Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Computational biology

SMMB – A stochastic Markov-blanket framework strategy for epistasis detection in GWAS

Clément Niel[1], Christine Sinoquet[1,*], Christian Dina[2] and Ghislain Rocheleau[3,4]

[1]Laboratoire des Sciences du Numérique de Nantes (LS2N), Centre National de la Recherche Scientifique UMR 6004, University of Nantes, Nantes, France, [2]Institut du Thorax, Institut National de la Santé et de la Recherche Médicale UMR 1087, Centre National de la Recherche Scientifique UMR 6291, University of Nantes, Nantes, France, [3]European Genomic Institute for Diabetes FR3508, Centre National de la Recherche Scientifique UMR 8199, Lille 2 University, Lille, France, [4]Current affiliation: Maelstrom Research, Montreal, Canada.

*To whom correspondence should be addressed.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Large scale genome-wide association studies (GWAS) are tools of choice for discovering associations between genotypes and phenotypes. To date, many studies rely on univariate statistical tests for association between the phenotype and each assayed single nucleotide polymorphism (SNP). However, interaction between SNPs, namely epistasis, must be considered when tackling the complexity of underlying biological mechanisms. Epistasis analysis at large scale entails a prohibitive computational burden when addressing the detection of more than two interacting SNPs. In this paper, we introduce a stochastic causal graph-based method, SMMB, to analyze epistatic patterns in GWAS data.

Results: We present SMMB (Stochastic Multiple Markov Blanket algorithm), which combines both ensemble stochastic strategy inspired from random forests and Bayesian Markov blanket-based methods. We compared SMMB with three other recent algorithms using both simulated and real datasets. Our method outperforms the other compared methods for a majority of simulated cases of 2-way and 3-way epistasis patterns (especially in scenarii where minor allele frequencies of causal SNPs are low). Our approach performs similarly as two other compared methods for large real datasets, in terms of power, and runs faster.

Availability: parallel version available on https://ls2n.fr/listelogicielsequipe/DUKe/128/
Contact: christine.sinoquet@univ-nantes.fr
Supplementary information: Supplementary data are available at Bioinformatics online.

1 Introduction

Research in human genetics has heavily relied on genome-wide case-control association studies over the past decade. The challenge of genome-wide association studies (GWAS) is to discover genetic variants which confer an increased risk of disease, with the clear objective of unravelling the genetic etiology of human diseases, leading to potentially improved patient diagnosis. Exploration of the genetic architecture underlying complex diseases, like type 2 diabetes or coronary heart disease, was mainly restricted to the examination of individual genetic variants acting independently and in an additive fashion. Interaction between genes or single nucleotide polymorphisms (SNPs), namely epistasis, represents a more recent line of inquiry with increased interest over the last five years. SNPs with null or small marginal effects, not detected using the standard single-SNP tests commonly run in GWAS, may show significant effects when interacting with genotypes at other SNPs. Unfortunately, exploration of all SNP combinations incurs a heavy computational burden due to the large size of typical GWAS datasets with hundreds of thousands to several millions of SNPs assayed for many thousand individuals.

The first definition of epistasis was coined by Bateson a century ago to portray a condition in which the effect of a gene is masked by the effect of another gene (Bateson,W., 1909). Later, Fisher defined epistasis as the (statistical) departure from the additive effects of two loci on the phenotype (Fisher,R.A., 1918). Many genetic interactions, supported by experimental validation, have already been identified, especially in plant species (Matsubara,K. *et al.*, 2015; Best,N.M. *et al.*, 2016; He,S. *et al.*, 2016; Press,M.O. and Queitsch C., 2017).

A recent review has described the main strategies employed to explore epistatic interactions (Niel,C. *et al.*, 2015). Exhaustive strategies, like GWIS (Genome Wide Interaction Search) (Goudey,B. *et al.*, 2013), can handle large datasets, but are limited to exploration of two-way interactions. By contrast, the popular MDR (Multifactor Dimensionality Reduction) approach (Hahn,L.W. *et al.*, 2003) can test every SNP combination up to a user-specified order of interaction, but cannot handle datasets containing many hundreds of SNPs. Although exhaustive search strategies represent by far the ideal investigation for many complex diseases, high-dimensional datasets comprised of millions of SNPs prohibit their practical application in real GWAS data.

At the sheer scale of GWAS, standard multivariate regression breaks down. In software Mendel, lasso penalized regression is used to select important marginal predictors in a first stage, and to look for interactions among the latter predictors afterwards (Wu,T.T *et al.*, 2009). Lasso penalty is adjusted to single out a certain number of predictors with non-zero regression coefficients, to provide a fixed number of predictors.

Other approaches, derived from the field of artificial intelligence, have been proposed: ensemble learning techniques based on random forests (SNPInterForest, Yoshida,M. and Koike,A. 2011; Random Jungle, Schwarz,D.F *et al.* 2010), metaheuristics designed for combinatorial optimization (AntEpiSeeker, Wang,Y. *et al.*, 2010; MACOED (Multi-objective Ant Colony Optimization algorithm for Epistasis Detection), Jing,P.J. and Shen,H.B., 2015; epiACO, Sun *et al.*, 2017), and Bayesian networks (BEAM (Bayesian Epistasis Association Mapping), Zhang,Y. and Liu,J.S., 2007; bNEAT, Han,B. and Chen,X.W., 2011). More recent approaches propose hybridization of previous methods: a two-stage approach which couples a two-locus combination filtering stage with ant colony optimization (HiSeeker, Liu *et al.*, 2017), or a combination of the K-Nearest Neighbors method with Multifactor Dimensionality Reduction (KNN-MDR, Abo Alchamlat,S. and Farnir,F., 2017). Even though various approaches have been put forward to identify epistatic interactions, their common drawback is the lack of detection power, especially when epistatic interactions involve SNPs with no marginal effect, a situation referred to as pure epistasis. This situation may entail missing interesting associations which could partly explain the remaining missing heritability in many complex traits (Maher,B., 2008; Manolio,T.A. *et al.*, 2009).

In the field of machine learning, feature selection algorithms are particularly relevant to analyze GWAS data. A solution to tackle the feature selection problem is to determine a subset of variables that can lead to a variable of interest being independent from the effect of all other remaining variables. In a Bayesian network context, such a variable subset is called a Markov blanket. A variation of the state-of-the-art IAMB (Incremental Association Markov Blanket) algorithm (Tsamardinos,I. *et al.*, 2003), called DASSO-MB (Detection of ASSOciations using Markov Blanket), was developed by Han and co-workers to learn the Markov blanket of a phenotype variable in a GWAS context (Han,B. *et al.*, 2010). However, like previously discussed approaches, the DASSO-MB algorithm is also prone to miss SNPs interacting in pure epistasis relationship.

In this work, we explore the Markov blanket approach, and we introduce an innovative hybrid approach which combines the Markov blanket methods with stochastic and ensemble features. Our approach is called SMMB (Stochastic Multiple Markov Blankets). To note, our method, as well as all other methods cited above, only deals with binary phenotypes, such as disease status. Hence, we conducted experiments on a wide range of simulated disease models, and compared the performance of SMMB to those of AntEpiSeeker, BEAM and DASSO-MB.

Section 2 provides some definitions related to Bayesian networks and Markov blankets and points out the shortcomings of existing algorithms when dealing with epistasis detection problems. Section 3 focuses on the detailed description of the SMMB method, while experimental results and discussion are presented in Section 4 and 5, respectively.

2 Background for the Markov blanket concept

In this section, we introduce specific properties of a Bayesian network. Then, we present theoretical aspects about the Markov blanket, which is a subnetwork of a Bayesian network. In the remainder of the article, notations $\boldsymbol{V}$ and T will refer to the SNPs and the phenotype (*i.e.* affected / unaffected status), respectively, in a GWAS context. Sets of variables (*i.e.* ensembles) will be denoted in bold font.

2.1. Bayesian network

In a Bayesian network B=<**V**, G, J>, random variables $\{X_1, ..., X_p\}$ of a set **V** are represented as nodes of a directed acyclic graph G. In G, edges between nodes represent the direct dependences between these variables. Moreover, the structure of G expresses conditional independence within variables in **V**. Based on the Markov condition (definition 1), the joint probability distribution J of the variables may be factorized.

Definition 1. Markov condition property.

Each variable is conditionally independent of its non-descendants given its parents in G. Therefore, the corresponding joint probability distribution J can be represented by equation (1), where $Pa(X_i)$ denotes the set of parents of X_i in G:

$$P(X_1, \dots, X_p) = \prod_{i=1}^{p} P(X_i \mid Pa(X_i)) \quad (1)$$

Definition 2. Conditional independence.

Conditional independence is observed if the occurrence of an event is not affected by another event under a set of conditions. We illustrate conditional independence through the following toy example. Given three random variables X, Y and Z, if the probability distribution of X conditional on Y and Z is equal to the probability distribution of X conditional on Z, then X is conditionally independent of Y given Z. This definition is formalized in equation (2):

$$P(X|Y,Z) = P(X|Z) \Leftrightarrow X \perp Y \mid Z \quad (2)$$

This definition can be extended to a set of variables **Z**.

2.2. Markov blanket

The Markov blanket of a target variable T, **MB**(T), is the minimal set of variables that can render T independent from all other variables that do not belong to **MB**(T). In other words, all variables not included in **MB(***T***)** are probabilistically independent of the variable T conditional on the Markov blanket of T (equation (3)).

$$\forall X \in \boldsymbol{V} \backslash \boldsymbol{MB}(\boldsymbol{T}), X \perp T \mid \boldsymbol{MB}(T) \quad (3)$$

Koller and Sahami determined that the Markov blanket of T is theoretically the optimal set of variables to predict the value of T (Koller,D. and Sahami,M., 1996). Therefore, any variable which is not in the Markov blanket of T can be ignored without significantly affecting the perfor-

mance of the learned predictor. In a genotype-phenotype association problem, a Markov blanket should be intuitively understood as the most parsimonious set of SNPs that predict the case-control status of an individual given its genotype. The learning stage of a Markov blanket is completed when no more SNP brings any additional predictive information about the case-control status of an individual. Since variables outside **MB***(T)* satisfy equation (3), a basic algorithm (*e.g.* IAMB) applied to epistasis detection successively attempts to identify the candidate SNP most associated with the phenotype, conditional on the current **MB***(T)*. This candidate is added to **MB***(T)* if the conditional dependence is statistically significant given some type I error threshold. This process is iterated as long as **MB***(T)* can be modified.

In equation (3), ***V*** denotes the whole set of variables represented in a Bayesian network, *i.e.* all SNPs involved in a GWAS context. As mentioned previously, a Markov blanket is a subnetwork of a Bayesian network structure: the Markov blanket of variable *X* is the set consisting of its parents, children, and co-parents (*i.e.* parents of common children).

Since learning a full Bayesian network is a NP-complete problem (Chickering,D.M. *et al*., 2004), heuristics are used in several algorithms to closely approximate the Markov blanket. During the past decade, various proposals were made to learn efficiently an optimal Markov blanket. The pioneer algorithm IAMB (Tsamardinos,I. *et al*., 2003) was further refined in variants such as inter-IAMB (Tsamardinos,I. *et al*., 2003) and Fast-IAMB (Yaramakala,S. and Margaritis,D., 2005). Limitations of IAMB regarding data efficiency were also overcame with MMPC/MB (Max-Min Parents and Children/Markov Blanket) (Peña,J.M. *et al*., 2005) and HITON-PC/MB (Aliferis,C.F. *et al*., 2003), by taking into account the Bayesian graph topology when learning a Markov blanket. DASSO-MB and IMBED (Improved Markov Blanket method for Epistasis Detection) (Yanlan,L. and Jiawei,L., 2012) were finally developed to deal with identification of disease susceptibility SNP associations with a Markov blanket approach. All these algorithms start from an empty Markov blanket and are based on two main stages. In a first stage, candidate variables are added to the Markov blanket. This is where heuristics are used to add the most promising variables. In a GWAS context, the heuristic used in DASSO-MB consists in iteratively including the SNP that shows the strongest statistical dependence with the phenotype conditional on the SNPs already present in the Markov blanket (provided the association is significant). However, since the first iteration of DASSO-MB performs a test conditional on an empty Markov blanket, this first iteration indeed includes the SNP that shows the greatest marginal effect. Moreover, the conditional independence tests in the remaining iterations will heavily rely on this first marginal effect-dependent inclusion. This general drawback is common to all methods cited above. The second stage of these algorithms is a direct application of equation (3), which translates into removing false positives that were prior included in the Markov blanket in the first stage.

Although the patterns of epistasis most frequently reported in publications involve SNPs with marginal effects (*e.g.* Crawford,L. *et al.*, 2017), our work addresses the epistasis detection problem defined as the situation where variables are not relevant for phenotype prediction when tested for association one after the other (pure epistasis). This restrictive situation is more complicated. Though, methods for pure epistasis detection are required since such patterns of epistasis were reported (Julià,A. *et al.*, 2008; Génin,E. *et al.*, 2013). Traditional Markov blanket-based techniques may fail to detect SNPs involved in epistatic interactions when such SNPs display little or no marginal effect. This is the reason why our heuristic combines Markov blanket learning with stochastic exploration of groups of SNPs, in a random forest-like fashion.

3 Methods

In this section, we first describe our proposed algorithm. Then we detail the statistic used to test conditional independence. This statistic is the fundamental measure ruling all decisions in our algorithm. Finally, we provide details about the method used to evaluate the power of SMMB on simulated data.

3.1. Stochastic Multiple Markov Blanket algorithm

The key idea of SMMB is to learn multiple suboptimal Markov blankets in order to obtain a Markov blanket consensus with high predictive performance. This principle is inspired from ensemble methods, *e.g.* random forests (Breiman,L., 2001), in which multiple weak learners are constructed and contribute to build the final learner by combining their contributions (Opitz,D. and Maclin,R., 1999; Jurek,A. *et al.*, 2014). Thus, it is often possible to learn simple learners while achieving great performance.

The SMMB algorithm is divided into two routines (Algorithms 1 and 2 in Figures 1 and 2, respectively). In each iteration of the top level routine, a Markov blanket is learned from *K* SNPs sampled from the complete dataset, where *K* is a user-specified parameter (Algorithm 1, lines 4 and 5). By default, the parameter *K* is set to $\sqrt{p}$, where *p* is the total number of SNPs in the dataset. This default setting is inspired from sampling parameters in Breiman's random forests (Breiman,L., 2001).

The second routine (learnMB, Algorithm 2) is the stochastic algorithm proposed to learn a Markov blanket **MB** taking into account potential epistatic interaction between SNPs. As in state-of-the-art algorithms, **MB** is initialized as an empty set (Algorithm 2, line 1). Then, a sampling is performed to obtain *k* SNPs out of the *K* SNPs previously drawn (Algorithm 2, line 4). The conditional association with the phenotype is then assessed for each of the $(2^k - 1)$ possible combinations within the set of *k* SNPs (Algorithm 2, line 5). When examining each such combination **s'**, candidate to inclusion in **MB**, function *assoc_score* (Algorithm 2, line 5) runs a series of conditional tests of independence as described in additional file 1 (Supplementary data), Algorithm 4. For each combination **s'**, function *assoc_score* examines the dependence between each variable X in **s'** and the phenotype, conditional on **MB** ∪ **(s'** \ **{**X**})**. In this way, combination **s'** is effectively considered as a group, as variables of **s'** are tested for dependence with the phenotype, conditional on the current MB enriched with the group **s'** but one variable. Indeed, if **s'** is added to **MB**, the conditioning set in *assoc_score* will become **MB** updated with **s'**. The statistical conditional independence test used in SMMB is the conditional G-test. The subset **s** that maximizes the conditional association is included in **MB** if the statistical independence (H_0) is rejected at type I error threshold α (Algorithm 2, line 6). Function $significant\_indep_{|MB}$ is briefly explained in additional file 1 (Supplementary data), Section Comments. In the G-test, support for conditional independence is provided based on the Chi-Squared distribution of the statistic under H_0. Once **s** has been included in **MB**, a backward phase is performed over **MB**, to discard false positives (Algorithm 2, lines 8 to 12). We must emphasize the fact that, for each variable X in **MB**, the backward phase successively considers subsets of **MB** \ {X}, as long as significant conditional dependence between X and T holds (Algorithm 2, lines 9 to 11).

Forward and backward steps are iterated until either the non-empty Markov blanket remains unchanged, or a maximal number of iterations, *m*, is reached while **MB** is still empty (Algorithm 2, line 15). The user-defined parameter *m* is used to remedy the following issue: the sampling of the set **S** of k variables (drawn from **X***) may not allow to identify a candidate subset **s** of **S** that can effectively be added to the MB. Thus, the

MB would be empty. To palliate this problem, SMMB coerces the exploration of subset **X*** as follows: whenever the MB stays empty, novel draws from **X*** are iterated (up to *m* draws). If, after the interleaved forward and backward steps, the resulting Markov blanket is no longer empty, it will be added to the list of Markov blankets (**MBs**) (Algorithm 1, line 6). The iterations continue until *r* Markov blankets are built or a maximal number of iterations, *t*, is reached. This number is considered as a parameter of the method. Motivation for parameter *t* is the following: owing to the sampling process, it might happen that SNPs irrelevant for epistasis detection (with regards to the phenotype of interest) be drawn at a given iteration of Algorithm 1. In this case, the resulting Markov blanket will be empty. As many SNP samples may lead to empty Markov blankets, a maximum number of iterations *t* must be set in Algorithm 1.

Once *r* Markov Blankets (at the most) are built, a consensus operation is performed over all learned Markov blankets (Algorithm 1, line 9). This operation consists in performing a union operation over all Markov blankets, followed by a backward phase on the resulting set of SNPs. The function *buildConsensus* is described in additional file 1 (Supplementary data), Algorithm 3. The backward phase in the consensus refinement follows the same scheme as in the backward phase previously explained for Algorithm 2. However, in function *buildConsensus*, correction for multiple testing is performed. This correction relies on adaptive permutations (Che,R. *et al.*, 2014). Additional file 2 (Supplementary data) briefly focuses on this aspect.

3.2. Statistical conditional independence test

The statistical test used in SMMB to test the independence between two variables, conditional on a set of variables, is the likelihood ratio test, also referred as the G-test. The conditional independence test is fundamental in SMMB: it allows statistical significance assessment for inclusions or removals of SNPs during the forward and backward phases of the Markov blanket learning approach. The conditional G-test (further defined) is used by function *significant_independence* in the backward phases of function learnMB (Algorithm 2, line 10) and of function buildConsensus (additional file 1, Supplementary data, Algorithm 3, line 4). In the forward phase, function *assoc_score* assesses the association between a group **s'** of SNPs and the phenotype, conditional on the current **MB** (Algorithm 2, line 5); for this purpose, *assoc_score* triggers a series of conditional (in)dependence tests: X ∈ **s**', X ⊥ T | **MB** ∪ (**s'** \ {X}) ?, as explained in additional file 1 (Supplementary data), Algorithm 4.

The G-test uses as the test statistic the logarithm of the ratio of two likelihoods related to the null (H_0) and alternative hypotheses. H_0 here refers to conditional independence. The G statistic is computed using equation (4), where *n* is the number of cells in the computed contingency table, O_i denotes the observed count in the i^{th} cell of the contingency table, and E_i denotes the expected count in the corresponding cell.

$$G = 2 \sum_{i=1 \text{ to } n,\, O_i \neq 0} O_i \, ln\left(\frac{O_i}{E_i}\right) \quad (4)$$

In this statistical test, expected counts are computed for fixed values of the conditioning set. For instance, to discover whether a SNP *s1* is independent from the phenotype conditional on set *S2*, the expected count of a contingency cell *ij* is computed following equation (5), where a value *k* is fixed for conditioning set *S2*:

$$E_{ijk} = \frac{O_{i\cdot k} O_{\cdot jk}}{O_{\cdot\cdot k}} \quad (5)$$

with *i* the value for SNP *s1*, and *j* the value for the phenotype.

Algorithm 1 Stochastic Multiple Markov blanket
SMMB($\mathbf{r}, \mathbf{t}, \mathbf{X}, \mathbf{T}, \mathbf{K}, \mathbf{k}, \mathbf{m}, \alpha$): **consensus_MB**

INPUT:
- **r**, maximal number of Markov blankets (|**MBs**|) output by the algorithm
- **t**, maximal number of iterations at top level,
- **X**, observed data for genotypes, data matrix of dimension $n * p$, with:
  - n, the number of individuals, and
  - p, the number of variables
- **T**, observed data for phenotype (vector of dimension n)
- **K**, total number of variables sampled from **X** to learn a Markov blanket (top level)
- **k**, number of variables sampled from **K** variables (inner level), $\mathbf{k} < \mathbf{K}$
- **m**, maximal number of resamplings (of **k** variables) as long as the Markov blanket remains empty
- α, type I error threshold

OUTPUT:
- **consensus_MB**, a consensus from all learned Markov blankets **MBs**, $|\mathbf{MBs}| \leq \mathbf{r}$

1: **MBs** ← ∅
2: $i \leftarrow 0$
3: **while** ($|\mathbf{MBs}| \leq r$ and $\mathrm{i} \leq t$)
4: $\mathbf{X}^* \leftarrow sampling\_without\_replacement(K, \mathbf{X})$
5: $\mathbf{MB}^* \leftarrow learnMB(\mathbf{X}^*, T, k, m, \alpha)$
6: **if not** $empty(\mathbf{MB}^*)$ **then** $add(\mathbf{MBs}, \mathbf{MB}^*)$ **end if**
7: $incr(i)$
8: **end while**
9: **consensus_MB** ← $buildConsensus(\mathbf{MBs}, \alpha)$
10: **return consensus_MB**

Figure 1. *SMMB algorithm. Algorithm 1 outlines how multiple Markov blankets are learned.*

Algorithm 2 learnMB($\mathbf{X}^*, \mathbf{T}, \mathbf{k}, \mathbf{m}, \boldsymbol{\alpha}$): **MB**

OUTPUT:
- **MB**, a Markov blanket, possibly empty

1: **MB** ← ∅ /*initialization of candidate Markov blanket*/
2: $i \leftarrow 0$
3: **repeat**
4: $\mathbf{S} \leftarrow sampling\_without\_replacement(\mathbf{k}, \mathbf{X}^*)$
/*Forward step*/
5: $\mathbf{s} \leftarrow argmax_{\mathbf{s}' \subseteq \mathbf{S}}\{assoc\_score(\mathbf{s}', T, \mathbf{MB})\}$
6: **if not** $significant\_indep_{|MB}(\mathbf{s}, T, \mathbf{MB}, \alpha)$ **then**
7: **MB** ← **MB** ∪ **s**
/*Backward step*/
8: **for each** $X \in \mathbf{MB}$
9: **for each** $\mathbf{S} \subseteq \mathbf{MB} \setminus \{X\}, \mathbf{S} \neq \emptyset$
10: **if**$(significant\_independence(X, T, \mathbf{S}, \alpha)$ **then MB** ← **MB** \ $\{X\}$; **break end if**
11: **end for**
12: **end for**
13: **end if**
14: $incr(i)$
15: **until** ((**not** $empty(\mathbf{MB})$) **and** (**MB** *does not change*)) **or** ($empty(\mathbf{MB})$ **and** $i = m$)
16: return **MB**

Figure 2. *SMMB algorithm. Algorithm 2 shows how each Markov blanket is learned.*

3.3. Evaluation of performance

Theoretically, in a GWAS epistasis detection problem, the discovery of a SNP combination that is truly associated with the disease is reported as a true positive (TP) result, the non-discovery of such a combination is reported as a false negative (FN) result. A true negative (TN) result corresponds to the case where there is effectively no discovery to report. In contrast, a false positive (FP) result corresponds to the situation where a discovery is reported, but should not be. We denote $\#D_T$ the total number of tests: $\#D_T = \#TN + \#FN + \#TP + \#FP$.

The power of an epistasis detection method is traditionally evaluated based on equation (6), where the denominator represents the total number of tests.

$$Power = \#\mathrm{TP} / \#\mathrm{D_T} \quad (6)$$

This classical criterion (equation (6)) is unbalanced since the more SNPs are being reported by a predictive algorithm, the greater is the chance to observe high power.

To obtain unbiased power evaluation, we considered the *F-measure* (equation 7), which combines recall or sensitivity (equation 8) and precision measure (equation 9). High recall means that mostly all truly associated SNP combinations are detected, but false positives may be detected as well. In contrast, high precision means that truly associated SNPs account for a large part of the overall detected SNPs.

$$F\text{-}measure = \frac{2}{1/recall + 1/precision} \quad (7)$$

$$recall = \#TP/(\#TP + \#FN) \quad (8)$$

$$precision = \#TP/(\#TP + \#FP). \quad (9)$$

Besides, the true negative rate (TNR) is defined as:

$$TNR = \#TN/(\#TN + \#FP). \quad (10)$$

4 Results

We first present the simulated disease models that were used to generate multiple datasets to compare performances over various methods. The corresponding results are then analyzed. The GWAS dataset used to provide an application of our algorithm on real biological data is depicted in the next subsection. Finally, the results obtained on this latter dataset are discussed.

4.1. Simulated datasets

The SMMB algorithm is first evaluated on simulated datasets with different disease models. A disease model is defined as the probability of being affected by the disease given a combination of SNPs. For a disease model, these probabilities are gathered in a penetrance table. A penetrance is denoted by $P(D \mid G_i)$, where D stands for the event "the individual is affected by the disease" and G_i indicates the i^{th} combination of SNPs. For each model, we generated 100 datasets. Each dataset contains 100 simulated genotyped markers for 2,000 cases and 2,000 controls. Two or three disease susceptibility SNPs were simulated, depending on the epistatic model chosen. The causal SNPs share the same minor allele frequency (MAF). For each non-causal SNP, the MAF was uniformly drawn from the [0.05, 0.5] interval.

The first two epistatic models used were recommended by Marchini and coworkers (Marchini,J. *et al*., 2005) and are widely used interaction models for performance comparison. We now provide a general description of the three models. In model 1, disease risk is increased when both SNPs present at least one disease susceptibility allele. Furthermore, disease risk increases with the number of disease susceptibility alleles at each locus. Model 1 is also referred to as the multiplicative model. In model 2, additional disease susceptibility allele at each locus does not further increase the disease risk. Model 2 is also referred to as the threshold model. In model 3, three disease loci jointly increase the disease risk. This latter model has already been used for performance comparison by Zhang and Liu (Zhang,Y. and Liu,J.S., 2015). For each model, we varied the common MAF of the causal SNPs in {0.05, 0.1, 0.2, 0.5}.

For these three models, the Java package GAMETES (version 2.1) was used to generate case-control simulated datasets (Urbanowicz,R.J. *et al*., 2012). The input parameters for GAMETES are the penetrance table, the total number of SNPs, the number of SNPs involved in the epistatic interaction, the MAF for each such SNP, the interval of MAFs for non-causal SNPs, the number of cases and the number of controls, the number of replicated datasets under these conditions. GAMETES automatically computes the heritability and the disease prevalence. Additional file 3 (Supplementary data) describes these models. Given the above constraints, we obtained a prevalence (P(D)) equal to 0.1. Heritability is $h^2 = 0.005$ in model 1, and $h^2 = 0.02$ in models 2 and 3.

We also simulated 100 datasets describing 4,000 individuals for 100 SNPs under the null hypothesis (absence of patterns of epistasis), to study the true negative rate.

Third, we simulated 100 datasets describing 4,000 individuals for 100 SNPs, each harboring 5 causal simulated SNPs in epistatic interaction. The 5-order interaction model (model 4) is described in additional file 3 (Supplementary data), together with the specific way to obtain it *via* GAMETES. For this model, the MAF common to the simulated interacting SNPs is equal to 0.2 whereas the other MAFs are in [0.05, 0.5], the prevalence (P(D)) is equal to 0.1, and the heritability is equal to 0.002.

4.2. Results on simulated datasets

The performance of SMMB was compared to those of three other methods, BEAM (Zhang,Y. and Liu,J.S., 2007), DASSO-MB (Han,B. et al., 2010) and AntEpiSeeker (Wang,Y. et al., 2010), all three dedicated to epistasis detection. For this purpose, we used the simulated datasets introduced in section 4.1. As SMMB, the reference method BEAM relies on a Bayesian framework, this time using Monte-Carlo Markov Chains (MCMC). We selected DASSO-MB since, as SMMB, it implements Markov blanket learning: the algorithm behind DASSO-MB is very similar to the IAMB algorithm. Finally, we selected a method coming from combinatorial optimization: AntEpiSeeker's strategy relies on ant colony optimization (Dorigo,M. and Stützle,T., 2003). Figure 3 shows the results for the three simulated epistatic models, when the six parameters of SMMB were fixed as follows: $t = 1000$, $r = 100$, $m = 30$, $K = 10$, $k = 3$, $\alpha = 0.05$. These results show that SMMB outperforms all other three methods in most simulated scenarii. Up to a MAF equal to 0.20, SMMB performs slightly better than the other methods for model 1 (Figure 3 a). Nevertheless, power remains poor for all four methods on this model. SMMB provides higher power than the other methods when model 2 is considered with MAFs equal to 0.10 and 0.50 (Figure 3 b). Besides, for this model, SMMB improves the reference software BEAM by 6 to 15% for all four MAFs considered. SMMB is ranked second behind DASSO-MB, for model 2, when MAF is minimized at 0.05 (Figure 3 b). Overall, for MAFs equal to 0.10 or more, the power of BEAM is smaller than those of SMMB, DASSO-MB and AntEpiSeeker. In contrast, BEAM performs slightly better than other methods when epistasis interaction occurs between three SNPs and minor allele frequency is low (Figure 2 c). Both DASSO-MB and AntEpiSeeker perform worse than SMMB and BEAM under the simulated three-way epistatic interaction model: SMMB always ranks second behind BEAM, except for MAF equal to 0.20, and shows a difference in power of 3% at the most with BEAM, while SMMB ranks first for a MAF equal to 0.20. These results suggest that SMMB and BEAM are more robust than DASSO-MB and AntEpiSeeker under three-way interaction models, and that SMMB performs equally or better than other methods when dealing with pairwise epistatic interaction models. Across the four methods compared, SMMB is the most robust method when no explicit order of interaction is assumed.

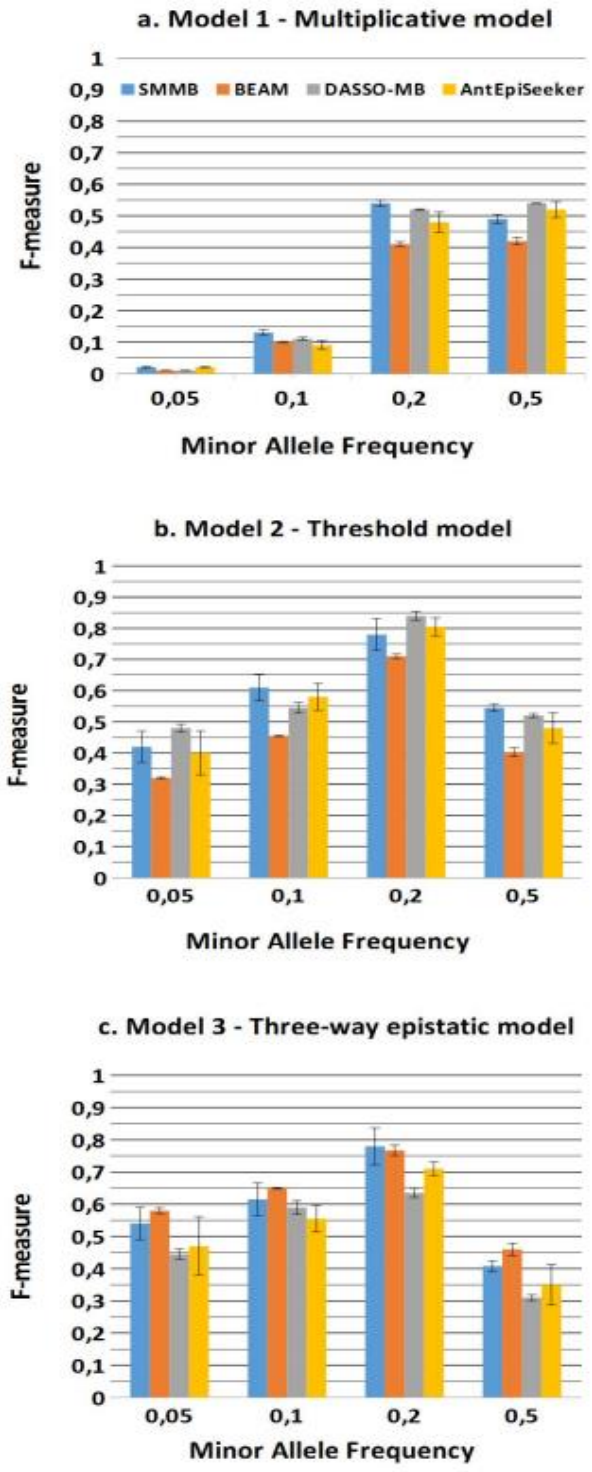

Figure 3. *Power comparison between SMMB, BEAM, DASSO-MB and AntEpiSeeker.*

Additional file 4 (Supplementary data) provides insights on the running times observed for the experiment described above. Fastest to slowest, one encounters DASSO-MB, SMMB, BEAM and AntEpiSeeker, with respective average running times of 5s, 30s, 93s and 469s (Model 1).

The results for the experiment under H_0 hypothesis are reported in additional file 5 (Supplementary data). The false positive rate greatly varies across the methods compared: with around 16%, SMMB ranks second behind BEAM (0%). DASSO-MB shows 21% of false positives. AntEpiSeeker output a false positive for each of the 100 simulated datasets.

When the simulated datasets harbor a 5-order epistatic interaction, all methods show practically null power, if not null (BEAM, DASSO-MB). Additional file 6 (Supplementary data) shows that SMMB is able to detect a subset (of size 2 to 4) for 5 simulated interactions. AntEpiSeeker detects a subset of size 2 for one simulated interaction. The difference between SMMB and AntEpiSeeker is that the former misses simulated interactions for half of the cases, whereas the latter always yields false interactions.

It is worth noting that existing published methods have been evaluated for 3-way epistasis models at the most, and have so far retrieved 3-way epistatic patterns at the best, on real GWAS datasets. An exception is the recently published method HiSeeker (Liu *et al.*, 2017), which was exceptionally tested for a 6-way epistatic model. However, this model was a model with *marginal* effect for each SNP, which does not correspond to the hypothesis SMMB deals with. Moreover, HiSeeker is a two-stage approach which first attempts to identify 2-way interactions having significant or intermediate association with the phenotype, to further extend them if possible. Thus HiSeeker tackles the problem of detecting embedded epistasis, whereas SMMB copes with pure epistasis detection.

4.3. Genome-wide real dataset

To run SMMB on large-scale datasets with a real genome-wide case-control study, we used the rheumatoid arthritis (RA) dataset provided by the Wellcome Trust Case Control Consortium (WTCCC, https://www.wtccc.org.uk/). This dataset consists of 469,612 SNP markers and 4,798 individuals (1,860 cases and 2,938 controls). Chromosomes were scanned separately.

4.4. Results on WTCCC GWAS dataset

More than 100 genetic susceptibility loci have now been identified for rheumatoid arthritis through standard single-SNP GWAS (Yarwood,A. *et al.*, 2016). In contrast, few works have investigated epistasis in RA (*e.g.* Julià,A. *et al.,* 2008; Wang,Y. *et al.*, 2010). We performed epistasis detection on WTCCC RA dataset using SMMB algorithm. A quality control phase based on specifications provided by the WTCCC Consortium was performed (The Wellcome Trust Case Control Consortium, 2007). SMMB was run with the following parameters : $t = 100$, $r = 4\times10^4$, $m = 30$, $K = 180$, $k = 3$, $\alpha = 0.05$. SMMB outputs a list of at most r "small" MBs learned by Algorithm 2, together with the refined consensus. To extract interactions from the consensus, SMMB proceeds as follows: if such "small" MB is included in the refined consensus, it is considered to be an interaction.

It took about 34h (average over 10 runs), for SMMB to handle the RA dataset, based on XEON biprocessors 5462 2.66 GHz, 6 cores (24 GB, 2 GB per core). This is a cumulative time for all 23 chromosomes. The running time for deterministic software DASSO-MB was 12h, and the running times for BEAM and AntEpiSeeker were 59h and 69h, respectively (average on 10 runs), in the same conditions. DASSO-MB is much faster, but performs poorly.

In contrast to an exhaustive approach, the stochastic feature of SMMB, BEAM and AntEpiSeeker helps to reduce the search space. However, it may lead to miss some interactions for some executions. For SMMB and AntEpiSeeker, it was necessary to run several executions to obtain seven 2-way interactions (see Table 1 in additional file 7, Supplementary data). Each run of BEAM identified these seven interactions. The deterministic method DASSO-MB only retrieved five interactions. None of the methods

yielded results outside this set of seven interactions. In 40% of the executions of SMMB (*i.e.* 4 out of 10 executions), all seven interactions were identified, in comparison to 50% for AntEpiSeeker, 0% for DASSO-MB, and 100% for BEAM. It has to be emphasized that over 10 runs, 70% of the runs output at least 6 of the 7 interactions for SMMB, versus 80% for AntEpiSeeker, which is on average twice as slow as SMMB. A well-known limitation of stochastic methods translates here in the impossibility to guarantee that the heuristics behind BEAM, SMMB or AntEpiSeeker will identify all true positives, given a dataset. Beyond detection power issues related to specific datasets, the restriction of the exploration of the search space may lead to missing true positives. Neither is it possible to specify a minimum number of executions guaranteeing the identification of all true interactions. As things stand, running several executions until no more interaction can be identified for an extra series of runs remains the only way to decrease the risk of missing true positives. In the case of the RA dataset, BEAM happens to show *a posteriori* the highest performance, but experiments on simulated datasets showed that BEAM is likely to miss true positives. In the recommended framework implementing several executions of the same stochastic software, SMMB is substantially faster than BEAM and AntEpiSeeker (respectively 1.73 and 2.02 times as fast).

The seven epistatic interactions detected are listed in additional file 7 (Supplementary data), Table 2. The p-value for the logistic regression of each epistatic pattern against the phenotype is smaller than 10^{-6}. None of the methods yielded results outside this set of seven interactions. In the case of the RA dataset, this comparison suggests excellent results regarding the false positive rates across the four methods.

5 Discussion

In this paper, we introduced a novel stochastic Markov blanket-based framework dedicated to the detection of epistatic interactions. SMMB successfully competes with state-of-the-art methods such as BEAM, AntEpiSeeker and DASSO-MB. The success of SMMB is due to the Markov blanket strategy combined with a stochastic feature to tackle the epistasis discovery when no marginal effect is displayed by causal variants. Our method was applied to a real GWAS dataset and we were able to recover the same results as those found by two other state-of-the-art software packages.

The algorithm SMMB is a variation of Markov blanket construction, with ensemble features. In ensemble methods, cross-validation allows to assert whether the training dataset is large enough, to avoid overfitting. In the context of GWAS, dealing with a too small cohort is likely to prevent generalization of the findings. If the purpose of the GWAS is not function study focused on the findings, but risk prediction, then performing a meta-analysis is considered the gold standard. SMMB will not escape this requirement.

Although the results show that our method performs well on simulated datasets and can be run on real GWAS datasets, some limitations remain. Indeed, despite the fact that the stochastic feature allows the analyst to identify epistatic interactions which involve SNPs that display no marginal effects, it raises the computational complexity and may lead to long execution times on real GWAS datasets. Though SMMB is about twice as fast as AntEpiSeeker and BEAM on the RA dataset, the user is compelled to consider epistasis one chromosome at a time, which hampers the detection of inter-chromosomal interactions. Furthermore, because of the non-deterministic nature of the algorithm, some of its executions on real GWAS data are shown to not detect some significant epistatic interactions while these interactions are correctly detected in other executions. This limitation is common to other stochastic methods (*e.g.* AntEpiSeeker). To decrease the risk of missing true positives, a recommendation is to run several executions of SMMB, which is affordable as SMMB is substantially faster than BEAM and AntEpiSeeker.

Another limitation appears when too many SNPs are gathered to build the consensus Markov blanket of SMMB. Indeed, when several thousands of SNPs have to be filtered in the final backward phase of the algorithm, the running time may be prohibitive. A theoretical upper bound for the number of tests in this phase is $O(q\ 2^q)$ (where q is an upper bound for the MB consensus size before filtering), and permutations are also performed during this phase. Nonetheless, in practice, it is observed that the size of the conditioning set (*i.e.* the MB consensus) rapidly decreases: generally, the first conditional test performed for a variable indicates that this variable must be discarded from the consensus MB. Thus the empirical complexity is instead close to $O(q\ e)$, where e denotes the complexity of the tests run on all permutations. Even if less time consuming *adaptive* permutations are run, the overall running time remains high.

Furthermore, all conditional independences are assessed through G-tests. When the number of observations is limited for a particular SNP combination, that type of test becomes non-relevant. This issue is often referred as the curse of dimensionality. As a consequence, potential true positive high-order epistatic interactions may be missed. One way to overcome this limitation could be to use permutations to assess the statistical significance of dependence between a SNP and the phenotype conditional on the Markov blanket, no matter the Markov blanket's size. However, permutations would lead to a very strong increase of the execution time that would not be affordable. In SMMB, the permutation approach is only used in the final backward phase (*i.e.* during the consensus Markov blanket construction) in order to apply the multiple test correction.

In this paper, we tackled the difficult subject of epistasis detection. There is still room for improvement in the situations where all the methods compared showed poor performances. In the light of this discussion, we intend to further improve SMMB in future studies, to best guide the exploration of the search space, and to reduce the complexity burden implied by the stochastic feature. In order to achieve this goal, we plan to guide the stochastic sampling procedure by incorporating an ant colony feature. In the near future, our directions of research will be investigating less computationally demanding strategies, to diminish the size of the MB consensus and refine it, together with accelerating MB stabilization during forward steps through ant colony optimization. In the same time, a reduction of the search space thanks to biological knowledge integration is also an interesting strategy which is still to be combine with ours.

Acknowledgements

The two co-first authors, CN and CS, would like to thank the anonymous reviewers for their constructive comments, which helped them to substantially improve the manuscript. This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of investigators who contributed to the generation of the data is available at https://www.wtccc.org.uk/. Part of the experiments were performed at the CCIPL (Centre de Calcul Intensif des Pays de la Loire). The development of the SMMB software is performed at the CCIPL.

Funding

Clément Niel is supported by the Regional Bioinformatics Research project GRIOTE granted by the Pays de la Loire region on the one hand, and the European Genomic Institute for Diabetes (EGID) Labex (Lille) on the

other hand. Ghislain Rocheleau's work is supported by a Chair in Biostatistics jointly sponsored by the Centre National de la Recherche Scientifique and Lille 2 University.

Conflict of Interest: none declared.

References

Abo Alchamlat,S. and Farnir,F. (2017) KNN-MDR: a learning approach for improving interactions mapping performances in genome wide association studies. *BMC Bioinformatics*, 18(1), 184.

Aliferis,C.F. *et al*. (2003) HITON: a novel Markov blanket algorithm for optimal variable selection. *Annual Symposium proceedings* / AMIA Symposium, 21-25.

Bateson,W. (1909) Mendel's principles of heredity. Cambridge, *UK: Cambridge University Press*.

Best,N.M. *et al.* (2016) nana plant2 encodes a maize ortholog of the Arabidopsis brassinosteroid biosynthesis gene DWARF1, identifying developmental interactions between brassinosteroids and gibberellins. *Plant. Physiol.*, 171(4), 2633-2647.

Breiman,L. (2001) Random forests. *Mach. Learn*., 45, 5-32.

Che,R. *et al.* (2014) An adaptive permutation approach for genome-wide association study: evaluation and recommendations for use. *BioData Min.*, 7, 9.

Chickering,D.M. *et al*. (2004) Large-sample learning of Bayesian networks is NP-hard. *J. Mach. Learn. Res*., 5, 1287-1330.

Crawford, L. *et al.* (2017) Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits. *PLOS Genet*.

Dorigo,M. and Stützle,T. (2003) The Ant Colony Optimization Metaheuristic: Algorithms, Applications, and Advances. *Handbook of Metaheuristics, International Series in Operations Research & Management Science*, 57, F. Glover and G. Kochenberger (eds.), 250-285.

Fisher,R.A. (1918) The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin*., 52, 399-433.

Goudey,B. *et al*. (2013) GWIS – model-free, fast and exhaustive search for epistatic interactions in case-control GWAS. *BMC Genomics*, 13 (Suppl.3): S10.

Hahn,L.W. *et al*. (2003) Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics*, 19, 376-382.

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. *et al.* (2010) A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinformatics*, 11(Suppl.3): S5.

He,S. *et al*. (2016) Genome-wide mapping and prediction suggests presence of local epistasis in a vast elite winter wheat populations adapted to Central Europe. *Theor. Appl. Genet.*, 1-13.

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

Julià,A. *et al.* (2008) Genome-wide association study of rheumatoid arthritis in the Spanish population: KLF12 as a risk locus for rheumatoid arthritis susceptibility. *Arthritis Rheum*., 58(8), 2275-2286.

Jurek,A. *et al.* (2014) A survey of commonly used ensemble-based classification techniques. *The Knowledge Engineering Review*, 29(5), 551-581.

Koller,D. and Sahami,M. (1996) Toward optimal feature selection. *In Proceedings of the 13th conference on machine learning* (Bari, Italy, July 3-6th, 1996). Morgan Kaufmann, San Fransisco, CA, 284-292.

Liu,J. *et al.* (2017) HiSeeker: detecting high-order SNP interactions based on pairwise SNP combinations. *Genes (Basel)*, 8(6), 153.

Maher,B. (2008) Personal genomes: the case of the missing heritability. *Nature*, 456, 18-21.

Manolio,T.A. *et al.* (2009). Finding the missing heritability of complex diseases. *Nature*, 461, 747-753.

Marchini,J. *et al*. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet*., 37, 413-417.

Matsubara,K. *et al*. (2015) Hybrid breakdown caused by epistasis-based recessive incompatibility in a cross of rice (Oryza sativa L.). *J. Hered*., 106(1), 113-122.

Niel,C. *et al.* (2015) A survey about methods dedicated to epistasis detection. *Front. Genet.*, 6, 285.

Opitz,D. and Maclin,R. (1999) Popular ensemble methods: an empirical study. *J. Artif. Intell. Res.*, 11, 169-198.

Peña,J.M. *et al*. (2005) Scalable, efficient and correct learning of Markov boundaries under the faithfulness assumption. *In Godo L. (eds) Symbolic and Quantitative Approaches to Reasoning with Uncertainty. Lect. Notes. Comput. Sc.*, 3571.

Press,M.O. and Queitsch C. (2017) Variability in a short tandem repeat mediates complex epistatic interactions in Arabidopsis thaliana. *Genetics*, 205(1), 455-464.

Schwarz,D.F *et al.* (2010) On safari to Random Jungle: a fast implementation of random forests for high-dimensional data. *Bioinformatics*, 26, 1752-1758.

Sun,Y., *et al* (2017) epiACO - a method for identifying epistasis based on ant colony optimization algorithm. *BioData Min.*, 10, 23.

The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447,661-678.

Tsamardinos,I. *et al.* (2003) Algorithms for large scale Markov blanket discovery. *In Proceedings of the the 16th International FLAIRS Conference* (St. Augustine, FL, May 11-15, 2003). AAAI Press, Menlo Park, CA, 376-380.

Urbanowicz,R.J. *et al*. (2012) GAMETES: a fast, direct algorithm for generating pure, strict, epistatic models with random architectures. *BioData Min.*, 5, 16.

Wang,Y. *et al*. (2010) AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Res. Notes*, 3, 117.

Wu,T.T *et al.* (2009) Genome-wide association analysis by lasso penalized logistic regression. Bioinformatics, 25(6), 714-721.

Yanlan,L. and Jiawei,L. (2012) An improved Markov blanket approach to detect SNPs-disease associations in case-control studies. *Int. J. Digit. Content Technol. Appl*., 6, 278-286.

Yaramakala,S. and Margaritis,D. (2005) Speculative Markov blanket discovery for optimal feature selection. *Fifth IEEE International Conference on Data Mining (ICDM)*.

Yarwood,A., *et al.* (2016) The genetics of rheumatoid arthritis: risk and protection in different stages of the evolution of RA. *Rheumatology*, 55,199-209.

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

Zhang,Y. and Liu,J.S. (2007) Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet*. 39, 1167-1173.