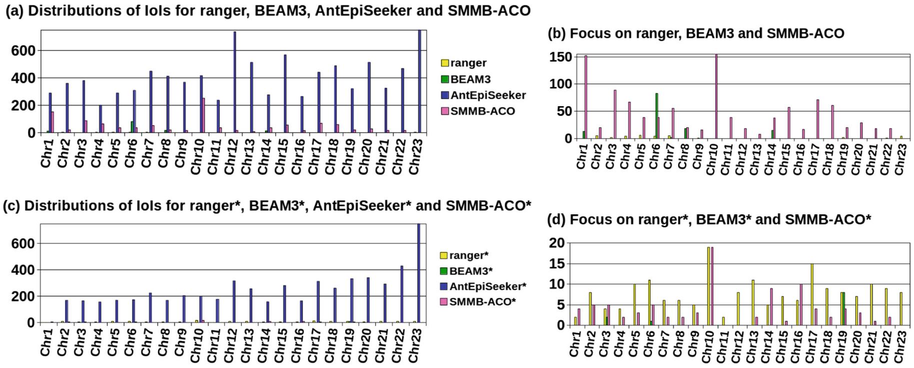


Fig. 3. Distributions of *interactions of interest* detected by ranger, BEAM3, AntEpiSeeker and SMMB-ACO, in the two protocols. "method*" denotes an approach with feature selection. AntEpiSeeker detected 13,062 IoIs which are spread over the 23 chromosomes (smallest number of IoIs for a chromosome: 202; 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. AntEpiSeeker* detected 7,633 IoIs distributed over all chromosomes but Chr1. An excess of IoIs in chromosome X is still observed for AntEpiSeeker* (35.2% of AntEpiSeeker's IoIs). 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)). SMMB-ACO* highlighted 88 IoIs spread over all chromosomes except Chr11, Chr12 and ChrX (smallest number of IoIs for a chromosome: 1; median number: 3; largest number: 19). The 131 IoIs detected by BEAM3 are located within 5 chromosomes only: Chr1, Chr6, Chr7, Chr8 and Chr14 respectively harbour 13, 83, 2, 18 and 15 IoIs. The 11 IoIs detected by BEAM3* are confined to Chr3 (2 IoIs), Chr6 (1 IoI) and Chr19 (8 IoIs). The 34 IoIs identified by ranger are distributed across 10 chromosomes: Chr2 to Chr7, Chr9, Chr19, Chr22 and Chr23 (minimum number of IoIs for these 10 chromosomes: 1; maximum number: 6). In contrast, the 180 IoIs highlighted by ranger* are spread over all chromosomes (smallest number of IoIs for a chromosome: 2; largest number: 19; median: 8).

is specific (five-fold increase): indeed, the relaxation of the significance threshold, from 5×10^{-4} (first protocol) to 5×10^{-2} (second protocol) was intended to put ranger on equal footing with the other methods' post-processings.

By construction of the protocols, the ratio of the number of IoIs to the total number of interactions is 100% for ranger. For the other methods, the feature selection does not allow to densify the number of IoIs in the outputs generated: the above ratio is constant through the two protocols, for each of the other methods: around 10% for BEAM3, close to 20% for SMMB-ACO and around 90% for AntEpiSeeker. This conclusion, which holds for three methods, is an important contribution of our study: it was not foreseeable that the much-vaunted credentials of feature selection for highlighting SNPs in epistasis detection would be undermined.

Table 4. Comparison of the distributions of p-values for the *interactions of interest* detected with the five approaches, in the two protocols. Four significance intervals are shown for $-log_{10}$(p-value). $-log_{10}(5 \times 10^{-2}) = 1.5$. The top section of the Table presents counts. The bottom section shows the corresponding percentages.

	First protocol					Second protocol			
	≥20	[10, 20[	[5, 10[	[1.5, 5[		≥20	[10, 20[	[5, 10[	[1.5, 5[
ranger	10	12	6	6	ranger*	3	3	1	173
BEAM3	0	0	0	131	BEAM3*	0	0	0	11
AntEpiSeeker	13	13	458	12,578	AntEpiSeeker*	10	8	315	7300
SMMB-ACO	0	0	6	1,136	SMMB-ACO*	0	0	2	86
ranger	29.40%	35.30%	17.65%	17.65%	ranger*	1.67%	1.67%	0.56%	96.10%
BEAM3	0%	0%	0%	100%	BEAM3*	0%	0%	0%	100%
AntEpiSeeker	0.10%	0.10%	3.51%	96.29%	AntEpiSeeker*	0.13%	0.11%	4.13%	95.63%
SMMB-ACO	0%	0%	0.53%	99.47%	SMMB-ACO*	0%	0%	2.27%	97.73%

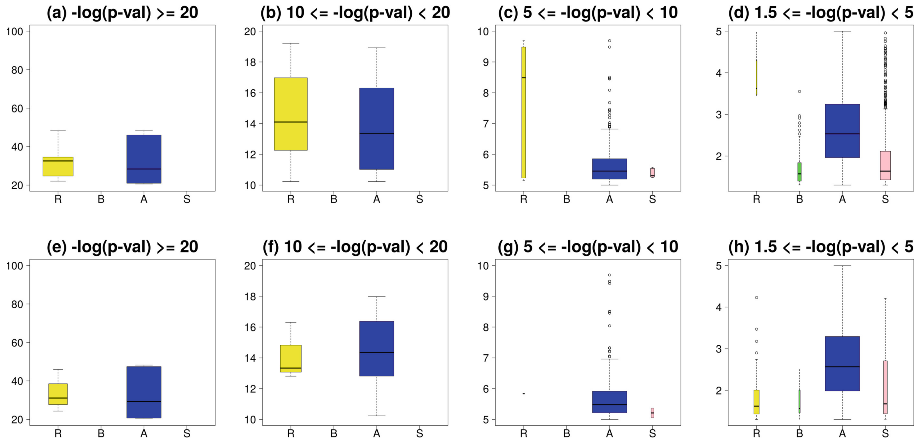


Fig. 4. Distributions of p-values for the *interactions of interest* detected by ranger, BEAM3, AntEpiSeeker and SMMB-ACO, in the two protocols. IoIs: interactions of interest. R: ranger. B: BEAM3. A: AntEpiSeeker. S: SMMB-ACO. $-log_{10}(5 \times 10^{-2}) = 1.5$.

Figure 3 focuses on the distribution of IoIs across the chromosomes. In the first protocol, 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 ten and five chromosomes respectively. Besides, the number of IoIs in BEAM3, around four times higher than in ranger, is circumscribed to a number of chromosomes that is two times less than for ranger. In the second protocol, the IoIs respectively detected by SMMB-ACO, AntEpiSeeker and ranger are present in nearly all chromosomes. The relaxation of the significance threshold explains the increase of IoIs in ranger. Again, the IoIs detected by BEAM3 are located in a few chromosomes (three chromosomes in the second protocol *versus* ten chromosomes in the first protocol).

Distributions of P-Values. Figure 4 and Table 4 allow to compare the distributions of IoI p-values obtained across ranger, BEAM3, AntEpiSeeker and SMMB-ACO, in the two protocols. We consider four intervals for the p-values.

We observe great discrepancies between the methods. A first remark is that AntEpiSeeker and ranger are the only two methods for which the p-values spread over the four intervals, for the two protocols: in contrast to the two other methods, AntEpiSeeker and ranger show p-values within the two first intervals (*i.e.*, below 10^{-10}) (subfigures (a), (b), (e) and (f)). A second observation is that BEAM3 is the only method whose 131 p-values (first protocol) and 11 p-values (second protocol) are all contained in the fourth interval (and are even confined to $[10^{-3.5}, 5 \times 10^{-2}]$ (subfigure (d)) and $[10^{-2.5}, 5 \times 10^{-2}]$ (subfigure (h)) for first and second protocols respectively. The overwhelming majority of IoIs detected by SMMB-ACO are also confined in the fourth interval. However, SMMB-ACO is able to highlight more significant IoIs than BEAM3: the SMMB-ACO p-values fall within ranges $[10^{-5}, 5 \times 10^{-2}]$ and $[10^{-4.4}, 5 \times 10^{-2}]$, respectively for the first and second protocols.

Besides, we already observed in Subsect. 6.2 that for each method except ranger, the percentages of IoIs (in the total set of interactions generated by the method) are identical for the two protocols. Again, the two protocols applied on the same method output close p-value distributions. The second conclusion to draw here is as follows: not only does feature selection not increase the rate of IoIs in the interactions generated by a method, feature selection does not enrich the IoIs generated with still more statistically significant IoIs.

6.3 Interactions of Interest Jointly Identified by Several Approaches

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 identified by AntEpiSeeker, ranger and SMMB-ACO. Under the second protocol, 4 IoIs were jointly identified by ranger* and AntEpiSeeker*.

Given the number of interactions output by AntEpiSeeker, an overlap was expected between AntEpiSeeker and some other method. However, an overlap was only observed between ranger and AntEpiSeeker. On the other hand, our study indicates that the mechanisms behind BEAM3, AntEpiSeeker and SMMB-ACO explore different sets of solutions. Finally, we observe that the selection discarded most of the SNPs that belonged to IoIs jointly identified by ranger and AntEpiSeeker. We emphasize here this impact of the feature selection: in the second protocol, ranger was run with a relaxed threshold (5×10^{-2} instead of 5×10^{-4}); we would therefore expect a larger overlap between AntEpiSeeker and ranger (which we did not check), but we verified that ranger* and AntEpiseeker* do not overlap much when the input dataset is reduced by feature selection.

Table 5. Statistics on the networks of *interactions of interest*, across all chromosomes, for ranger, BEAM3, AntEpiSeeker and SMMB-ACO, in the two protocols. "method*" denotes an approach with feature selection. IoI: interaction of interest. n_{IoIs}: total number of IoIs detected by a method; nb_{chr}: number of chromosomes in which IoI networks were found; nb_{net}: number of such IoI networks; N_i, N_g and N_s: respectively, number of IoIs, genes and SNPs in a network. Q1, Q2 and Q3 respectively denote the first quartile, the median and the third quartile.

Method	n_{IoIs}	nb_{chr}	nb_{net}	N_i			N_g			N_s		
				Q1	Q2	Q3	Q1	Q2	Q3	Q1	Q2	Q3
ranger	**34**	7	8	2	**3.5**	4	0	**0.5**	1	3	**4.5**	5
ranger*	**180**	23	25	4	**6**	8	2	**3**	6	5	**7**	9
BEAM3	**131**	4	4	14.3	**16.5**	33.8	10	**10.5**	13.8	12	**13**	22
BEAM3*	**11**	1	1	8	**8**	8	6	**6**	6	8	**8**	8
AntEpiSeeker	**13,062**	23	929	2	**3**	4	2	**3**	4	3	**4**	5
AntEpiSeeker*	**7,633**	23	611	2	**3**	4	1	**2**	4	3	**4**	5
SMMB-ACO	**1,142**	22	87	2	**4**	8.5	2	**4**	8	3	**5**	9.5
SMMB-ACO*	**88**	20	15	2	**2**	3	1.8	**3**	4	3	**3**	4

6.4 Networks of Interactions of Interest

This subsection is devoted to the detection of networks of interactions of interest (IoIs). Some statistics on the number of networks identified per method are first provided. Then we focus on the distributions of the networks' sizes. We end this subsection by comparing across all methods the spatial distributions of the networks of IoIs across chromosomes.

Number of Networks Detected and Distribution of Their Sizes. For each method and for each chromosome, we have identified all pairs of IoIs whose members share a SNP. This led us to build networks of IoIs. Table 5 allows to compare the four methods, in the two protocols, with respect to the numbers of IoIs, genes and SNPs involved in each of the networks identified across all chromosomes. We used the R package biomaRt to identify the genes associated with SNPs [7]. We could only identify genes for SNPs whose RefSNP label (*e.g.*, rs1996546) is known for the corresponding SNP provided by the WTCCC Consortium. For example, a network involving 6 IoIs and 7 SNPs was identified in chromosome 5 through ranger; however, none of the corresponding genes could be retrieved in this case.

We first observe that the number of IoI networks is more or less related to the total number of IoIs detected by the method considered. This observation was expected for statistical reasons. The second remark to draw from Table 5 is the existence of a contrast between the networks in ranger*, BEAM3* and BEAM3, and the networks in the other methods. For the three former methods, the medians for the number of IoIs in a network are respectively 6, 8 and 16.5. All the other methods show a median in interval $[2, 4]$. The maxima observed for the 25, 1 and 4 networks respectively identified in ranger*, BEAM3* and BEAM3 are respectively 18, 8 and 81 IoIs in a network.

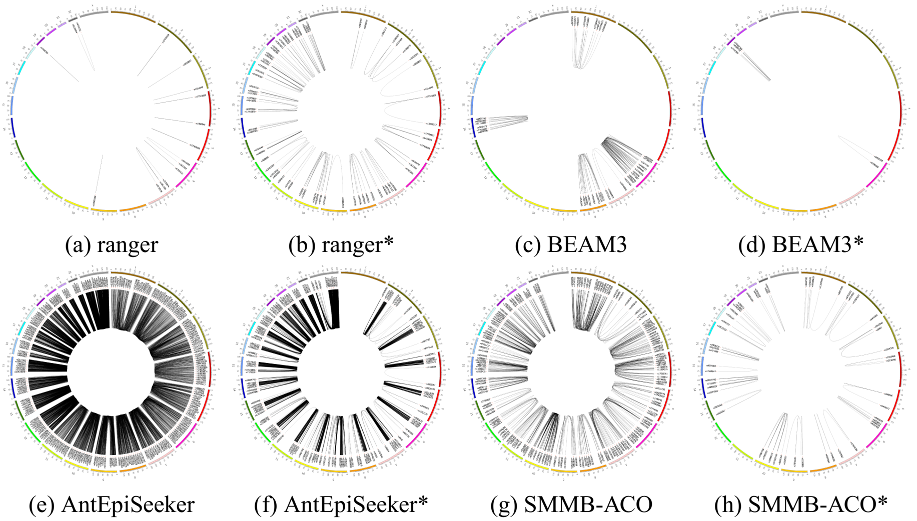
The explanation for these high medians lies in the small number of networks identified, which gives weight to the few networks of large sizes. Indeed, ranger and SMMB-ACO* only detected 8 and 15 IoI networks respectively, but no outlier exists for ranger (maximum number of IoIs in a network: 6), or only one exists for SMMB-ACO* (maximum number of IoIs in a network: 9). On the other hand, it is remarkable that the *median* number of IoIs in a network is not inflated for SMMB-ACO, AntEpiSeeker* and AntEpiSeeker, which yielded over a thousand IoIs. Therefore, we conclude that the number of IoIs detected by a method impacts the number of IoI networks identified, but not the size of the networks identified.

Besides, as highlighted in Sect. 6.2 (caption of Fig. 3), a specific behavior was shown for AntEpiSeeker and AntEpiSeeker*: they detected a third of their IoIs in chromosome X. Again, a specific characteristic is shown: the first method identified a *single* network containing 4,330 IoIs, whereas the second method detected a *unique* network of size 2,575 IoIs. All the remarks provided in this paragraph hold when we consider N_s, the number of SNPs, to measure the size of an IoI network. The conclusions are similar if we consider N_g, the number of genes, except that we could not list genes related to SNPs with unknown RefSNP labels, on the one hand, and that some SNPs are connected to several genes, on the other hand.

Spatial Distribution of the Interactions of Interest Detected Across Chromosomes. Figure 5 allows the visual comparison of the spatial distributions of IoIs across chromosomes, for ranger, BEAM3, AntEpiSeeker and SMMB-ACO, in the two protocols. As an illustration, Fig. 6 focuses on SMMB-ACO, for which a SNP may belong to 10 IoIs and even up to 19 (chromosome 10). These two latter figures were drawn using the R software package circos dedicated to data visualization through circular layouts [23].

Illustration with a Network Detected by SMMB-ACO, and Biological Insights. As an illustration, we show in Table 6 the 19 IoIs constituting one of the networks identified in chromosome 10 by SMMB-ACO run downstream feature selection. This network involves 13 SNPs and is related to 6 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.

Besides a number of standard single-SNP GWASs, the few AISs devoted to Crohn's disease (CD) focus on genes or pathways already known to contribute to the disease onset, such as NOD on Chr16, CCNY and NKX2-3 on Chr10, LGALS9 and STAT3 on Chr17, and SBNO2 on Chr19 [21,26]. It is not a surprise that among the six genes highlighted in the network of Fig. 7, two genes are already known to impact CD onset: CCNY and NKX2-3. It was also expected that our protocol designed for AIS investigation without prior biological knowledge would detect novel interaction candidates, which it does.


(a) ranger (b) ranger* (c) BEAM3 (d) BEAM3*

(e) AntEpiSeeker (f) AntEpiSeeker* (g) SMMB-ACO (h) SMMB-ACO*

Fig. 5. Spatial distributions of the *interactions of interest* detected across the chromosomes, for ranger, BEAM3, AntEpiSeeker and SMMB-ACO under the two protocols. "method*" denotes an approach with feature selection.

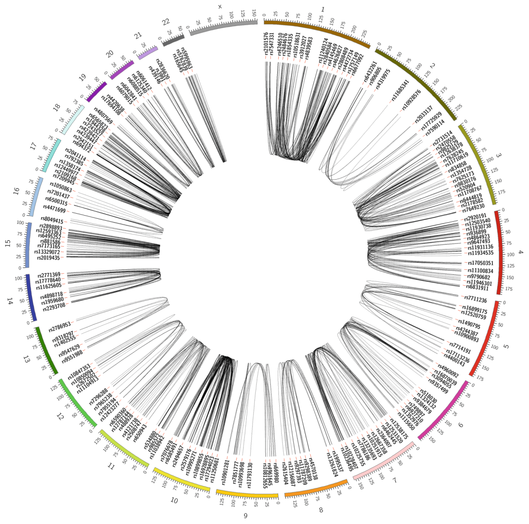


Fig. 6. Spatial distribution of the interactions of interest detected across the chromosomes, for SMMB-ACO.

6.5 Discussion

We gained considerable and unforeseeable insights from our study. First, on the CD dataset, DASSO-MB is of no help. The verbose AntEpiSeeker provides a

Table 6. Network of 19 interactions of interest, 13 SNPs and 6 known genes, identified by SMMB-ACO* in chromosome 10. "SMMB-ACO*" denotes the approach with feature selection. A letter in first column and a letter in second column denote an interaction (for instance, G-F in first line). iv: intron variant; gutv: genic upstream transcript variant; utv: upstream transcript variant; gdtv: genic downstream transcript variant; nctv: non coding transcript variant; 3puv: 3 prime UTR variant. CREM 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). CUL2 is a major component of multiple cullin-RING-based ECS (ElonginB/C-CUL2/5-SOCS-box protein) E3 ubiquitin-protein ligase complexes; these complexes mediate the ubiquitination of target proteins. CUL2 is ubiquitous (27 tissues, including colon, small intestine and appendix). NKX2-3 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. LINC01475 (long intergenic non-protein coding RNA 1475) is expressed in 7 tissues including colon, small intestine, duodenum and appendix. CPXM2, a protein of the carboxypeptidase X, M14 family member 2, is broadly expressed in 21 tissues. CCNY belongs to the cyclins, which control cell division cycles and regulate cyclin-dependent kinases (27 tissues including colon, small intestine, duodenum and appendix).

SNP	RefSNP label	Location	Gene	SNP	RefSNP label	Location	Gene
G	rs7095491	99514301	—	F	rs2505639	35185493	CREM (gutv, gdtv, iv)
				H	rs11010067	35006503	—
				I	rs4934709	35050396	CUL2 (iv)
				K	rs17582416	34998722	—
				D	rs10761659	62685804	—
D	rs10761659	62685804	—	E	rs7078219	99514608	—
				J	rs10883371	99532698	NKX2-3, LINC01475 (utv)
				L	rs1548964	99529896	LINC01475 (iv)
				B	rs7067790	123917521	CPXM2 (gutv, iv)
				M	rs3936503	35260329	CCNY (gutv, iv)
L	rs1548964	99529896	LINC01475 (iv)	F	rs2505639	35185493	CREM (gutv, gdtv, iv)
				I	rs4934709	35050396	CUL2 (iv)
				A	rs10995271	62678726	—
B	rs7067790	123917521	CPXM2 (gutv, iv)	A	rs10995271	62678726	—
				C	rs6601764	3820350	—
J	rs10883371	99532698	NKX2-3, LINC01475 (utv)	F	rs2505639	35185493	CREM (gutv, gdtv, iv)
				I	rs4934709	35050396	CUL2 (iv)
I	rs4934709	35050396	CUL2 (iv)	A	rs10995271	62678726	—
M	rs3936503	35260329	CCNY (gutv, iv)	E	rs7078219	99514608	—

wealth of results, in which we suspect a high rate of false positives. Moreover, under each protocol, 30% of the IoIs generated by AntEpiSeeker are discovered in chromosome X, a chromosome not related to Crohn's disease. Besides, it appears

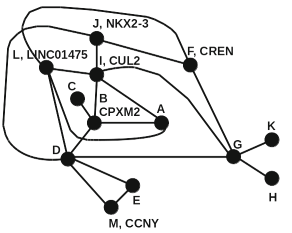


Fig. 7. Network of 19 interactions of interest, 13 SNPs and 6 known genes, identified by SMMB-ACO* in chromosome 10. See Table 6 for complementary information.

that the only way to reduce this verbosity is decreasing the p-value threshold. The widely cited software BEAM3 cannot 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. The reason lies in BEAM3's low number of 2-way interactions. In contrast, SMMB-ACO notably detects IoIs in chromosome 10, a chromosome which harbours genes connected to CD. Besides, in spite of dimension reduction, this result holds when feature selection is applied upstream SMMB-ACO. To note, ranger is the only method in the first protocol to output a high proportion (around 65%) of IoIs with low p-values (below 10^{-10}). However, with a relaxed threshold and upstream feature selection, this phenomenon is marginal.

The case of ranger set apart, feature selection does not help increase the rate of IoIs (significant 2-way interactions) in the interactions generated by any method. In BEAM3, AntEpiSeeker and SMMB-ACO, the ratio of the number of IoIs to the total number of interactions detected remains constant through the two protocols: around 10% for BEAM3, around 90% for AntEpiSeeker, and close to 20% for SMMB-ACO. This conclusion is an important contribution of our study since feature selection is often put forth as a means to not only reduce the search space, but reduce it to a subspace of interest. In the case of the CD dataset, feature selection just implemented data dimension reduction. Besides, the three methods differ in the ratios of IoIs identified, which shows that these methods do not explore the same solution space. Moreover, feature selection does not enrich the IoIs generated with still more statistically significant IoIs.

Finally, the number of IoIs detected by a method impacts the number of IoI networks identified, but not the size of the networks identified.

The two experimental protocols implemented in this extensive analysis allow us to highlight a great heterogeneity between the methods compared, in all domains: running times, numbers of IoIs detected, distributions of p-values for the IoIs identified, numbers of IoI networks and distributions of the sizes of the latter. Some methods, which fall into the category of widely cited approaches in the literature, however showed weaknesses. BEAM3 is extremely time consuming for large chromosome datasets. At the opposite, a flaw was also evidenced since BEAM3 produced surprisingly high running times for small datasets obtained *via* feature selection. The verbosity of AntEpiSeeker, even on reduced datasets of around 5,000 SNPs, renders its use questionable for practitioners: it is not affordable to biologically validate IoIs whose number scales in thousands. Thus, the quickness of AntEpiSeeker is impeded by this verbosity. The more recent approach SMMB-ACO, a complex method, is nonetheless faster than the reference software BEAM3. However, SMMB-ACO still requires memory management improvements since it was shown to consume fluctuating memory across several runs on the same large dataset.

7 Conclusion and Further Work

For computational reasons, in the GWAS field, simulations are performed using data whose dimension is not comparable with real genome-wide datasets'. Consequently, these simulations reveal nothing about the effectiveness and efficiency of methods in true conditions. Moreover, the ratio between the number of SNPs and the number of subjects observed is not comparable between simulated and real datasets.

This work departs from the standard framework in genetic association studies as it reports an unprecedented extensive comparative analysis of five approaches on large-scale real data, following two experimental protocols. In the first protocol, the native methods are used straightforwardly. In the second protocol, feature selection is performed upstream of these methods. Our analysis rapidly discarded DASSO-MB, to focus on the two remaining state-of-the-art approaches designed to detect epistasis from scratch, AntEpiSeeker and BEAM3. An unavoidable reference in GWAS, ranger was used in combination with logistic regression, to detect epistasis. A more recent approach, SMMB-ACO, was included in the comparison. We designed the two experimental protocols, taking care to output comparable sets of (2-way) interactions across the approaches. In the second protocol, ranger was used upstream any of the former methods (including ranger itself), to implement feature selection. Using 23 chromosome-wide case control datasets related to Crohn's disease, we achieved 1,150 feature selection phases together with 1,886 genetic analyses. We observed a great heterogeneity across methods in all aspects: running times and memory requirements, numbers of interactions of interest (IoIs) output, p-value ranges, numbers of IoI networks and distributions of the sizes of the latter.

The insights gained in the present work will lead us to discart feature selection in our future work. We plan to extend the comparative analysis to six additional genome-wide real 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. We also plan to consider various genetic models.

Acknowledgment. This work was supported by the GRIOTE Research project funded by the Pays de la Loire Region. In this study, we have processed real-world data generated by the Wellcome Trust Case Control Consortium. The experiments reported in this paper were performed at the CCIPL (Centre de Calcul Intensif des Pays de la Loire).

Appendix

Table 7. Parameter adjustment for the five methods. (Table published in [3]).

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
AntEpiSeeker	**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_a above variables ($k < K_a$)	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

References

1. Aflakparast, M., Salimi, H., Gerami, A., Dubé, M.-P., Visweswaran, S., et al.: Cuckoo search epistasis: a new method for exploring significant genetic interactions. Heredity **112**, 666–764 (2014)
2. Ayers, K., Cordell, H.: SNP selection in genome-wide and candidate gene studies via penalized logistic regression. Genet. Epidemiol. **34**(8), 879–891 (2010)
3. Boisaubert, H., Sinoquet, C.: Detection of gene-gene interactions: methodological comparison on real-world data and insights on synergy between methods. In: Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019), vol. 3, pp. 30–42. BIOINFORMATICS (2019)
4. Breiman, L.: Bagging predictors. Mach. Learn. **24**(2), 123–140 (1996). https://doi.org/10.1023/A:1018054314350
5. Chang, Y.-C., Wu, J.-T., Hong, M.-Y., Tung, Y.-A., Hsieh, P.-H., et al.: GenEpi: gene-based epistasis discovery using machine learning (2018). bioRXiv, https://doi.org/10.1101/421719
6. Chatelain, C., Durand, G., Thuillier, V., Augé, F.: Performance of epistasis detection methods in semi-simulated GWAS. BMC Bioinform. **19**(1), 231 (2018)
7. Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., Moor, B.D., et al.: Biomart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics **21**, 3439–3440 (2005)
8. Fergus, P., Montanez, C., Abdulaimma, B., Lisboa, P., Chalmers, C.: Utilising deep learning and genome wide association studies for epistatic-driven preterm birth classification in African-American women (2018). arXiv preprint, arXiv:1801.02977
9. Furlong, L.: Human diseases through the lens of network biology. Trends Genet. **29**, 150–159 (2013)
10. Gao, H., Granka, J., Feldman, M.: On the classification of epistatic interactions. Genetics **184**(3), 827–837 (2010)
11. Gibert, J.-M., Blanco, J., Dolezal, M., Nolte, V., Peronnet, F., Schlötterer, C.: Strong epistatic and additive effects of linked candidate SNPs for Drosophila pigmentation have implications for analysis of genome-wide association studies results. Genome Biol. **18**, 126 (2017)
12. Gilbert-Diamond, D., Moore, J.: Analysis of gene-gene interactions. Current Protocols in Human Genetics, 0 1: Unit1.14 (2011)
13. Gola, D., Mahachie John, J., van Steen, K., König, I.: A roadmap to multifactor dimensionality reduction methods. Briefings Bioinform. **17**(2), 293–308 (2016)
14. Graham, D., Xavier, R.: From genetics of inflammatory bowel disease towards mechanistic insights. Trends Immunol. **34**, 371–378 (2013)
15. Han, B., Chen, X.-W.: bNEAT: a Bayesian network method for detecting epistatic interactions in genome-wide association studies. BMC Genomics **12**(Suppl. 2), S9 (2011)
16. Han, B., Chen, X.-W., Talebizadeh, Z.: FEPI-MB: identifying SNPs-disease association using a Markov blanket-based approach. BMC Bioinform. **12**(Suppl. 12), S3 (2011)
17. Han, B., Park, M., Chen, X.-W.: A Markov blanket-based method for detecting causal SNPs in GWAS. BMC Bioinform. **11**(Suppl. 3), S5 (2010)
18. Hohman, T., Bush, W., Jiang, L., Brown-Gentry, K., Torstenson, E., et al.: Discovery of gene-gene interactions across multiple independent datasets of Late Onset Alzheimer Disease from the Alzheimer Disease Genetics Consortium. Neurobiol. Aging **38**, 141–150 (2016)

19. Jiang, X., Neapolitan, R., Barmada, M., Visweswaran, S., Cooper, G.: A fast algorithm for learning epistatic genomic relationships. In: Proceedings of the Annual American Medical Informatics Association Symposium (AMIA 2010), pp. 341–345 (2010)
20. Jing, P., Shen, H.: MACOED: a multi-objective ant colony optimization algorithm for SNP epistasis detection in genome-wide association studies. Bioinformatics **31**(5), 634–641 (2015)
21. Khor, B., Gardet, A., Ramnik, J.: Genetics and pathogenesis of inflammatory bowel disease. Nature **474**(7351), 307–317 (2011)
22. Koller, D., Sahami, M.: Toward optimal feature selection. In: Proceedings of the 13th Conference on Machine Learning (ICML 1996), pp. 284–292. Morgan Kaufmann, San Fransisco (1996)
23. Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., et al.: Circos: an information aesthetic for comparative genomics. Genome Res. **19**(9), 1639–1645 (2009)
24. Li, J., Malley, J., Andrew, A., Karagas, M., Moore, J.: Detecting gene-gene interactions using a permutation-based random forest method. BioData Min. **9**, 14 (2016)
25. Lunetta, K., Hayward, L., Segal, J., Eerdewegh, P.V.: Screening large-scale association study data: exploiting interactions using random forests. BMC Genet. **5**, 32 (2004)
26. McGovern, D., Kugathasan, S., Cho, J.: Genetics of inflammatory bowel diseases. Gastroenterology **149**(5), 1163–1176 (2015)
27. Nicodemus, K., Law, A., Radulescu, E., Luna, A., Kolachana, B., et al.: Biological validation of increased schizophrenia risk with NRG1, ERBB4, and AKT1 epistasis via functional neuroimaging in healthy controls. Arch. Gen. Psychiatry **67**(10), 991–1001 (2013)
28. Pearl, J.: Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference. Morgan Kaufmann Publishers Inc., San Francisco (1988)
29. Sackton, T., Hartl, D.: Genotypic context and epistasis in individuals and populations. Cell **166**(2), 279–287 (2016)
30. Schwarz, D., König, I., Ziegler, A.: On safari to random jungle: a fast implementation of random forests for high-dimensional data. Bioinformatics **26**(14), 1752–1758 (2010)
31. Shen, Y., Liu, Z., Ott, J.: Support vector machines with L1 penalty for detecting gene-gene interactions. Int. J. Data Min. Bioinform. **6**, 463–470 (2012)
32. Sinoquet, C., Niel, C.: Enhancement of a stochastic Markov blanket framework with ant colony optimization, to uncover epistasis in genetic association studies. In: Proceedings of the 26th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN 2018), pp. 673–678 (2018)
33. Stanislas, V., Dalmasso, C., Ambroise, C.: Eigen-Epistasis for detecting gene-gene interactions. BMC Bioinform. **18**, 54 (2017). https://doi.org/10.1186/s12859-017-1488-0
34. Sun, Y., Shang, J., Liu, J.-X., Li, S., Zheng, C.-H.: epiACO - a method for identifying epistasis based on ant colony optimization algorithm. BioData Min. **10**, 23 (2017)
35. Uppu, S., Krishna, A., Gopalan, R.: Towards deep learning in genome-wide association interaction studies. In: Proceedings of the 20th Pacific Asia Conference on Information Systems (PACIS2016), p. 20 (2016)

36. Urbanowicz, R., Meeker, M., LaCava, W., Olson, R., Moore, J.: Relief-based feature selection: introduction and review. J. Biomed. Inform. **85**, 189–203 (2018)
37. Vineis, P., Pearce, N.: Missing heritability in genome-wide association study research. Nat. Rev. Genet. **11**, 589–589 (2010)
38. Visscher, P., Wray, N., Zhang, Q., Sklar, P., McCarthy, M., et al.: 10 years of GWAS discovery: biology, function, and translation. Am. J. Hum. Genet. **101**(1), 5–22 (2017)
39. Wang, Y., Liu, X., Robbins, K., Rekaya, R.: AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. BMC Res. Notes **3**, 117 (2010)
40. Wright, M., Ziegler, A.: ranger: a fast implementation of random forests for high dimensional data in C++ and R. J. Stat. Softw. **77**(1), 1–17 (2017)
41. Zhang, Y.: A novel Bayesian graphical model for genome-wide multi-SNP association mapping. Genet. Epidemiol. **36**(1), 36–47 (2012)
42. Zhang, Y., Liu, J.: Bayesian inference of epistatic interactions in case-control studies. Nat. Genet. **39**, 1167–1173 (2007)
43. Zhu, Z., Tong, X., Zhu, Z., Liang, M., Cui, W., et al.: Development of MDR-GPU for gene-gene interaction analysis and its application to WTCCC GWAS data for type 2 diabetes. PLOS ONE **8**(4), e61943 (2013)
44. Zuk, O., Hechter, E., Sunyaev, S., Lander, E.: The mystery of missing heritability: genetic interactions create phantom heritability. Proc. Nat. Acad. Sci. **109**, 1193–1198 (2012)