A survey about methods dedicated to epistasis detection

Clément Niel, Christine Sinoquet, Christian Dina, Ghislain Rocheleau

▶ **To cite this version:**

Clément Niel, Christine Sinoquet, Christian Dina, Ghislain Rocheleau. A survey about methods dedicated to epistasis detection. Frontiers in Genetics, 2015, 6 (Article 285), pp.19. 10.3389/fgene.2015.00285 . hal-01205577

HAL Id: hal-01205577
https://hal.science/hal-01205577v1

Submitted on 12 Jul 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

frontiers in Genetics

REVIEW
published: 10 September 2015
doi: 10.3389/fgene.2015.00285

A survey about methods dedicated to epistasis detection

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

[1] Computer Science Institute of Nantes-Atlantic (Lina), Centre National de la Recherche Scientifique UMR 6241, Ecole Polytechnique de l'Université de Nantes, Nantes, France, [2] Computer Science Institute of Nantes-Atlantic (Lina), Centre National de la Recherche Scientifique UMR 6241, University of Nantes, Nantes, France, [3] 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, [4] European Genomic Institute for Diabetes FR3508, Centre National de la Recherche Scientifique UMR 8199, Lille 2 University, Lille, France

During the past decade, findings of genome-wide association studies (GWAS) improved our knowledge and understanding of disease genetics. To date, thousands of SNPs have been associated with diseases and other complex traits. Statistical analysis typically looks for association between a phenotype and a SNP taken individually via single-locus tests. However, geneticists admit this is an oversimplified approach to tackle the complexity of underlying biological mechanisms. Interaction between SNPs, namely epistasis, must be considered. Unfortunately, epistasis detection gives rise to analytic challenges since analyzing every SNP combination is at present impractical at a genome-wide scale. In this review, we will present the main strategies recently proposed to detect epistatic interactions, along with their operating principle. Some of these methods are exhaustive, such as multifactor dimensionality reduction, likelihood ratio-based tests or receiver operating characteristic curve analysis; some are non-exhaustive, such as machine learning techniques (random forests, Bayesian networks) or combinatorial optimization approaches (ant colony optimization, computational evolution system).

Keywords: epistasis detection, genome-wide association study, complex disease, biological data mining, feature selection

OPEN ACCESS

Edited by:
David A. Rosenblueth, Universidad Nacional Autónoma de México, Mexico

Reviewed by:
Jingyi Jessica Li, University of California, Los Angeles, USA
Xiaodan Fan, The Chinese University of Hong Kong, Hong Kong

***Correspondence:**
Clément Niel, Computer Science Institute of Nantes-Atlantic (Lina), Centre National de la Recherche Scientifique UMR 6241, Ecole Polytechnique de l'Université de Nantes, Rue Christian Pauc, BP 50609, 44306 Nantes, France
clement.niel@univ-nantes.fr

Specialty section:
This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics

Received: *13 May 2015*
Accepted: *27 August 2015*
Published: *10 September 2015*

Citation:
Niel C, Sinoquet C, Dina C and Rocheleau G (2015) A survey about methods dedicated to epistasis detection. Front. Genet. 6:285. doi: 10.3389/fgene.2015.00285

Introduction

Genome-wide association studies (GWAS) have generated huge datasets in the past 8 years in order to find association between genetic polymorphisms and phenotypes. Individual risk prediction based on those discoveries was promising. Nevertheless, genetic architecture of complex diseases, such as type II diabetes, is still largely misunderstood (Vassy et al., 2014). Indeed, gene-environment and gene-gene interactions must be considered to better understand etiology of such phenotypes. In other words, various joint effects of genetic variations, namely epistasis, are likely to partly determine the disease state (Mackay and Moore, 2014). While common genome-wide association analysis checks for potential SNP-disease associations in a one-SNP-at-a-time fashion, looking for all potential epistatic interactions in such datasets will quickly result in combinatorial overload. This is why classical GWAS often left behind the daunting task of epistasis detection.

Several strategies came up to overcome the epistasis intricacy. After a first section dealing with epistasis generalities, we will present in this review the main categories of methods dedicated to epistasis detection. These methods are classified as follows. First, some exhaustive approaches for searching significant genetic marker combinations will be introduced. As some of these, like

Multi-Dimensional Reduction (MDR), are not manageable at a genome-wide scale, we will next turn our attention to filtering strategies which aim at reducing the size of the dataset, thereby decreasing the size of the search space. A final section will deal with machine learning and data mining techniques. This review does not intend to provide an exhaustive list of all software programs designed to find epistatic interactions, but rather to give an overview of the main categories of strategies put forward in the last 5 years.

Background—Epistasis

During the past decade GWAS have played a central role in the discovery of genotype-phenotype associations. In GWAS analyses, geneticists rely on DNA polymorphism markers to detect these associations. One of the most popular classes of genetic markers, Single Nucleotide Polymorphism (SNP), allows comparison of allelic frequencies between a sample of cases ascertained for a disease and a sample of controls. In the standard approach, SNPs are tested one by one for statistical association with the disease (Hirschhorn, 2009). Genetic variants are considered to have independent effects on the phenotype. As a result, only additive effects are considered under this approach. This kind of analysis has been widely used for years, but results are often not as appealing as expected. Indeed, with the "one locus at a time" strategy, only a little part of the genetic variance explains the phenotype, the remaining part being referred to "missing heritability" (Maher, 2008; Manolio et al., 2009).

It has been commonly admitted that missing heritability is partly due to genetic variants showing effects when they interact with one or more other variants (Eichler et al., 2010). Epistasis refers to the combinatorial effect of one or more genetic variants (**Figure 1**). These effects might interactively contribute besides existing marginal effects or they can also exist in absence of any marginal effect. In the last case, traditional statistical parametric methods will likely miss those interactions owing to the inflexibility of parametric models (Culverhouse et al., 2002; McKinney et al., 2006). For instance, in complex diseases like asthma (Howard et al., 2002), diabetes (Cho et al., 2004) or hypertension, additive genetic variation involves many SNPs, among which a vast majority have very small effect sizes (odds ratio less than 1.2, see **Box 1**) (Ritchie, 2015). As complex traits are poorly explained by additive models, one expects gene-environment or gene-gene interactions to substantially contribute to the genetics of these diseases.

Thus, epistasis detection has become an important field of research in human genetics: more complex models are studied nowadays, where combinations of genetic variants are examined for association with a trait. From a biological point of view, it seems unlikely that some phenotypes are only driven by genetic variants acting independently. For instance, large and complex networks of gene-gene and protein-protein interactions are well known in systems biology for their high connectivity, density and resistance to variation (Boone et al., 2007). Moreover, it has been observed that consequences of induced mutations are greatly variable in different genetic backgrounds (Mackay, 2014). Once aware of all this, it seems inconsistent to see gene-gene interactions as rare events.

Biological Epistasis and Statistical Epistasis

First, it is essential to distinguish biological epistasis (also called functional epistasis) from statistical epistasis (Cordell, 2002). The term biological epistasis was coined by Bateson (1909). In its original definition, it only involved allele effect at one locus concealed by the effect of another allele at a second locus. This can be seen as a broadening of the dominance concept at an inter-loci level. A more recent definition also allows genetic variant effects to be enhanced by effects of other genetic variants (Siemiatycki and Thomas, 1981). Generally, speaking, an epistatic effect exists when the effect of an allele at a genetic variant depends either on the presence or absence of another genetic variant.

On the other hand, statistical epistasis refers to the departure from additive effects of genetic variants at different loci with regard to their global contribution to the phenotype (Wang et al., 2010a). This definition was proposed by Fisher (1918). One relies on this definition when one wants to detect epistatic interactions with computational methods. Ultimately, the goal consists in interpreting interactions found to be statistically relevant in order to get closer to their biological definition and to apprehend the underlying functional mechanisms. This last step is undoubtedly the more difficult one (Moore and Williams, 2005) and is often disregarded.

A recent concrete example of epistasis has been described by Gertz et al. (2010), where three SNPs were shown to be involved in an epistatic interaction in yeast *Saccharomyces cerevisiae* (**Figure 2**). In the following, italic characters refer to the gene while normal characters refer to the corresponding protein. One SNP is located in the promoter region of *RME1* which encodes a transcription factor repressing the transcription of *IME1*, a gene coding for a transcription factor which promotes sporulation. State of this SNP influences the production rate of RME1. The second SNP is located in the promoter region of *IME1*. Its state affects the binding specificity of RME1-*IME1*. The third SNP lies in the coding region of *IME1* and its state conditions the binding specificity of IME1-kinase, which is the active form of IME1. Gertz and coworkers showed that the allele combination of these SNPs have a non-additive effect on the RME1-*IME1* binding and on the sporulation efficiency. Consequently, sporulation efficiency is partly ruled by epistasis. Many other cases of epistasis have been evidenced recently (Smith et al., 2014; Ellis et al., 2015; Huang et al., 2015; Liu et al., 2015; Matsubara et al., 2015).

Origin of Epistasis: an Evolutionary Point of View

Canalization is a theory proposed by Waddington (1942). It is based on a generally admitted assumption: natural selection maintains the majority of a population into a healthy condition. Thus, in response to genetic and environmental variations, phenotypic modifications are buffered. This is especially true for vital physiological levels, such as blood glucose or blood pressure. To this end, evolution has favored complex robust systems resistant to variations (Moore and Williams, 2009). A compelling argument in favor of this hypothesis is the redundancy rate in biological networks. This feature is well known in systems biology

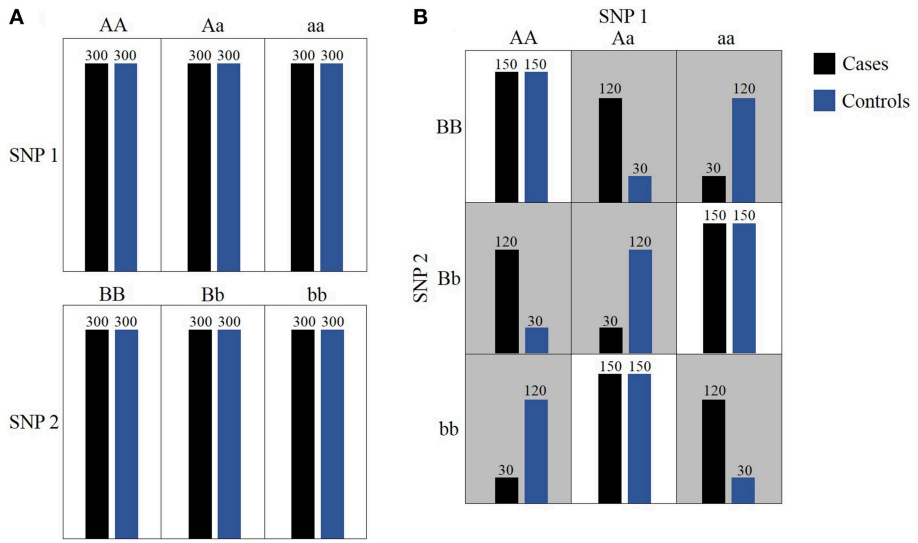

FIGURE 1 | Toy example of epistasis. (A) Neither SNP 1 nor SNP 2 presents a marginal effect. **(B)** In gray cells, allele combinations between SNP 1 and SNP 2 induce statistically significant epistatic effect on the phenotype distribution.

BOX 1 | Logistic regression and odds ratios.

A logistic regression model is a statistical model that depicts the relationship between a linear combination of variables (e.g., SNPs in a GWAS) and a binary trait, the disease phenotype (i.e., affected/unaffected status). The probability p of being affected is expressed in the log scale as:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

where x_1 and x_2 each correspond to the at-risk genetic variants, $x_1 x_2$ accounts for the interaction between them, and β_i are parameters being estimated from the data.

Odds ratios are highly related to logistic regression models. Indeed, $\exp(\beta_x)$ is an estimate of the odds ratio between the outcome and predictor variable x when values of other predictor variables are fixed. This is interesting because interpretation of odds ratios is intuitive. An odds is a measure related to probabilities. If an event has some non-null probability to occur in a particular experiment, odds for this event can be viewed has the ratio of the number of events to the number of non-events if the experiment were repeated multiple times. Thus, high odds correspond to high probability for this event, and *vice versa*. Given a probability p of occurrence for this event, an odds is defined as follows: $Odds = \frac{proportion\ of\ success}{proportion\ of\ failure} = \frac{p}{1-p}$.

An odds ratio (OR) is then simply the ratio of two odds. It evaluates association between disease occurrence and predictor variables. As such, this measure is closely related to statistical independence: if two variables (in the example below, SNP genotype and disease status) are statistically independent, their OR reduces to 1. Note that an OR not equal to 1 does not necessarily imply a statistically significant association.

Table 1 | Example of 2 × 3 frequency table to compute an allelic odds ratio.

		SNP genotype		
		AA	**Aa**	**aa**
Disease status	**Affected**	a	b	c
	Unaffected	d	e	f

Based on **Table 1** above, the odds ratio might be calculated using $OR = \frac{(2*a+b)/(2*d+e)}{(2*c+b)/(2*f+e)}$, assuming allele A is the at-risk allele. This OR is also called the allelic odds ratio (Sasieni, 1997).

where protein-protein interaction and gene-gene interaction networks exhibit redundant pathways making them resistant to variations (e.g., to deletion of a network node). A disease state would then be due to accumulation of mutations in the genetic network such that its robustness is outstripped. Therefore, all these network interactions are likely to involve epistatic effects. Canalization theory thus explains why so many variants only provide small contributions to the phenotype (Moore, 2003).

Challenges in Epistasis Detection

Challenges in epistasis detection are threefold. The first one is statistical. Statistical methods traditionally used in univariate SNP-phenotype associations are not adequate to find epistasis.

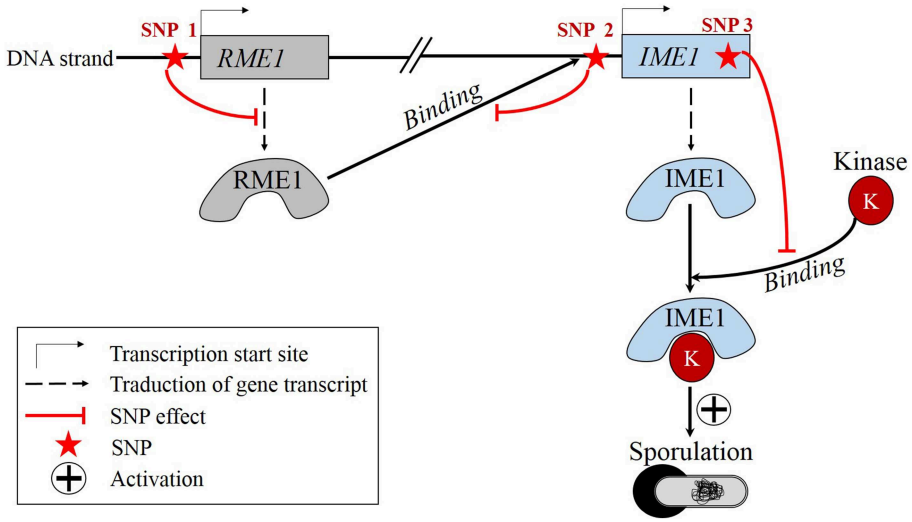

FIGURE 2 | Real example of epistasis: *S. cerevisiae* sporulation is regulated by epistatic effects among three SNPs. State of SNP 1 modulates the production rate of RME1. State of SNP 2 influences the binding specificity of RME1. State of SNP 3 conditions the binding specificity of IME1-kinase.

Finding epistatic interactions is a typical case of the *large p, small n* problem (Johnstone and Titterington, 2009). In practice, the aim is to balance the false-positive rate—produced by the astronomic number of tests performed—and the false-negative rate—a consequence of applying too much stringent significance thresholds. Moreover, SNPs involved in epistatic interactions may have very low minor allele frequencies (MAFs) whereas the number of variants to be tested might be huge. As a result, data is often sparse, leading to the so-called *curse of dimensionality*. The second challenge is computational. Though the overall complexity is linear with the number of individuals in the studied population, it becomes exponential when the interaction order increases. In 2-way interactions, this complexity corresponds to quadratic complexity. The number of combinations to be tested within a dataset containing 1 million SNPs is tremendous: 5×10^{11} pairwise interactions, 1.7×10^{17} 3-way interactions, 4.2×10^{22} 4-way interactions, 8.3×10^{27} 5-way interactions, and so on (Ritchie, 2015). Hence, an exhaustive search of epistatic interactions of order 3 or more would lead to a computational burden too prohibitive. Finally, the third challenge is the interpretation of the analytical results. To interpret statistical results biologically is not straightforward, for statistical interaction does not automatically entails interaction at the biological or mechanistic level (Cordell, 2002).

Exhaustive Search for Epistasis

In this section, we will discuss strategies of detection that exhaustively test all combinations of variants. Exhaustive search has been proposed to circumvent the local optimality problem, a drawback of heuristic techniques. Most exhaustive methods are designed to detect only pairwise interactions and those directed at higher order detection are simply not scalable. Despite their shortcomings, traditional parametric regression methods serve as a foundation in the field, as emphasized in the following subsection. Then, we will present a strategy derived from such regression methods and designed to be faster than traditional methods. Finally, we will discuss two model-free approaches.

Parametric Regression Methods

Traditionally, the most common framework for exploring GWAS data is parametric regression models. A parametric algorithm has a fixed number of parameters that has to be estimated from the data, and relies on strong assumptions about the probability distribution generating the data. This class of algorithms makes accurate predictions when those assumptions are sufficiently close to reality, but performs badly when proved incorrect. Logistic regression (see **Box 1**) has been widely used as a parametric method for exhaustive search of interactions in association analysis. For example, software PLINK (Purcell et al., 2007) has implemented logistic regression models to detect epistasis. But, in high dimensional data, parameter estimation is a costly and non-accurate procedure that introduces large standard errors because sample sizes are too small compared to genome-wide data size. As a consequence, many false positives are generated when dealing with such data. To overcome this problem, *p*-values are usually corrected with Bonferroni multiple-test correction (see **Box 2**). This correction being overly conservative, only interactions with very strong effects will be detected and many other interactions will be missed. Hence, the logistic regression strategy has been widely portrayed as unsuitable for handling genome-wide datasets (Cordell, 2009; Moore and Williams, 2009; Steen, 2012). Highly related to standard regression methods, penalized regression techniques, such as the LASSO (least absolute shrinkage and selection operator) or SCAD (smoothly clipped absolute deviation) gained some popularity to detect SNP-SNP interactions. However, those techniques are restricted to two-way interactions and are still

BOX 2 | Bonferroni correction.

Problem - Hypothesis-based statistical tests (e.g., *t*-test) are subject to false positive inflation when multiple tests are performed. For example, at a traditional 5% threshold set for statistical significance, there is a 5% chance to falsely reject the null hypothesis. Hence, if this test is performed 100 times when the null hypothesis is in fact true, and 5 tests are found to be statistically significant, then all 5 represent false positive associations. In this case, it is said that the risk is high and uncontrolled. This issue is known as the *problem of multiple tests*.

Answer - Bonferroni correction is applied to properly adjust the type I error rate. It consists in dividing the significance threshold by the total number of tests performed. For instance, if a study involves testing for 100 000 hypotheses at a desired global 5%significance level, the corrected significance level for each test is set at $\frac{0.05}{100\,000} = 5 \times 10^{-7}$.

Shortcoming - This method tends to reject non-null hypotheses due to its conservativeness. This conservative feature is also a shortcoming. It becomes inaccurate because it only favors strongly significant associations. As a result, many true positive associations will be missed (i.e., creating false negatives), thereby leading to a loss in statistical power.

prone to inflated false positive rate. Moreover, they are too computationally intensive to exhaustively search through all the pairwise interaction search space. In that case, feature selection techniques are required (further discussed in Section Two-stage Approach: Filters to Obtain Reduced Search Space). The interested reader is referred to Gou et al. (2014) for a recent detailed application of penalized regression-based approach for epistasis detection.

Bitwise Representation of Data and Likelihood Ratio-based Testing

We will introduce the Boolean operation-based testing and screening (BOOST) software program to exemplify this section. Designed to be fast, BOOST runs an exhaustive analysis of all potential pairwise SNP-SNP interactions (Wan et al., 2010). The main feature of BOOST is to build contingency tables and use them to calculate log-likelihood ratios for evaluating interaction effects. For two SNPs, a contingency table is a 3×3 matrix displaying the frequency distribution of all nine possible genotypes (**Figure 1B**). However, computing all potential contingency tables at a genome-wide scale is a time-consuming process. In fact, there are as many contingency tables as there are pairwise interactions to test (see Section Challenges in Epistasis Detection). In order to boost the procedure in terms of time and space efficiency, GWAS data is first transformed in a binary way. In usual data representation, each row symbolizes a SNP and each column symbolizes a subject (**Figure 3A**). In binary representation, each SNP is depicted by three rows, each of them describing the genotype status (i.e., 0, 1, or 2), and two columns depict cases and controls subjects respectively (**Figure 3B**). Each table cell contains a bit string where each bit represents one subject and its genotype: 1 if it corresponds to the genotype status encoded by the current row, 0 otherwise. Even if the binary matrix seems three times larger than the usual one, its space usage is smaller because one bit is an eighth of a byte, and bytes are the usual units (i.e., non binary) used for storing information. That representation also sticks closer to machine-language, which means that building a contingency table from it only involves fast bitwise (i.e., Boolean) operations.

Once contingency tables are constructed, the program is ready to test for pairwise interactions. The way to detect epistasis complies with Fisher's epistasis definition (see Section Biological Epistasis and Statistical Epistasis) since authors look for a difference between the independent effect model (i.e., marginal effects) and the model which includes both marginal and interaction effects. In other words, for each SNP pair, BOOST tests for a departure from the linear additive model. Under the assumption of equivalence between a logistic regression model and its corresponding log-linear model (Agresti, 2002), this departure is expressed in terms of log-likelihoods. However, the traditional log-likelihood of marginal effect model is constructed via computationally costly iterations that are not tractable at a genome-wide scale. Hence, authors use a non-iterative approximation of the log-likelihood ratio called Kirkwood superposition approximation (KSA) (Matsuda, 2000). On the basis of contingency tables, all pairwise interactions are tested with this indulgent KSA. As it is an approximation, too many false positives are deemed significant with respect to a threshold specified by the user. Therefore, after this first quick screening phase, interaction effects of the selected SNP pairs are again evaluated in a second phase. The number of SNP pairs is supposed to be reduced enough during the first phase in such a way that evaluation of interaction effects via a classical log-likelihood ratio on the remaining pairs is now affordable. Finally, significance of evaluated effects is assessed with a χ^2 test. One could say that the use of the χ^2 statistic discredits the method with the following argument: testing interaction effects of a SNP that shows high marginal effect with a χ^2 statistics may lead to evidence of a statistically significant epistatic effect while that perceived signal could solely be due to noise induced by high marginal effect. For instance, the latter issue has been reported in 2013 by Goudey and coworkers in their result section (Goudey et al., 2013). As a consequence, this phenomenon could favor the selection of many false positive interactions that have little to no epistatic effect. However, even if BOOST uses the χ^2 statistics to ultimately assess significance of epistatic interactions, tested SNP pairs already show significant association with a log-likelihood difference between the model which does not consider interactions (reduced model) and the model that does consider them (full model).

This approach is faster than its contemporary Bayesian method BEAM (see Section Bayesian Networks) and shows comparative power of detection. A year later, an even faster version that relies on graphic processing units (GPU) instead of central processing units (CPU) was developed. However, an important shortcoming arises because BOOST heavily relies on contingency table construction: low minor allele frequencies (MAF) generate sparse contingency tables, which hampers the detection power of BOOST. Indeed, in each cell of the contingency table, a minimal number of individuals is required so

A

	Controls			Cases		
	I1	I2	I3	I4	I5	I6
SNP 1	0	**1**	**1**	2	2	0
SNP 2	0	1	2	2	0	2
SNP 3	1	1	0	1	**2**	**2**

B

		Controls			Cases		
		I1	I2	I3	I4	I5	I6
SNP 1	= 0	1	0	0	0	0	1
	= 1	0	**1**	**1**	0	0	0
	= 2	0	0	0	1	1	0
SNP 2	= 0	1	0	0	0	1	0
	= 1	0	1	0	0	0	0
	= 2	0	0	1	1	0	1
SNP 3	= 0	0	0	1	0	0	0
	= 1	1	1	0	1	0	0
	= 2	0	0	0	0	**1**	**1**

FIGURE 3 | Representations of GWAS data. (A) Classical representation: cell (i, j) corresponds to status of SNP i for individual j. **(B)** Binary representation: cell (i, j) corresponds to the true (1) or false (0) assertion that a SNP i has a specific value (0, 1, or 2) for individual j. For ease of comprehension, the link between these two representions is highlighted in gray.

that the χ^2 test is statistically valid. But when contingency tables are sparse, this requirement is not met, thus leading to failure of epistatic interactions detection. Despite the fact that nearly all true positives are detected (i.e., the detection power is high), BOOST is sensitive to type I errors (Yoshida and Koike, 2011). Finally, a notable shortcoming is that the method only analyzes pairwise interactions and no higher order interactions.

ROC Curve Analysis

Goudey et al. introduced the genome-wide interaction search (GWIS) model-free approach in 2013 with the purpose of pairwise epistasis detection (Goudey et al., 2013). While BOOST compares a difference in segregation between two regression models, GWIS tests the difference in segregation power between a SNP pair and the corresponding SNPs taken individually. GWIS is not based on regression analysis, but exploits receiver operating characteristic (ROC) curves to test the discrimination power of SNP pairs. A ROC curve plots the true positive rate (i.e., sensitivity) against the false positive rate (i.e., 1 – specificity) of a classification model. In the context of GWAS, a ROC curve represents the performance of some model designed in classifying individuals according to their affected or unaffected status. For each pair of SNPs, GWIS considers three classification models and builds the respective ROC curves: two for each SNP taken individually, and one for the SNP pair. When the ROC curve corresponding to a SNP pair lies over the other two curves corresponding to individual SNPs, the SNP pair is said to have better prediction power than SNPs taken individually. The next question is to assess if the departure in prediction power between these classification models is significant. To answer this question, Goudey et al. proposed a model-free hypothesis test called *difference in sensitivity and specificity* (DSS). The goal is to quantify the gain in sensitivity and specificity of a ROC curve over another one (Goudey et al., 2013). It seems important to the authors to perform exhaustive search rather than heuristics, in order to avoid being trapped in local optima, then missing significant pairs. GWIS is also designed to be fast (e.g., faster than BOOST) and to scale up to datasets containing millions of SNPs.

The BOOST and GWIS strategies are designed to run exhaustive genome-wide fast scans of epistatic interactions. However, they are restricted to the detection of interacting SNP pairs, which is a substantial limitation. All epistatic models assuming interaction with order greater than two will be missed by these two methods. In the next section, we present a technique that overcomes this problem and exhaustively looks for higher order epistasis.

A Full Combinatorial Approach

Multifactor Dimensionality Reduction (MDR) is now a reference in the epistasis detection field. No parameters are estimated (i.e., nonparametric) and no assumptions are made on the genetic model (i.e., model-free) under this supervised classification approach. This strategy could detect interactions even when independent main effects are inexistent (Ritchie et al., 2001, 2003; Hahn et al., 2003). It is not constrained to identification of pairwise interactions but also searches for higher order interactions (Moore et al., 2006).

First, MDR partitions the dataset for cross-validation. By default, nine tenths of the dataset (training set) is used to build the model and the remaining tenth (testing set) is used to evaluate this model. The model is built following the steps presented in **Figure 4**. For an interaction order specified by the user, the corresponding number of SNPs is drawn (**Figure 4A**). Genotype combination counts are then distributed into a contingency table (**Figure 4B**). For instance, in a two-SNP biallelic interaction model, the nine possible two-locus genotype combinations are allotted into their respective table cells. For a three-SNP interaction model, twenty-seven table cells would

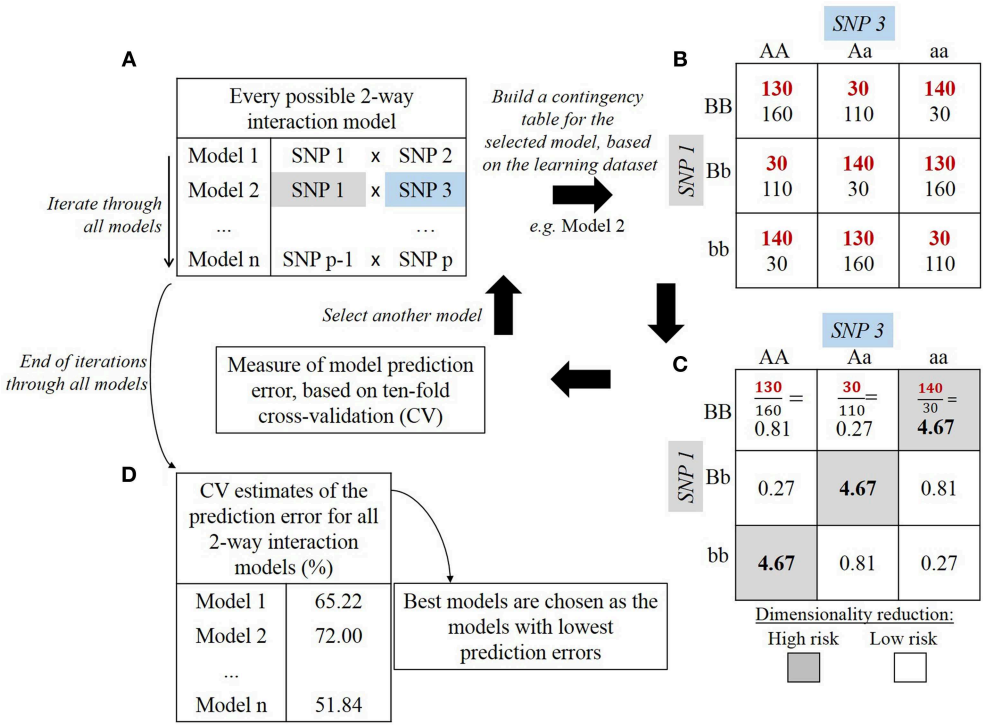

FIGURE 4 | Steps of multifactor dimensionality reduction (MDR) algorithm: example of 2-way interaction model. Description of one iteration of the cross-validation process. In **(A)**, a SNP combination is drawn among all potential SNP combinations. In **(B)**, numbers in red denote counts for cases whereas numbers in black denote counts for controls. In **(C)**, each cell displays the ratio of cases over controls. In **(D)**, the prediction error is estimated over the 10 iterations.

be needed. Then, the count of cases and controls is reported for each genotype combination and each cell is evaluated with the following ratio: $\frac{number\ of\ cases\ sharing\ this\ genotype\ combination}{number\ of\ controls\ sharing\ this\ genotype\ combination}$ (**Figure 4C**). This way, each genotype combination is classified either as high-risk if the above ratio lies beyond a specified threshold (e.g., 1.0), or as low-risk if it lies below that threshold (De et al., 2014). The classification model is then formed by merging cells marked high-risk in one group and all cells marked low-risk in another group. This explains why that method refers to "Dimensionality Reduction": starting with a problem where dimensionality equals the chosen interaction order, only one dimension remains in the end with high-risk and low-risk values. These steps are repeated for every possible combination of SNPs at a given interaction order, and each combination results in one prediction model. A 10-fold cross-validation process allows to assess the quality of such models. In other words, for each of the 10 iterations of the cross-validation, the models are trained to discriminate between low-risk and high-risk groups through the learning step (on nine tenths of the data). The proportion of ill-classified affected and unaffected individuals is evaluated on the testing set (one tenth of the data). Finally, the prediction error of each model is estimated over the 10 iterations (**Figure 4D**). The top best models over the 10-fold cross-validation are retained.

As the main feature of MDR is to reduce the data dimension, it can easily be combined with other classification methods (Moore and Andrews, 2015). This flexibility is also a good point to emphasize because since 2006, many extensions of MDR have been proposed so that it is applicable to quantitative traits (Gui et al., 2013). Besides, other variants of the MDR algorithm have been proposed that rely on parallel implementations to boost MDR computing time performance (Bush et al., 2006), to handle missing data (Namkung et al., 2009), or to implement permutation tests (Greene et al., 2009a). However, MDR remains a brute-force search algorithm that induces a prohibitive computational burden when the number of SNPs to analyze exceeds several hundreds. This lack of scalability is its most critical shortcoming in a genome-wide analysis context.

Most exhaustive strategies cannot afford screens of higher order interaction space search since they are not designed to scale up (Taylor and Ehrenreich, 2015). Even the aforementioned GWIS method is restricted to pairwise interaction detection. Exhaustive methods allowing exploration of higher order interactions, like MDR, cannot handle a genome-wide analysis and are constrained to several hundreds of SNPs. To overcome this shortcoming, a common technique is to preprocess data, reducing the entire SNP set to a smaller subgroup that has a tractable size for exhaustive higher order genetic interaction analysis. However, the type of filter is also important. Choosing a marginal-effect dependent filter would be indeed counterproductive with a method like MDR which is most effective in detecting interactions showing pure epistatic effects.

Two-stage Approach: Filters to Obtain Reduced Search Space

To address the computational burden issue, the overarching goal of some methods is to restrict the analysis to a small subset of candidate markers so that the exhaustive investigation of the remaining combinations is computationally tractable, even for higher order interactions. One approach is to conduct a single SNP-SNP analysis to keep only SNPs with significant marginal effects. SNP combinations are then tested among the remaining marker subset. For example, this strategy has been used in combination with stepwise logistic regression to pre-select a small fraction of SNPs (e.g., pre-determined 10%) based on single-SNP associations significance, before testing for interactions between the selected markers (Marchini et al., 2005). But such filtering leads to an obvious bias where epistatic interactions exclusively induced by combinatorial effects (i.e., with no marginal effect) are not picked up. Nevertheless, there are other ways to reduce the number of SNP combinations down to an informative subgroup. There also exists data mining and data integration techniques dedicated to filter and score downsized genetic variant sets, where null marginal effect is not a rejection condition. We will illustrate each technique in the next two subsections.

Filtering Based on Data Mining Techniques

We will illustrate this category with the ReliefF method. ReliefF approach consists in learning informative features from the dataset without any *a priori* knowledge (Robnik-Šikonja and Kononenko, 2003). The algorithm computes a proximity measure between individuals on the basis of genome-wide genetic similarity. The goal is to evaluate the quality of genetic variants according to how well their values distinguish individuals near to each other.

The algorithm is quite simple (**Figure 5**). For each individual (noted *I*), the procedure determines the nearest individuals (i.e., neighbors) sharing the same phenotype (set noted *S* for *same*), and also the nearest individuals that show up the opposite phenotype (set noted *O* for *opposite*). If *I* and *S* show different values for a marker, then this variant discriminates individuals having the same phenotype, thus decreasing its importance. On the contrary, if *I* and *O* show different values for a marker, this variant discriminates individuals having different phenotypes, thereby its importance is increased. These steps are then repeated over a predefined number of individuals. Moore and coworkers showed in 2007 that ReliefF algorithm is scalable (Moore and White, 2007).

The popularity of ReliefF gave rise to several variations (Kononenko, 1994) that we will quickly present below. RReliefF (Regressional ReliefF) was designed to study quantitative traits like eQTL epistasis (Huang et al., 2013). When applied to a genome-wide dataset, noisy genetic markers may be attributed too much weight, hence inflating their importance estimates. To alleviate this problem, TuRF (Tuned ReliefF) proposed to eliminate from the SNPs set considered for epistasis detection, SNPs with no or very low importance. These SNPs rarely discriminate individuals from their neighbors having a different phenotype (Moore and White, 2007). Importance of remaining SNPs is then re-estimated, without considering these noisy SNPs. Results are encouraging since TuRF power of detection is identical to or better than ReliefF. ECRF (Evaporate Cooling ReliefF) also attempts to solve the noisy variable problem (McKinney et al, 2007). It significantly outperforms ReliefF for detecting epistasis. Its algorithm combines information theory and ReliefF. In ReliefF and its above extensions, the user-defined number of nearest individuals to consider (i.e., *S* and *O*) is usually fixed at 10. Using such a predefined number may be considered as a selection bias since the information coded in the data is not fully exploited. To tackle this issue, SURF (Spatially Uniform ReliefF) proposes to take into account all neighbors within a given distance rather than a fixed number of neighbors (Greene et al., 2009b). SURF generally takes into consideration much more neighbors than ReliefF, labeling 25–50% of all individuals as neighbors. So when applied to a GWAS dataset, SURF has higher power of detection than ReliefF, albeit this may become a cumbersome procedure. A latest variation, SURF∗ (Greene et al., 2010), also considers information of farthest individuals to build importance scores. In terms of detection power of epistatic interactions, the performance of TuRF and ReliefF has been compared in Moore and White (2007). ECRF has also been compared to ReliefF in McKinney et al (2007). Finally, SURF has been compared to both ReliefF and TuRF in Greene et al. (2009b). However, ECRF and SURF have not been compared to each other, as well as ECRF and TuRF. ECRF and TuRF show improved performance over ReliefF, whereas SURF and SURF* show improved performance over both ReliefF and TuRF.

Filtering Based on Data-integration Techniques

Another research area advocates the use of knowledge from external databases, in order to select SNP groups that are relevant to the phenotype of interest (Grady et al., 2011). Even if this approach is hindered by a lack of epistasis understanding in complex organisms, it avoids the black box effect of data mining techniques that may hamper the interpretation of underlying biological mechanisms.

One way to do that is to query information in online public protein-protein interaction databases like IntAct (Kerrien et al., 2012), BioGRID (Chatr-Aryamontri et al., 2015), STRING (Franceschini et al., 2013) or ChEMBL (Willighagen et al., 2013). It is then possible to narrow all SNPs down to a reduced list of markers located in genes that encode for proteins involved in relevant interactions. When markers are mapped to an interacting gene pair, tests are exhaustively conducted on interactions between each SNP of the first gene against each SNP of the second gene. Unfortunately, one would probably fail at discovering new biological models by selecting SNPs in such a direct way. A more promising strategy is to come up with a score for each SNP (Ritchie, 2015), based on assessed relative importance of the proteins encoded by the genomic region encompassing the SNP. Novel findings are within reach by running a prioritization scheme rather than a strict removal (Pattin and Moore, 2008).

Resorting to pathways is also interesting. For instance, this approach has already been applied with information drawn

ReliefF algorithm

```
Input: GWAS dataset
       n: number of closest individuals to consider (with respect to the genotypic data)
       t: number of iteration

Output: Importances of all SNPs of GWAS dataset

For i=1 to t:
       Randomly select an individual I
       Find the n nearest neighbors S
       Find the n nearest neighbors O
       For each SNP in the GWAS data:
              Decrease SNP importance each time the SNP genotype differs between
                     I and a neighbor in S
              Increase SNP importance each time the SNP genotype differs between
                     I and a neighbor in O
       End for
End for
```

FIGURE 5 | ReliefF algorithm.

from pathways involved in lipid synthesis (Ma et al., 2015), by including evidence from public databases like KEGG Pathway (Kanehisa et al., 2012), Reactome (Croft et al., 2014) or BioCarta (Nishimura, 2001). For a pathway of interest, one first looks at the involved genes, and then maps SNPs to these genes. The technique is similar to the above protein interaction-guided analysis. But there is a bias as certain pathways are more deeply studied than others: genes (and SNPs therein) involved in a very well-known pathway may be given more weight than those involved in a less studied one. Instead of relying on guidance restricted to pathways or to protein-protein interactions, the comprehensive knowledge approach (Pendergrass et al., 2013a) is more global as it exploits pathways, protein interactions, gene expression, gene ontology, etc. As appealing as this approach might be, it is not currently possible to accurately evaluate results found by this strategy because implementing pathway simulations is not a trivial task. This would require a tool designed to simulate pathways and protein-protein interaction networks, and then simulate GWAS data where several SNPs are involved in these networks. Such a tool does not exist yet. Therefore, for this kind of filter based on comprehensive knowledge, we cannot properly and objectively assess its scientific relevance.

One software program worth mentioning is Biofilter. It gathers information from 13 databases (Pendergrass et al., 2013a), which contain experimental evidence of interaction, pathway or ontological similarity relationships. On the basis of biological plausibility, Biofilter models interactions that will be tested irrespective of the marginal effects. So it creates polygenic models, thanks to gene-disease and gene-gene connection knowledge (Pendergrass et al., 2013b). The statistical and computational challenges are also addressed since not all combinations of interactions are examined. Statistical relevance is based on the statement that the more two genes are involved in a relationship, the more likely they are to share an important biological link (Bush et al., 2009).

Although data-integration techniques yield meaningful and biologically relevant results, exploiting external information sources like pathways or protein-protein interaction networks is controversial. Online databases are incomplete and so is our understanding of biological pathways. Thus, making use of them to build filters would in most cases results in a flawed analysis. Moore and Hill recently recommended (Moore and Hill, 2015) to combine both the *biased* approach (from a biologist point of view) based on expert knowledge, and computational approaches solely driven by GWAS data (neither immune to bias from a statistician point of view). Similarly to computational exhaustive methods, this combined approach is taking advantage of artificial intelligence methods, which we discuss in the next section.

Non-exhaustive Searches Enhanced by Artificial Intelligence

Machine learning and combinatorial optimization represent alternatives to parametric statistical methods for detecting combinations of variants that are associated with a phenotype. Machine learning methods build non-parametric models to compile information further used for epistatic detection. Combinatorial optimization techniques consider a search space of solutions (i.e., combinations of potentially interacting SNPs) and browse through this space to find the more relevant combinations. Heuristics are commonly used in these algorithms, especially when dealing with genome-wide datasets in search of higher order genetic interactions. Identification of classification variables and interactions between them which allows outcome prediction is a well-known hurdle addressed by the machine learning and data mining fields of artificial intelligence (Cordell, 2009). In such non-parametric models, precautions must be taken to avoid overfitting (see **Box 3**). It has to be noted that if the model complexity of the underlying genetic mechanisms is too high compared to the sample size, using non-parametric methods

BOX 3 | Overfitting.

The aim of machine learning is to explain a system by learning a model with a training dataset. But dataset's particularities result in an overly tuned model adjusted for very specific features (Leinweber, 2007). In other words, overfitting happens when the training stage gives too much importance to the noise within data. Overfitting is detected when a simpler and more accurate model exists. However, identifying what to ignore in the overfitting model is a non-trivial task. Overfitting typically arises when model complexity is too high compared to the size of the training data. In practice, cross-validation possibly combined with pruning is used to avoid overfitting.

may not be affordable. In this case, parametric methods are the only practical alternative, assuming that the model assumptions are not severely violated.

A majority of these heuristics test for associations of variants allowing interactions, rather than testing for interactions themselves. The distinction lies in the following: besides SNPs involved in epistatic interactions, a model representing associations allowing for interactions also includes SNPs which have marginal effects. Therefore, although it is not a straight proof of epistasis, it is nonetheless an examination of polygenic models. Thus, if such procedures heavily rely on marginal effects for association findings, they will detect multiple SNPs with independent effect. But if they do not rely on marginal effects, they will also consider epistatic interactions.

With regard to machine learning techniques, we will first take a look at random forests and their variants, then move on to Bayesian network-based strategies. As for combinatorial optimization strategies, ant colony optimization and computational evolution system approaches will be presented.

Random Forests and their Variants

A tree-based algorithm generates a tree where each tree-node represents a predictor variable and a path designates a sequence of predictor variables from the root to the leaves of the tree. When the tree is constructed from GWAS data, each node represents a SNP. A basic tree-growing algorithm is deterministic in that each step looks for the predictor variable that optimally segregates the population. So a grown tree is a classifier which represents a SNP set allowing prediction of the phenotype of interest. This approach can handle SNPs that are associated in a non-linear way, dealing with interactions encoded in a hierarchical fashion between layers of the tree. A notable shortcoming of tree-based methods is that they are quite dependent of marginal effects. At the beginning of the tree learning step, the algorithm looks for a single SNP that well discriminates cases from controls. In practice, this is equivalent to looking for SNPs with high marginal effects.

Random forests were designed to avoid bias generated by growing a single tree. The random forest strategy creates multiple—generally thousands—classification or regression trees (e.g., CART) in order to apply an ensemble procedure. An ensemble procedure aggregates the predictions of all trees to produce a powerful and robust prediction tool (Breiman, 2001). The SNP set output is defined as the most important variable set of the random forest (to be further explained in this section). Although growing a random forest is a relatively computationally intensive procedure, it has been evaluated as a good strategy for detecting the most predictive SNPs in large-scale association studies (Bureau et al., 2005) and was applied to GWAS several times in the last 5 years with epiForest (Jiang et al., 2009), random Jungle (Schwarz et al., 2010) and SNPInterforest (Yoshida and Koike, 2011).

A classification tree is grown using the following steps (Jiang et al., 2009). First, a bootstrap sampling is performed from the GWAS dataset comprised of N individuals and M SNPs. It consists in randomly selecting, with replacement, N individuals from the N original individuals. Individuals not drawn are called out-of-the-bag (OOB) individuals. So a new dataset and an OOB set are created for each grown tree. Then a random feature selection is applied to construct each node of the tree. To do so, instead of considering all variables from the initial GWAS dataset, a random subset of variables is picked out without replacement. A recursive data splitting procedure is next executed, such that a parent node results in two child nodes given a rule that leads to a better discrimination of the current set of individuals (from the parent node) with regards to the disease status. This discrimination score is measured as a goodness of split or a decrease in impurity Δi. The tree is then grown up to its largest extent. These previous steps are repeated until a forest is built (**Figure 6**).

For each node, a so-called variable importance is assessed to evaluate its contribution to the trait either individually or via multi-way interactions with other predictor markers. In other words, variable importance represents weight approximating the causal effect of a predictor variable. There are several ways to measure variable importance (Schwarz et al., 2010). One is the *Gini importance*, a second one is the *permutation importance*, and a third one is the *conditional variable importance*, based on permutation importance. The conditional variable importance seems to be more suitable when applied to genetic data while the other two are biased in presence of linkage disequilibrium (correlation between SNPs) (Strobl et al., 2008). Compared to the original random forest construction, algorithms readjusted for epistasis detection include multiple SNPs at each tree-node during tree building (Botta et al., 2014). It is intended to detect SNP combinations even when marginal effects are very weak or inexistent (Yoshida and Koike, 2011). The readjusted method is less sensitive to SNPs presenting little marginal effects than an exhaustive approach like MDR. However, even if random forests reveal associations potentially pointing at interactions, they cannot make a distinction between a scenario of interacting SNPs and a scenario of several independent SNPs additively contributing to the phenotype. As a result, random forests are lacking clear interpretation.

More recently, another tree assembling software program was developed: GWGGI (Wei and Lu, 2014). It differs from the previous methods in two points. First, it uses a tree-growing algorithm which is more computationally efficient (Lu et al., 2012): the standard variable selection procedure is replaced with

A **Random forest algorithm**

```
Input: GWAS dataset with M SNPs to analyze
       t: number of trees to build

Output: An ensemble of trees

For i=1 to t:

      - Select a bootstrap sample from the GWAS dataset

      - Grow a tree from the bootstrapped data,
        by repeating the following steps for
        each terminal node of the tree:

            1. Select a random subset of variables
               from the bootstrap sample

            2. Find the variable that best splits
               the data into homogeneous groups

            3. Split the terminal node into two child nodes

End for
```

B **Toy example of the three steps to execute to grow a tree**

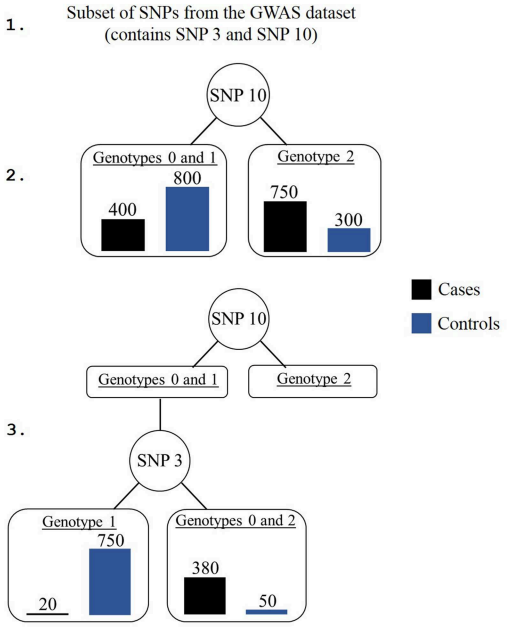

FIGURE 6 | Random forest algorithm. (A) Algorithm of a random forest procedure. **(B)** An example of the three steps needed to grow one tree.

a forward algorithm. The principle of a forward algorithm is to take into account previously selected variables. The novel variable identified is the one, when added to the previous set of variables, allowing for the most accurate prediction. Secondly, the GWGGI algorithm relies on likelihood ratios and the Mann-Whitney statistics to assess the predictors' importance in order to facilitate the statistical significance assessment of selected association models. Since each tree can be considered as a multi-locus genotype model, each individual is confronted to each grown tree and a likelihood ratio is generated: $LR_i^t = \frac{P(G_i^t|D)}{P(G_i^t|\overline{D})}$ where G_i^t is the genotype of individual i mapped t, and D (*resp.*$\overline{D}$) is the control status (*resp.* case status). Then for each individual, all likelihood ratios are assembled into a unique one by averaging the total number of trees. Finally, a U-statistic is constructed with comparisons between assembled likelihood ratios of cases vs. controls in order to evaluate the joint association of the selected SNPs with the phenotype (Wei et al., 2013). The U-statistic is calculated in the following way: $U = \frac{\sum_{i=1}^{N\_cases}\sum_{j=1}^{N\_controls}\psi(LR_i, LR_j)}{N\_cases * N\_controls}$. The ψ function is a kernel function defined as:

$$\psi\left(LR_i, LR_j\right) = \begin{cases} 1 & if\ LR_i > LR_j \\ 0.5 & if\ LR_i = LR_j \\ 0 & if\ LR_i < LR_j \end{cases}$$

The null hypothesis states that there is no association between the selected SNPs and the phenotype.

Bayesian Networks

Bayesian networks provide a compact representation of dependencies between variables. A Bayesian network consists of two components: a graphical one and a probabilistic one. In the former—directed acyclic graph (DAG)—variables are represented by nodes and dependencies between them are represented by directed edges. The probabilistic component of a Bayesian network associates a probability distribution with each node of the DAG, thus accounting for uncertainty. A Bayesian network encodes the Markov property: each variable is independent of its non-descendants, given its parents in the DAG. The governing theorem of a Bayesian network is the following. Let X, Y, and Z be variables of the Bayesian network. If $P(X|Y, Z) = P(X|Y)$, then X is conditionally independent of Z, given Y (noted $X \perp Z|Y$). When applied to genetic data, variables are typically SNPs and phenotypic values. Bayesian networks offer an appealing and intuitive way to capture relationships existing between genetic markers and disease status. The structure learning of a Bayesian network amounts to a model selection problem. Because this learning is an NP-hard problem (Chickering et al., 2004), specific techniques have to be used to reduce the computational burden.

A famous Bayesian network-based software program called BEAM (Bayesian Epistasis Association Mapping) (Zhang and Liu, 2007) is often used as a Bayesian-based reference for performance comparisons. BEAM relies on a Markov Chain Monte Carlo (MCMC) algorithm to test iteratively each marker, conditional on the current status of other markers. For each marker, the algorithm outputs its posterior probability of association with disease. Markers are then distributed into three groups: group 0 for markers unlinked with the phenotype, group 1 for SNPs that contribute independently to the phenotype (additive model) and group 2 for SNPs that influence the disease risk given particular allele combinations (epistasis model). After

that partitioning phase, a B-statistic is used to further filter detected SNP groups. When the BEAM method was originally published, the B-statistic was a new alternative to the usual χ^2 test of association between a phenotype and a set of SNPs. A detailed explanation of its computation would require a much deeper presentation of BEAM, which is not the aim of this section. The interested reader is referred to Zhang and Liu (2007) for a comprehensive explanation of how to build a B-statistic. Although the B-statistic enables to get rid of expensive permutation tests, MCMC iterations make this method inadequate when handling datasets containing more than 500,000 genetic markers, which is now commonplace in GWAS studies.

More recently, Han et al. (2012) also worked with Bayesian networks to capture SNP-disease associations with EpiBN. As these authors consider that SNPs are causal with respect to the phenotype, the Bayesian network built here is composed of two layers: one layer with the phenotype as a unique node, connected to parent nodes of the phenotype in the second layer which represents the SNPs associated with the phenotype. Edges between nodes representing SNPs can exist, thus allowing detection of interactions between genetic variants in the model. Instead of a MCMC-based algorithm, they use a Branch-and-Bound iterative procedure to learn the structure of the Bayesian network. At each iteration, the algorithm adds, deletes or reverses an edge. Then a score function is called to find the best network structure evolution since the previous iteration. The network is iteratively constructed and at each iteration, the current network structure goodness is assessed with a score function. The goal is to maximize this score. The score function is made of two terms that indicate how well the current structure fits the data—on the basis of a maximum likelihood ratio—and how complex the Bayesian network is. In Han et al. (2012), it has been shown through multiple simulations that the EpiBN software program seems to outperform BEAM in interactions detection power.

A different but not less appealing Bayesian strategy is the Markov blanket-based method. It allows discovery of SNPs in the local pathway of the phenotype, also referred to as "local causal SNPs" (Alekseyenko et al., 2011). In the context of GWAS, this strategy is used to avoid the time-consuming training processes like tree-growing of random forests or structure learning of a full Bayesian network. The principle is to find a minimal set of variables that completely shield the disease status from all other variables, thus resulting in a local Bayesian network fraction that borders the phenotype node in the graph: this set is defined as the Markov blanket. In other words, each SNP will be statistically independent of the case-control status when conditioned on the SNPs forming the Markov Blanket. A Markov blanket-based strategy can be applied for causal findings because the Markov Blanket contains direct causal variables (i.e., parent nodes), direct effect variables (i.e., child nodes), and direct causal variables of direct effect variables (i.e., spouses) (**Figure 7A**).

With the goal of finding a minimal SNP set, this strategy is expected to minimize the number of false positives. Besides its classification accuracy, this strategy has been put forward for its compactness (Aliferis et al., 2010a). Moreover, the Markov blanket-based strategy has proved to properly address the combinatorial hurdle raised by epistasis analysis at the GWAS scale (Aliferis et al., 2010b). The Markov blanket construction algorithm will generally go through two stages (**Figure 7B**). The first one, called "forward phase," adds new relevant variables to the candidate Markov blanket (noted *canMB*). In practice, this stage consists in finding the SNP X which is the most associated with the phenotype, given *canMB* (e.g., tested with a G^2 test, which is a subclass of likelihood-ratio tests and is similar to a χ^2 test, McDonald, 2014), and including X in *canMB* if X is dependent of the phenotype, given *canMB* (e.g., if the G^2 statistics is lower than some user-specified threshold): $\neg(X \; Phenotype \mid canMB) \Rightarrow add \; X \; in \; canMB$. This operation is repeated until *canMB* no longer changes from one iteration to the other. The second phase, called "backward phase," aims at removing false positives that were included in the previous step. To achieve it, each SNP of the candidate Markov blanket is checked. A SNP Y is detected as a false positive if it is independent of the phenotype given a SNP subset of *canMB*. Three implementations of this approach were recently developed: DASSO-MB (Han et al., 2010), TIE* (Alekseyenko et al., 2011; Statnikov et al., 2013) and IMBED (Yanlan and Jiawei, 2012), and all proved to be more sample-efficient than BEAM, i.e., less samples are needed to reach the same power of detection as BEAM. In DASSO-MB (Han et al., 2010, Han and coworkers postulate that, in epistatic interaction studies, only causal SNPs are sought, and consequently only parent nodes of the phenotype have to be detected. Hence, DASSO-MB represents a more specific application of the Markov Blanket approach. Considering a set of 19 SNPs already known to be associated with rheumatoid arthritis, an application of TIE* (Target Information Equivalency) showed that a Markov blanket-based approach could make the whole SNP set independent of the phenotype when conditioned on three other SNPs identified in the Markov blanket (Alekseyenko et al., 2011). In other words, the reported SNP set does not provide any predictive information about the disease status beyond that brought by the three SNPs identified with the Markov blanket.

The bias of this approach is that the first SNP added to the candidate Markov blanket is picked on the basis of a univariate test. So the detection of marker combinations when marginal effects are slight or nonexistent is still a major obstacle (Han and Chen, 2011). Markov blanket-based strategies also heavily rely on the *faithfulness* assumption, defined with respect to the sample, as follows: every conditional independence in the Bayesian network also exists in the probability distribution of the variables. In practice, this hypothesis is rarely met in GWAS.

Ant Colony Optimization

Ants communicate with each other through pheromone levels to find the optimal path leading to food. If an ant finds a shorter path, it will produce and increase the pheromone concentration along this path. Other ants will more likely follow that path showing increased pheromone concentrations, thereby creating a positive feedback to find the best path to food. In 2010, AntEpiSeeker algorithm (Wang et al., 2010b) was derived from the generic ant colony optimization (Dorigo and Gambardella, 1997) (ACO) algorithm. AntEpiSeeker performs the search of

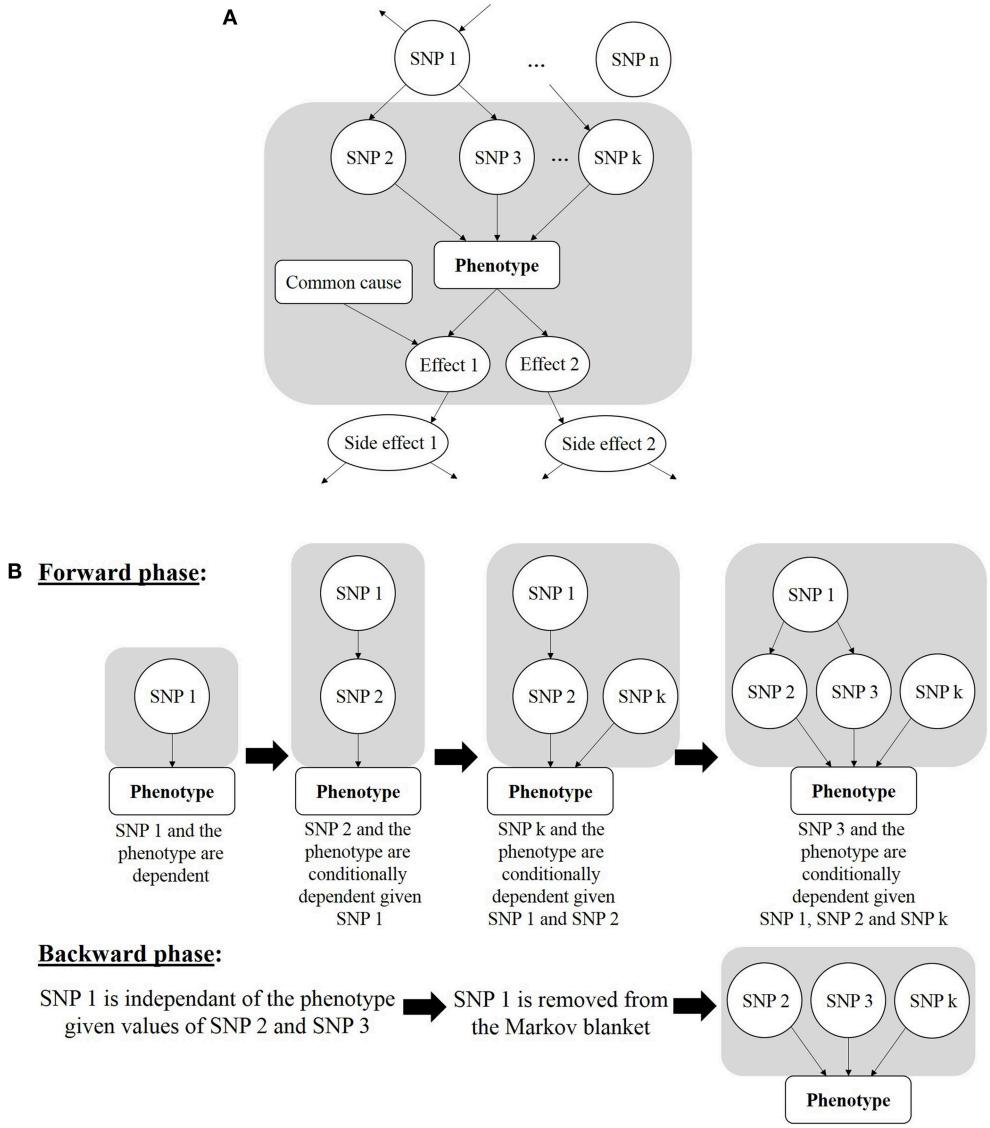

FIGURE 7 | (A) Markov blanket of a phenotype, in gray area. It is made of the parents (SNP 2, SNP 3, and SNP k), of the children (Effect 1 and Effect 2) and of the spouses (Common cause of Effect 1 with respect to the Phenotype). **(B)** Two stages of Markov blanket learning. For ease of reading, the Markov blanket is reduced to parents from **(A)**.

multiple groups of SNPs associated with the disease in parallel. The algorithm is an iterative procedure where artificial ants cooperate at each iteration to update knowledge about the propensity of SNPs to be related to the disease (**Figure 8**). From a computational point of view, ants represent SNP sets that have potential epistatic effects, and a pheromone concentration is a weight evaluated by epistatic interaction significance of the selected set of SNPs. Communication between ants is mimicked by a probability distribution function (PDF) shared by all ants. The PDF is a function describing the probability of selecting a specific SNP at a specific iteration. This probability depends on the pheromone concentration for this SNP at this iteration, and on another factor which allows to weight SNPs according to expert knowledge drawn from additional biological data. At each iteration, multiple SNPs are picked up, depending on the PDF, to build each ant. Then a χ^2 test is used as a score function to measure the association between an ant and the phenotype. Results are used to update the PDF for the next iteration. Once highly suspected sets of SNPs are assembled, AntEpiSeeker conducts a second analysis stage: an exhaustive search of epistasis within each built ant is performed, as well as within the set of SNPs that have the highest pheromone levels. The ant colony strategy was also exploited more recently in MACOED (Jing and Shen, 2015).

The positive feedback effect represents an interesting feature of the algorithm. Unfortunately, many parameters require fine tuning, like the number of iterations, the order of interactions, the number of SNPs in each ant, or the evaporation rate of

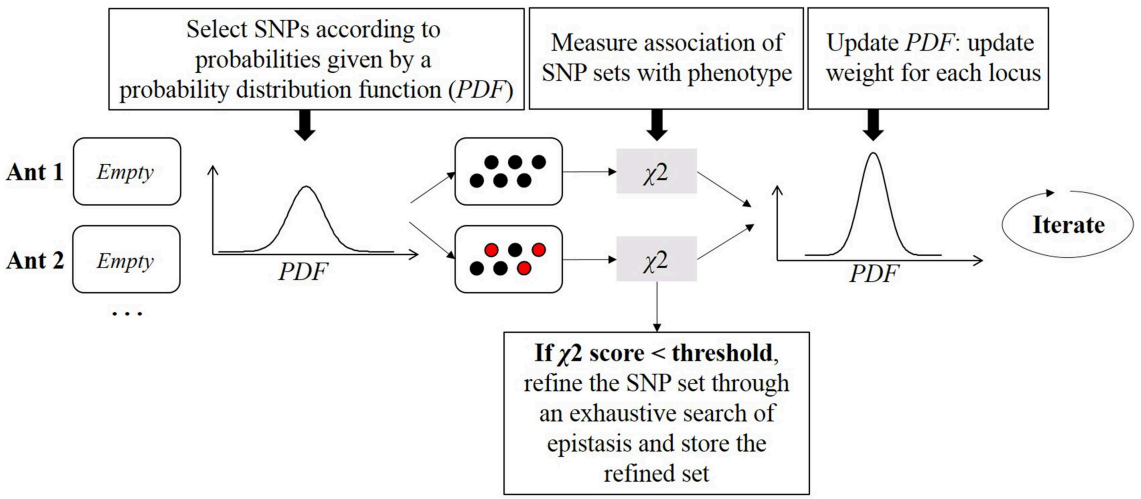

FIGURE 8 | Ant colony optimization procedure. For each ant, multiple SNPs are drawn. The probability distribution function (PDF) gives the probability of each SNP to be drawn. Once an ant is filled with a SNP set, joint association of this SNP set with the phenotype is evaluated with a χ^2 test. For each ant, the PDF is updated according to *p*-values resulting from χ^2 tests, such that SNPs efficiently classifying individuals will have a higher probability of being drawn in the next iteration.

pheromones which is an ingredient of the update function of the PDF. Those parameters must be estimated a priori, which is considered as a limitation of this algorithm.

Computational Evolution System

The algorithm behind the Computational Evolution System (CES) is an original strategy based on natural selection and Darwinian evolution. The goal is to grow a computer program from several basic building blocks, similar to a DNA strand emerging from a composition of the four basic nucleotides. This program tries to reproduce the natural evolution process underlying complex real biological systems. The first question is what the building blocks are, whenever one wants to build such a computer program. The answer is non-trivial and is decisive in epistasis analysis when trying to avoid dependence to marginal effects. In a recent application of CES (Moore and Hill, 2015), the building blocks were defined as basic functions involving SNPs. A basic function is an operator (add, delete, and copy) aggregating SNPs in combinations, and the resulting composition of building blocks is called a solution. In other words, a solution can be perceived as a set composed of various elements, where each element is a function dealing with genetic polymorphisms. A solution is thus a classifier designed to predict the case-control status of an individual given its genotype.

A CES is governed by a pyramidal architecture where each level is probabilistically controlled by its upper layer. The lowest level is a two-dimensional grid of solutions where each solution is a list of building blocks. The second level is a grid of solution operators influencing the lower layer. Each cell consists of a combination of add, delete, and copy operators having a given probability of being executed. Attributes can be added, deleted or copied either randomly or using expert knowledge. A third level of computation is used to introduce changes in execution probabilities of the latter operators. A last level controls the variation rate of the third layer. Uncertainty is injected in this architecture in order to mimic a realistic natural evolution system. As a result, there is high flexibility in model creation based on CES.

The stage during which all solutions are modified is called a generation. From one generation to the next, accuracy of each solution is modified as follows: an operator is drawn according to the execution probability distribution; this operator is then applied to each solution. It has to be noted that the initialization of the CES grid of solutions is either random or guided with expert knowledge. This last option has been highly recommended (Greene et al., 2009a; Payne et al., 2010). The accuracy of each solution is assessed in the following way. Each solution is applied to case and control individuals to obtain two distinct score distributions: one for cases and one for controls. A threshold is determined as the arithmetic mean between the medians of the two distributions. Then individuals are predicted to be cases or controls given this threshold. The solution accuracy is computed afterwards as an error ratio between predicted and actual status. Once one knows how to compare solutions, one can select the optimal solution which maximizes the prediction accuracy. The solution is selected among all generations (e.g., 1000 generations) in the following way. Each solution occupies a lattice position in the two-dimensional grid and competes with its neighborhood composed of eight adjacent solutions. Within this neighborhood, the solution with the highest accuracy is selected to replace the central position of that neighborhood.

This approach is interesting in that it allows modeling of complex interactions with few hypotheses. It also has the capability to use expert knowledge, and is well suited for parallelization. However, the computational complexity of the CES strategy precludes a direct analysis of GWAS data with hundreds of thousands SNPs. Such datasets will require a preprocessing step with filtering methods introduced in Section Two-stage Approach: Filters to Obtain Reduced Search Space.

TABLE 2 | Summary table of strategies reviewed to detect epistasis along with representative software programs and datasets applications.

Strategy		Software program	Exhaustivity	Pairwise-restricted	Dataset	# SNIPs	# Individuals	Runtime		References
								Sequential	Parallel	
Bitwise operations and Likelihood ratio tests		BOOST	Yes	Yes	WTCCC–multiple diseases	459,019	5000	NA	23 h (4 CPUs)	Wan et al., 2010
ROC curve analysis		GWIS	Yes	Yes	WTCCC–multiple diseases	459,019	5000	60 h	10.9 h (4 CPUs)	Goudey et al., 2013
Combinatorial		MDR	Yes	No	Simulated	50	500	36 min	NA	Moore et al., 2006
Random forest		Random jungle	No	No	Crohn's Disease	275,153	1003	12.7 h	0.53 h (40 CPUs)	Strobl et al., 2008
		SnpInterforest	No	No	WTCCC - RA	10,000	3500	98 h	NA	Yoshida and Koike, 2011
		GWGGI–TAMW	No	No	WTCCC - CAD	459,019	4864	10 h	NA	Wei and Lu, 2014
		GWGGI–LRMW	No	No	WTCCC - CAD	459,019	4864	3.5 h	NA	Wei and Lu, 2014
Bayesian		BEAM	No	No	AMD	47,727	3500	8 days	NA	Zhang and Liu, 2007
		epiBN	No	No	AMD	96,933	146	NA	NA	Han et al., 2012
	Markov blanket	Dasso-MB	No	No	AMD	91 495	14G	NA	NA	Han et al., 2010
		FEPI-MB	No	No	Simulated	500	4000	0.5 s	NA	Han et al., 2011
		IMBED	No	No	AMD	96,933	146	NA	NA	Yanlan and Jiawei, 2012
		TIE*	No	No	NARAC	490,073	2 044	NA	NA	Statnikov et al., 2013
Ant colony optimization		AntEpiSeeker	No	No	WTCCC – RA	332,831	3503	NA	5 days (2 CPUs)	Wang et al., 2010b
Computational evolution system		CES	No	No	Prostate cancer	219	2286	NA	NA	Moore and Hill, 2015

Runtimes were not always available (NA) and are indicated for simulated datasets when no real data application is available. The notation "WTCCC—multiple diseases" stands for "WTCCC—Bipolar Disorder (BD), Coronary Artery Disease (CAD), Crohn's Disease (CD), Hypertension (HT), Rheumatoid Arthritis (RA), Type 1 Diabetes (T1D), and Type 2 Diabetes (T2D).

Discussion

While an exhaustive epistasis analysis has become a quite straightforward task for SNP pairs, higher order interactions search in an exhaustive way is not conceivable at the moment. In this paper, we reviewed main current strategies for epistasis detection: exhaustive ones based on brute-force approach, filtering ones aiming at reducing genome-wide SNP set size, and different machine learning and combinatorial optimization procedures to find SNP associations yielding the best classification power. **Table 2** summarizes categories of methods described in this paper and gives representative software programs illustrating each category. In particular, this table highlights characteristics of the largest GWAS dataset analyzed using each software program. Runtimes are indicated, when available, for sequential and parallel versions of each program, for information about scalability.

Despite efforts for developing novel methods dedicated to epistasis detection, genetic variance of complex traits is weakly explained by detected epistatic interactions. This may be due to low detection power of pure and strict epistatic interactions for many of these methods. Much remains to be done to improve power of detection using model-free searches. For instance, the TURF method (see Section Filtering Based on Data Mining Techniques) which excludes SNPs with low predictive power, prior to performing epistasis detection, could be extended to other strategies like random forests, thereby improving detection of epistasis.

Precision of association measure estimates between epistatic interactions and phenotypes can be enhanced by increasing the $\frac{number\ of\ samples}{number\ of\ SNPs}$ ratio. First, increasing the sample size is a way to improve power of epistasis detection. Federating data from laboratories in the context of meta-analysis is a widespread approach, though subject to biases due to heterogeneity of laboratory practices. Second, reducing the number of SNPs to analyze might improve the statistical power under a given hypothesis. For instance, such a reduction of the search space size is possible thanks to systematic methods, like using significant pairwise interactions as a prior basis for the search of higher order interactions. Regarding data integration approaches, biological expert knowledge based-filters are often proposed to guide epistasis analysis. Being a biased approach, it is recommended to run at the same time a procedure without any *a priori* knowledge (Ritchie, 2015). Although development of epistasis detection methods is growing, many methods are hampered in presence of genetic heterogeneity or incomplete penetrance. Random forest-based techniques have been described to efficiently deal with genetic heterogeneity because data is split in different subsets in early stages of the algorithm (Koo et al., 2013). Besides, some of the existing software programs, like BEAM, will soon become unsuitable to GWAS datasets which will keep growing in size so that several millions of SNPs will be the rule rather than the exception. On the other hand, such a huge number of SNP might increase power of existing strategies tailored to handle massive datasets.

An interesting fact rarely discussed in literature describing the strategies reviewed in this survey is the confusing boundary between epistasis and linkage disequilibrium. Because linkage disequilibrium is by definition a phenomenon involving dependence between genetic variants, its frontier with epistatic interactions may be blurred since the aforementioned software programs are designed to detect SNPs that jointly affect the phenotype. This issue is particularly acute for case-only approaches. For standard case/control studies, if estimation of linkage disequilibrium within controls provides the same result as within cases, then the observed linkage disequilibrium does not originate from epistatic interactions.

Development of simulation models dealing with epistasis is also an active research area (Moore et al., 2015). Even if some authors already use various simulation models to estimate efficiency of their algorithms (Beam et al., 2014), these simulation tools lack the complexity of genetic mechanisms observed in real data. For instance, simulation models used in most software programs introduced in the previous sections only generate pairwise epistatic interactions. As a consequence, strategies dealing with higher order interaction detection are not confronted to simulation scenarios involving those types of interactions. Hopefully, such a gap will certainly be filled in the future.

With regard to evaluating association strength several authors rely on *p*-values to sort the best candidate SNPs. However, *p*-values alone do not allow any straightforward statement about the association strength. A *p*-value only estimates the probability of having observed the value of the test statistic under the null hypothesis (i.e., there is no association between the tested SNP and the phenotype) (du Prel et al., 2009). Odds ratio combined with confidence intervals are also widely used measures in GWAS reports.

The need for scalable and powerful strategy to detect SNP-SNP interactions is clearly unmet today. This is especially true for detection of higher order interactions. Massive testing of SNPs combinations should no longer be a tedious task, but rather a routine operation in a GWAS analysis workflow.

Conclusion

Currently, no strategy to detect epistasis stands out: all must strike balance between time efficiency and detection power. However, different techniques are available to reduce running times. Some authors improved time efficiency through parallelization of their strategies, e.g., random forests, ant colony optimization and approaches based on computational evolution. Other authors implemented versions of their software programs which use graphic processing units (GPU) instead of traditional central processing units (CPU).

Acknowledgments

CN 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. GR's work is supported by a Chair in Biostatistics jointly sponsored by the Centre National de la Recherche Scientifique and Lille 2 University. We also thank two anonymous reviewers for very helpful comments and valuable improvement of the manuscript.

References

Agresti, A. (2002). *Categorical Data Analysis, 2nd Edn*. Hoboken, NJ: John Wiley & Sons, Inc.

Alekseyenko, A. V., Lytkin, N. I., Ai, J., Ding, B., Padyukov, L., Aliferis, C. F., et al. (2011). Causal graph-based analysis of genome-wide association data in rheumatoid arthritis. *Biol. Direct*. 6:25. doi: 10.1186/1745-6150-6-25

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. (2010a). Local causal and markov blanket induction for causal discovery and feature selection for classification part I: algorithms and empirical evaluation. *J. Mach. Learn. Res.* 11, 171–234.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., and Koutsoukos, X. D. (2010b). Local Causal and markov blanket induction for causal discovery and feature selection for classification part II: analysis and extensions. *J. Mach. Learn. Res.* 11, 235-284.

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

Beam, A. L., Motsinger-Reif, A., and Doyle, J. (2014). Bayesian neural networks for detecting epistasis in genetic association studies. *BMC Bioinformat*. 15:368. doi: 10.1186/s12859-014-0368-0

Boone, C., Bussey, H., and Andrews, B. J. (2007). Exploring genetic interactions and networks with yeast. *Nat. Rev. Genet*. 8, 437–449. doi: 10.1038/nrg2085

Botta, V., Louppe, G., Geurts, P., and Wehenkel, L. (2014). Exploiting SNP Correlations within Random Forest for genome-wide association studies. *PLoS ONE* 9:e93379. doi: 10.1371/journal.pone.0093379

Breiman, L. (2001). Random forests. *Mach. Learn*. 45, 5–32. doi: 10.1023/A:1010933404324

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., et al. (2005). Identifying SNPs predictive of phenotype using random forests. *Genet. Epidemiol*. 28, 171–182. doi: 10.1002/gepi.20041

Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2006). Parallel multifactor dimensionality reduction: a tool for the large-scale analysis of gene-gene interactions. *Bioinformatics* 22, 2173–2174. doi: 10.1093/bioinformatics/btl347

Bush, W. S., Dudek, S. M., and Ritchie, M. D. (2009). Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. *Pac. Symp. Biocomput*. 368–379. doi: 10.1142/9789812836939_0035

Chatr-Aryamontri, A., Breitkreutz, B. J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., et al. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Res*. 43, D470–D478. doi: 10.1093/nar/gku1204

Chickering, D. M., Heckerman, D., and Meek, C. (2004). Large-sample learning of Bayesian Networks is NP-Hard. *J. Mach. Learn. Res*. 5, 1287–1330.

Cho, Y. M., Ritchie, M. D., Moore, J. H., Park, J. Y., Lee, K.-U., Shin, H. D., et al. (2004). Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus. *Diabetologia* 47, 549–554. doi: 10.1007/s00125-003-1321-3

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet*. 11, 2463–2468. doi: 10.1093/hmg/11.20.2463

Croft, D., Mundo, A. F., Haw, R., Milacic, M., Weiser, J., Wu, G., et al. (2014). The Reactome pathway knowledgebase. *Nucleic Acids Res*. 42, D472–D477. doi: 10.1093/nar/gkt1102

Culverhouse, R., Suarez, B. K., Lin, J., and Reich, T. (2002). A Perspective on Epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet*. 70, 461–471. doi: 10.1086/338759

De, R., Bush, W. S., and Moore, J. H. (2014). Bioinformatics challenges in genome-wide association studies (GWAS). *Methods Mol. Biol*. 1168, 63–81. doi: 10.1007/978-1-4939-0847-9_5

Dorigo, M., and Gambardella, L. M. (1997). Ant colonies for the travelling salesman problem. *Biosystems* 43, 73–81. doi: 10.1016/S0303-2647(97)01708-5

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H., et al. (2010). Missing heritability and strategies for finding the underlying causes of complex disease. *Nat. Rev. Genet*. 11, 446–450. doi: 10.1038/nrg2809

Ellis, J. A., Scurrah, K. J., Li, Y. R., Ponsonby, A. L., Chavez, R. A., Pezic, A., et al. (2015). Epistasis amongst PTPN2 and genes of the vitamin D pathway contributes to risk of juvenile idiopathic arthritis. *J. Steroid Biochem. Mol. Biol*. 145, 113–120. doi: 10.1016/j.jsbmb.2014.10.012

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edin*. 52, 399–433. doi: 10.1017/S0080456800012163

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., et al. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Res*. 41, D808–D815. doi: 10.1093/nar/gks1094

Gertz, J., Gerke, J. P., and Cohen, B. A. (2010). Epistasis in a quantitative trait captured by a molecular model of transcription factor interactions. *Theor. Popul. Biol*. 77, 1–5. doi: 10.1016/j.tpb.2009.10.002

Gou, J., Zhao, Y., Wei, Y., Wu, C., Zhang, R., Qiu, Y., et al. (2014). Stability SCAD: a powerful approach to detect interactions in large-scale genomic study. *BMC Bioinformatics*. 15:62. doi: 10.1186/1471-2105-15-62

Goudey, B., Rawlinson, D., Wang, Q., Shi, F., Ferra, H., Campbell, R. M., et al. (2013). GWIS–model-free, fast and exhaustive search for epistatic interactions in case-control GWAS. *BMC Genomics*. 13 (Suppl. 3):S10. doi: 10.1186/1471-2164-14-S3-S10

Grady, B. J., Torstenson, E. S., McLaren, P. J., DE Bakker, P. I., Haas, D. W., Robbins, G. K., et al. (2011). Use of biological knowledge to inform the analysis of gene-gene interactions involved in modulating virologic failure with efavirenz-containing treatment regimens in ART-naïve ACTG clinical trials participants. *Pac. Symp. Biocomput*. 253–264.

Greene, C. S., Hill, D. P., and Moore, J. H. (2009a). Environmental sensing of expert knowledge in a computational evolution system for complex problem solving in human genetics. *Genet. Evolut. Comput*. 19–36. doi: 10.1007/978-1-4419-1626-6_2

Greene, C. S., Himmelstein, D. S., Kiralis, J., and Moore, J. H. (2010). The informative extremes: using both nearest and farthest individuals can improve relief algorithms in the domain of human genetics. *Evolut. Comput. Mach. Learn. Data Min. Bioinform*. 6023, 182–193. doi: 10.1007/978-3-642-12211-8_16

Greene, C. S., Penrod, N. M., Kiralis, J., and Moore, J. H. (2009b). Spatially uniform relieff (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Min*. 2:5. doi: 10.1186/1756-0381-2-5

Gui, J., Moore, J. H., Williams, S. M., Andrews, P., Hillege, H. L., van der Harst, P., et al. (2013). A simple and computationally efficient approach to multifactor dimensionality reduction analysis of gene-gene interactions for quantitative traits. *PLoS ONE* 8:e66545. doi: 10.1371/journal.pone.0066545

Hahn, L. W., Ritchie, M. D., and Moore, J. H. (2003). Multifactor dimensionality reduction software for detecting gene-gene and gene-environment interactions. *Bioinformatics* 19, 376–382. doi: 10.1093/bioinformatics/btf869

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. doi: 10.1186/1471-2164-12-S2-S9

Han, B., Chen, X.-W., and Talebizadeh, Z. (2011). FEPI-MB: identifying SNPs-disease association using a Markov Blanket-based approach. *BMC Bioinform.* 12 (Suppl. 12):S3. doi: 10.1186/1471-2105-12-S12-S3

Han, B., Chen, X. W., Talebizadeh, Z., and Xu, H. (2012). Genetic studies of complex human diseases: characterizing SNP-disease associations using Bayesian networks. *BMC Syst Biol.* 6 (Suppl. 3):S14. doi: 10.1186/1752-0509-6-S3-S14

Han, B., Park, M., and Chen, X. W. (2010). A Markov blanket-based method for detecting causal SNPs in GWAS. *BMC Bioinform*. 11 (Suppl. 3):S5. doi: 10.1186/1471-2105-11-S3-S5

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nat. Rev. Genet*. 10, 392–404. doi: 10.1038/nrg2579

Hirschhorn, J. N. (2009). Genomewide association studies–illuminating biologic pathways. *N. Engl. J. Med*. 360, 1699–1701. doi: 10.1056/NEJMp0808934

Howard, T. D., Koppelman, G. H., Xu, J., Zheng, S. L., Postma, D. S., Meyers, D. A., et al. (2002). Gene-gene interaction in Asthma: IL4RA and IL13 in a Dutch population with Asthma. *Am. J. Hum. Genet*. 70, 230–236. doi: 10.1086/338242

Huang, C. H., Pei, J. C., Luo, D. Z., Chen, C., Chen, Y. W., and Lai, W. S. (2015). Investigation of gene effects and epistatic interactions between Akt1 and neuregulin 1 in the regulation of behavioral phenotypes and social functions in genetic mouse models of schizophrenia. *Front. Behav. Neurosci*. 8:455. doi: 10.3389/fnbeh.2014.00455

Huang, Y., Wuchty, S., and Przytycka, T. M. (2013). eQTL Epistasis - challenges and computational approaches. *Front. Genet.* 4:51. doi: 10.3389/fgene.2013.00051

Jiang, R., Tang, W., Wu, X., and Fu, W. (2009). A random forest approach to the detection of epistatic interactions in case-control studies. *BMC Bioinform*. 10:S65. doi: 10.1186/1471-2105-10-S1-S65

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. doi: 10.1093/bioinformatics/btu702

Johnstone, I. M., and Titterington, D. M. (2009). Statistical challenges of high-dimensional data. *Philos. Trans. A. Math. Phys. Eng. Sci.* 367, 4237–4253. doi: 10.1098/rsta.2009.0159

Kanehisa, M., Goto, S., Sato, Y., Furumichi, M., and Tanabe, M. (2012). KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40, D109–D114. doi: 10.1093/nar/gkr988

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., et al. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Res.* 40, D841–D846. doi: 10.1093/nar/gkr1088

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. *Lect. Notes Comp. Sci.* 784, 171–182. doi: 10.1007/3-540-57868-4_57

Koo, C. L., Liew, M. J., Mohamad, M. S., and Salleh, A. H. (2013). A Review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *Biomed. Res. Int.* 2013:432375. doi: 10.1155/2013/432375

Leinweber, D. J. (2007). Stupid data miner tricks:overfitting the S&P 500. *J. Invest.* 16, 15–22. doi: 10.3905/joi.2007.681820

Liu, J., Martin-Yken, H., Bigey, F., Dequin, S., François, J. M., and Capp, J. P. (2015). Natural yeast promoter variants reveal epistasis in the generation of transcriptional-mediated noise and its potential benefit in stressful conditions. *Genome Biol. Evol.* 7, 969–984. doi: 10.1093/gbe/evv047

Lu, Q., Wei, C., Ye, C., Li, M., and Elston, R. C. (2012). A likelihood ratio-based Mann-Whitney approach finds novel replicable joint gene action for type 2 diabetes. *Genet. Epidemiol.* 36, 583–593. doi: 10.1002/gepi.21651

Ma, L., Keinan, A., and Clark, A. G. (2015). Biological knowledge-driven analysis of epistasis in human GWAS with application to lipid traits. *Methods Mol. Biol.* 1253, 35–45. doi: 10.1007/978-1-4939-2155-3_3

Mackay, T. F. (2014). Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.* 15, 22–33. doi: 10.1038/nrg3627

Mackay, T. F., and Moore, J. H. (2014). Why epistasis is important for tackling complex human disease genetics. *Genome Med.* 6, 42. doi: 10.1186/gm561

Maher, B. (2008). Personal genomes: the case of the missing heritability. *Nature* 456, 18–21. doi: 10.1038/456018a

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., et al. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747–753. doi: 10.1038/nature08494

Marchini, J., Donnelly, P., and Cardon, L. R. (2005). Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.* 37, 413–417. doi: 10.1038/ng1537

Matsubara, K., Yamamoto, E., Mizobuchi, R., Yonemaru, J., Yamamoto, T., Kato, H., et al. (2015). Hybrid breakdown caused by epistasis-based recessive incompatibility in a cross of rice (Oryza sativa L.). *J. Hered.* 106, 113–122. doi: 10.1093/jhered/esu065

Matsuda, H. (2000). Physical nature of higher-order mutual information: intrinsic correlations and frustration. *Phys. Rev. E.* 62, 3096–3102. doi: 10.1103/PhysRevE.62.3096

McDonald, J. H. (2014). *Handbook of Biological Statistics, 3rd Edn.* Baltimore, MD: Sparky House Publishing.

McKinney, B. A., Reif, D. M., Ritchie, M. D., and Moore, J. H. (2006). Machine learning for detecting gene-gene interactions. *Appl. Bioinformat.* 5, 77–88. doi: 10.2165/00822942-200605020-00002

McKinney, B. A., Reif, D. M., White, B. C., Crowe, J. E. Jr., and Moore, J. H. (2007). Evaporative cooling feature selection for genotypic data involving interactions. *Bioinformatics* 23, 2113–2120. doi: 10.1093/bioinformatics/btm317

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum. Hered.* 56, 73–82. doi: 10.1159/000073735

Moore, J. H., Amos, R., Kiralis, J., and Andrews, P. C. (2015). Heuristic identification of biological architectures for simulating complex hierarchical genetic interactions. *Genet Epidemiol.* 39, 25–34. doi: 10.1002/gepi.21865

Moore, J. H., and Andrews, P. C. (2015). Epistasis analysis using multifactor dimensionality reduction. *Methods Mol. Biol.* 1253, 301–314. doi: 10.1007/978-1-4939-2155-3_16

Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N., et al. (2006). A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *J. Theor. Biol.* 241, 252–261. doi: 10.1016/j.jtbi.2005.11.036

Moore, J. H., and Hill, D. P. (2015). Epistasis analysis using artificial intelligence. *Methods Mol. Biol.* 1253, 327–346. doi: 10.1007/978-1-4939-2155-3_18

Moore, J. H., and White, B. C. (2007). Tuning ReliefF for genome-wide genetic analysis. *Evol. Comput. Mach. Learn. Data Min. Bioinformat.* 4447, 166–175. doi: 10.1007/978-3-540-71783-6_16

Moore, J. H., and Williams, S. M. (2005). Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *Bioessays* 27, 637–646. doi: 10.1002/bies.20236

Moore, J. H., and Williams, S. M. (2009). Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* 85, 309–320. doi: 10.1016/j.ajhg.2009.08.006

Namkung, J., Elston, R. C., Yang, J. M., and Park, T. (2009). Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method. *Genet Epidemiol.* 33, 646–656. doi: 10.1002/gepi.20416

Nishimura, D. (2001). BioCarta. *Biotech Softw. Internet Rep.* 2, 117–120. doi: 10.1089/152791601750294344

Pattin, K. A., and Moore, J. H. (2008). Exploiting the proteome to improve the genome-wide genetic analysis of epistasis in common human diseases. *Hum. Genet.* 124, 19–29. doi: 10.1007/s00439-008-0522-8

Payne, J. L., Greene, C. S., Hill, D. P., and Moore, J. H. (2010). Sensible initialization of a computational evolution system using expert knowledge for epistasis analysis in human genetics. *Exploitation Link. Learn. Evol. Algorithms* 3, 215–226. doi: 10.1007/978-3-642-12834-9_10

Pendergrass, S. A., Frase, A., Wallace, J., Wolfe, D., Katiyar, N., Moore, C., et al. (2013a). Genomic analyses with biofilter 2.0: knowledge driven filtering, annotation, and model development. *Bio. Data Min.* 6:25. doi: 10.1186/1756-0381-6-25

Pendergrass, S. A., Verma, S. S., Holzinger, E. R., Moore, C. B., Wallace, J., Dudek, S. M., et al. (2013b). Next-generation analysis of cataracts: determining knowledge driven gene-gene interactions using Biofilter, and gene-environment interactions using the PhenX Toolkit. *Pac. Symp. Biocomput.* 147–58. doi: 10.1142/9789814447973_0015

du Prel, J.-B., Hommel, G., Röhrig, B., and Blettner, M. (2009). Confidence interaval or *p*-value? *Dtsch. Arztebl. Int.* 106, 335–339. doi: 10.3238/arztebl.2009.0335

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* 81, 559–575. doi: 10.1086/519795

Ritchie, M. D. (2015). Finding the epistasis needles in the genome-wide haystack. *Methods Mol. Biol.* 1253, 19–33. doi: 10.1007/978-1-4939-2155-3_2

Ritchie, M. D., Hahn, L. W., and Moore, J. H. (2003). Power of multifactor dimensionality reduction for detecting gene-gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet. Epidemiol.* 24, 150–157. doi: 10.1002/gepi.10218

Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., et al. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am. J. Hum. Genet.* 69, 138–147. doi: 10.1086/321276

Robnik-Šikonja, M., and Kononenko, I. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Mach. Learn.* 53, 23–69. doi: 10.1023/A:1025667309714

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics* 53, 1253–1261. doi: 10.2307/2533494

Schwarz, D. F., König, I. R., and Ziegler, A. (2010). On safari to Random Jungle: a fast implementation of Random Forests for high-dimensional data. *Bioinformatics* 26, 1752–1758. doi: 10.1093/bioinformatics/btq257

Siemiatycki, J., and Thomas, D. C. (1981). Biological models and statistical interactions: an example from multistage carcinogenesis. *Int. J. Epidemiol.* 10, 383–387. doi: 10.1093/ije/10.4.383

Smith, S. B., Reenilä, I., Männistö, P. T., Slade, G. D., Maixner, W., Diatchenko, L., et al. (2014). Epistasis between polymorphisms in COMT, ESR1, and GCH1 influences COMT enzyme activity and pain. *Pain* 155, 2390–2399. doi: 10.1016/j.pain.2014.09.009

Statnikov, A., Lytkin NI, Lemeire, J., and Aliferis, C. F. (2013). Algorithms for discovery of multiple markov boundaries. *J. Mach. Learn. Res.* 14, 499–566.

Steen, K. V. (2012). Travelling the world of gene-gene interactions. *Brief Bioinform.* 13, 1–19. doi: 10.1093/bib/bbr012

Strobl, C., Boulesteix, A.-L., Kneib, T., Augustin, T., and Zeileis, A. (2008). Conditional variable importance for random forests. *BMC Bioinformat.* 9:307. doi: 10.1186/1471-2105-9-307

Taylor, M. B., and Ehrenreich, I. M. (2015). Higher-order genetic interactions and their contribution to complex traits. *Trends Genet.* 31, 34–40. doi: 10.1016/j.tig.2014.09.001

Vassy, J. L., Hivert, M. F., Porneala, B., Dauriz, M., Florez, J. C., Dupuis, J., et al. (2014). Polygenic type 2 diabetes prediction at the limit of common variant detection. *Diabetes* 63, 2172–2182. doi: 10.2337/db13-1663

Waddington, C. H. (1942). Canalization of development and the inheritance of acquired characters. *Nature* 150, 563–565. doi: 10.1038/150563a0

Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L., et al. (2010). BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am. J. Hum. Genet.* 87, 325–340. doi: 10.1016/j.ajhg.2010.07.021

Wang, X., Elston, R. C., and Zhu, X. (2010a). The meaning of interaction. *Hum. Hered.* 70, 269–277. doi: 10.1159/000321967

Wang, Y., Liu, X., Robbins, K., and Rekaya, R. (2010b). AntEpiSeeker: detecting epistatic interactions for case-control studies using a two-stage ant colony optimization algorithm. *BMC Res. Notes* 3:117. doi: 10.1186/1756-0500-3-117

Wei, C., and Lu, Q. (2014). GWGGI: software for genome-wide gene-gene interaction analysis. *BMC Genet.* 15:101. doi: 10.1186/s12863-014-0101-z

Wei, C., Schaid, D. J., and Lu, Q. (2013). Trees Assembling Mann-Whitney approach for detecting genome-wide joint association among low-marginal-effect loci. *Genet. Epidemiol.* 37, 84–91. doi: 10.1002/gepi.21693

Willighagen, E. L., Waagmeester, A., Spjuth, O., Ansell, P., Williams, A. J., Tkachenko, V., et al. (2013). The ChEMBL database as linked open data. *J. Cheminform.* 5:23. doi: 10.1186/1758-2946-5-23

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. doi: 10.4156/jdcta.vol6.issue15.32

Yoshida, M., and Koike, A. (2011). SNPInterForest: a new method for detecting epistatic interactions. *BMC Bioinformat.* 12:469. doi: 10.1186/1471-2105-12-469

Zhang, Y., and Liu, J. S. (2007). Bayesian inference of epistatic interactions in case-control studies. *Nat. Genet.* 39, 1167–1173. doi: 10.1038/ng2110

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Copyright © 2015 Niel, Sinoquet, Dina and Rocheleau. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) or licensor are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.