A survey of statistical methods for gene-gene interaction in case-control genome-wide association studies

Mathieu Emily

Journal de la Société Française de Statistique
Vol. 159 No. 1 (2018)

A survey of statistical methods for gene-gene interaction in case-control genome-wide association studies

Titre: Une revue des méthodes statistiques pour détecter des interaction gène-gène dans les études pangénomiques

Mathieu Emily[1]

Abstract: Over the last few years, case-control genome-wide association studies (GWAS) have proven to be a successful tool to identify genomic regions associated with complex diseases. Nevertheless, current GWAS still heavily rely on a single-marker strategy, in which each biological marker (or SNP for single nucleotide polymorphism) is tested individually for association with the disease. However, it is widely admitted that this is an oversimplified approach to tackle the complexity of underlying biological mechanisms and gene-gene interaction must be considered. Unfortunately, gene-gene interaction detection gives rise to complex statistical challenges, arising from the high-dimensionality and the complex architecture of the data as well as the size of the space of interaction models. The purpose of this survey is to provide a critical overview of the numerous statistical methods proposed to detect gene-gene interaction detection in GWAS. Those methods have been developed to detect interaction at various scales of the data and we decompose our survey in three main classes: SNP-SNP interaction methods, Gene-Gene interaction methods and large-scale methods. For each class of methods, we identify relative strengths and weaknesses in terms of statistical power and provide perspectives to the future of statistical strategies in gene-gene interaction analysis.

Résumé : Ces dernières années ont confirmé l'intérêt des études pangénomiques (GWAS) pour l'identification de régions génomiques associées à des maladies complexes. Néanmoins, les études actuelles reposent sur une stratégie simple-point, dans laquelle chaque marqueur biologique est testé individuellement pour l'association avec la maladie. Cependant, il est largement admis que cette approche est trop simpliste pour s'attaquer à la complexité des mécanismes biologiques sous-jacents et qu'il est important d'inclure l'interaction gène-gène dans l'analyse. Malheureusement, la détection de l'interaction gène-gène soulève des défis statistiques complexes, issus de la grande dimension et de l'architecture complexe des données ainsi que de la taille de l'espace des modèles d'interaction. Le but de cette étude est de fournir un aperçu des nombreuses méthodes statistiques proposées pour détecter une interaction gène-gène dans les GWAS. Ces méthodes ont été développées pour détecter l'interaction à différentes échelles des données et nous décomposons notre étude en trois classes principales : les méthodes d'interaction SNP-SNP, les méthodes d'interaction Gene-Gene et les méthodes à grande échelle. Pour chaque classe de méthodes, nous identifions les forces et les faiblesses en termes de puissance statistique et proposons des pistes de développements dans la modélisation statistique de l'interaction gène-gène.

Keywords: Gene-gene interaction, Regression models, Machine learning, Information theory, Statistical power

Mots-clés : Interaction gène-gène, Modèles de régression, Apprentissage, Théorie de l'information, Puissance statistique

AMS 2000 subject classifications: 35L05, 35L70

[1] Agrocampus Ouest - IRMAR UMR 6625
E-mail: mathieu.emily@agrocampus-ouest.fr

 ISSN: 2102-6238

1. Introduction

Case-control genome-wide association studies (GWAS) aim at investigating the genetic components of binary traits like major diseases (cancer, diabetes, Alzheimer, Parkinson, ...). GWAS typically compare the probabilistic distribution of hundreds of thousands of variables, called Single Nucleotide Polymorphisms (SNPs), between a population of affected individuals and a population of unaffected individuals. Since SNPs are precisely located in the genome and cover the entire genome, GWAS are used to identify SNPs or genomic regions that influence the risk of disease with the hope of accelerating drug and diagnostics development (Balding, 2006). Single-locus approaches, whereby each SNP is tested individually for association, have first been developed to analyse GWAS (Lewis, 2002). Although such single-locus approaches have successfully identified regions of disease susceptibility (Hindorff et al., 2009), findings were of modest effect and a large proportion of the genetic heritability is still not covered for common complex diseases (Maher, 2008; Manolio et al., 2009). To overcome such a lack, the search for gene-gene interaction, also refers to as epistasis, has gained in popularity (Moore, 2003; Phillips, 2008). Since human complex diseases are generally caused by the combined effect of multiple genes, the detection of genetic interactions is indeed essential to improve our knowledge of the etiology of complex diseases (Cordell, 2009; Hindorff et al., 2009).

However, the detection of gene-gene interaction in GWAS remains very challenging. First, from a computational point-of-view, the number of SNPs (or variables) in a GWAS can reach up to 1,000,000, thus generating $\binom{1000000}{2} \approx 5 \times 10^{11}$ possible interaction tests. An exhaustive testing requires extensive computing resources to perform, store and post-process the analysis (Ritchie, 2015). Next, from a biological point-of-view, gene-gene interaction has to face with is the lack of clear definition of what epistasis means. This has generated a big controversy regarding the ability of methods based on a statistical definition of interaction to detect biological interaction (Cordell, 2002). As a result, there is no consensus regarding the underlying null hypothesis of statistical methods which make the formal comparison between methods complex. Finally, from a statistical point-of-view, the detection of gene-gene interaction raises issues related to the statistical power of proposed methods, such as the data structure and the complexity of the models of interaction. GWAS data are first characterized by their high-dimension and by the correlation between variables inherited from the complex architecture of the genome. Furthermore, the lack in power is enhanced by the number of factors known to influence the power of statistical methods in GWAS (Emily and Friguet, 2015; Emily, 2016b) and by the vast amount of epistatic models (Li and Reich, 2000; Hallgrimsdottir and Yuster, 2008).

The purpose of the review is to provide a comprehensive comparison of the detection ability of statistical methods used to search for gene-gene interaction in susceptibility with a binary outcome. Compared to previous reviews proposed in the literature (Cordell, 2009; Steen, 2011; Ritchie, 2015; Niel et al., 2015), where attention was paid to the computational burden relative to the application of these methods at the genome scale, we focus on the statistical power of each method. Although the lack of consensus on the statistical null hypothesis upon which each method is based (which is not possible because of the limited biological knowledge of gene-gene interaction) prevents from a nice and clear formal power comparison, we focus here on several features known to influence power function. According to their relative scope, we first propose a classification of statistical methods and distinguish between methods dedicated to (1) SNP-SNP interaction, (2)

 ISSN: 2102-6238

gene-gene interaction and (3) genome-wide interaction. Due the complexity of living organisms, the variability of an outcome may indeed be related to variations observed at different biological levels. More specifically, GWAS data have the ability to investigate association at (1) the smallest scale of a single site in DNA using SNPs, (2) the functional level of the gene using SNP-sets and (3) the organism level through whole genome scans. From a statistical point-of-view, in each class, methods aim at addressing the issue of interaction by considering different sets of predictors. First for SNP-SNP interaction methods, only two predictors are included in the models. For gene-gene interaction methods, only two SNP-sets (with variable sizes, namely m_1 and m_2) are considered. However, SNP-set sizes are assumed to be moderate and does not scale up to the genome size. Finally, for genome-wide interaction the whole set of predictors is included in the model thus resulting in a deep modification of the scale of interactions that can be considered. Therefore, several statistical strategies have been proposed to deal with whole genome data by (1) restricting to exhaustive pairwise testing by combining all possible SNP-SNP interaction tests or all possible gene-gene interaction tests, (2) estimating multidimensional models that include interaction terms possible of order higher than 2 or (3) adapting machine learning techniques able to cope with GWAS data.

Due to the lack of clear definition of gene-gene interaction, our comparative analysis is not based on a formal power studies but rather focus on qualitative factors that influence the statistical power of each method. To successfully detect gene-gene interaction, a method is expected to be a good trade-off between flexibility, interpretability and computational efficiency. In this review, we summarized such a trade-off by ten main qualitative features, reported in Table 1, that have been stated in previous reports (Cordell, 2009; Niel et al., 2015; Emily, 2016b). Those features can be grouped into four main classes whether they are related to (1) the impact of the correlation structure among predictors, (2) the nature of the underlying signal (*i.e.* the relationship between between the response Y and the set of predictors), (3) the computational aspects of the method and (4) the interpretability of the results.

First, the architecture of the genome induces a complex correlation structure among the predictors. It is therefore important for a statistical method to account for such a structure. However, since correlation structure can be very different from a genomic region to another, statistical methods are expected to consider a large panel of structure while not overfitting the data. We focus our comparative analysis on the following qualitative features: ♯1: the robustness to predictors structure (*i.e.* the capacity for a method to perform well regardless of the predictor structure), ♯2: the control of the Type-I error rate and ♯3: the need for parameter hand-setting (*i.e.* the capacity for a method not to rely on parameters that are likely to generate overfitting).

Next, because of the complexity of the space of epistatic models, the true relationship between the response and the predictor (also called nature of the signal) can have very diverse form. It is therefore crucial for a method to be flexible enough to catch a large variety of interaction signals. We therefore consider 3 additional qualitative features: ♯4 the reliance of an underlying model (*i.e.* the model-free characteristic of a method), ♯5 the capacity to detect non-linear relationship between the response and predictors and ♯6 the capacity to detect pure epistatic signal (*i.e.* signal without marginal effect).

Finally, we consider qualitative features that are of importance when searching for gene-gene interactions in practice. We focus on the computational performances of each method by considering ♯7 the possibility for a method to a have a closed formulation that fasten the

 ISSN: 2102-6238

computation of the method, ♯8 the computational efficiency of each method (*i.e.* the computational cost), ♯9 the scalability of the method to determine whether a method is applicable at the genome scale or not and ♯10 the interpretability of the results. The latter feature is important in the sense that interpretability allows to give insights in further investigation to be performed to propose therapeutic development.

TABLE 1. *Summary of the ten qualitative features that serve as a basis for our comparative analysis. Those features are related to four characteristics : the correlation among predictors, the nature of the signal, the computational aspects and the interpretability of the results.*

Correlation among predictors	Nature of the signal	Computational aspects	Interpretability
♯1 Robustness to predictors structure	♯4 Reliance of an underlying model	♯7 Derivation in closed form	♯10 Capacity to detect true causal predictors
♯2 Control of the Type-I error rate	♯5 Capacity to detect non-linear association	♯8 Computational efficiency	
♯3 Parameter hand-setting	♯6 Capacity to detect pure epistasis	♯9 Scalability	

The remainder of the article is organized as follows. In section 2, the main notations used throughout the paper as well as the statistical characteristics of a GWAS dataset are introduced. Then, Section 3 is devoted to statistical methods proposed to detect interaction between two SNPs, therefore focusing on the smallest possible scale. In Section 4, statistical procedures motivated to detect interaction at the gene level are presented and evaluated. In the next Section 5, statistical methods and strategies dedicated to large-scale (up to genome-wide) analysis are presented. Finally, the paper is ended by Section 6 where the remaining limitations as well as some perspectives in gene-gene interaction are discussed.

2. Notations, modeling assumptions and statistical hypotheses

Let $Y \in \{0,1\}$ be the random binary variable corresponding to the disease status (*i.e.* Y is the phenotype variable) where $Y = 0$ stands for an healthy individual (control status) and $Y = 1$ a diseased individual (case status). Let consider that the genotype of an individual is measured through a collection of p SNPs. In more details, for $i = 1,\ldots,p$, let X_i be a random variable modeling the genotype of the i^{th} SNP. Various modeling of the X_i's can be considered. First, in its raw representation, X_i is a categorical variable with three states denoted by $X_i \in \{AA, Aa, aa\}$. States AA and aa correspond to the two homozygote genotypes while Aa is the heterozygote state, where A is the major (resp. minor) allele of the SNP i. In a second representation, it can be considered that genotypes are allele counts so that $X_i \in \{0,1,2\}$. In another modeling, each X_i can be treated in terms of allele. Therefore, at the allele level, $X_i \in \{A,a\}$ is a binary variable and, since each individual has two copies of an allele, the sample size is twice the number of individuals. Finally, a fourth representation of X_i is a continuous allelic modeling where $X_i \in \{0,1\}$. It is noteworthy that each of these four modeling are considered throughout the paper. The choice of a specific modeling is mostly driven by mathematical arguments but it is important to keep in mind that it deeply impacts the interpretation of the results.

 ISSN: 2102-6238

Let further consider a sample of n individuals with n_c controls and n_d cases ($n_c + n_d = n$) and $\mathbf{Y} = [y_1, \ldots, y_n]'$ the vector of the observed binary phenotypes. The observed genotypes can be represented by a $n \times p$ matrix: $\mathbf{X} = [x_{ij}]_{i\in 1\ldots n; j\in 1\ldots p}$ where x_{ij} is the observed genotype for SNP j carried by individual i. Therefore, $x_{i,j}$ is the realization of a random variable characterized by one of the four probabilistic distributions introduced in the previous paragraph. Therefore a typical dataset can be summarized as in the following Equation (1).

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_c} \\ y_{n_c+1} \\ \vdots \\ y_{n_c+n_d} \end{bmatrix} \quad \mathbf{X} = \begin{bmatrix} x_{1,1} & \cdots & x_{1,p} \\ \vdots & \ddots & \vdots \\ x_{n_c,1} & \cdots & x_{n_c,p} \\ x_{(n_c+1),1} & \cdots & x_{(n_c+1),p} \\ \vdots & \ddots & \vdots \\ x_{(n_c+n_d),1} & \cdots & x_{(n_c+n_d),p} \end{bmatrix} \tag{1}$$

2.1. SNP-SNP interaction statistical hypotheses

In a first part we focus our attention to statistical methods that aim at detecting an association between Y and a pair (X_i, X_j) where $i \neq j$ and $i = 1, \ldots, p$, $j = 1, \ldots, p$. Four different types of association are considered in the literature, whether the association is investigated at the genotype or the allele level and whether the type of investigated association is statistical or biological. The four types of association, called *StatAllele*, *StatGeno*, *BioAllele* and *BioGeno* relies on four different null hypotheses : $\mathcal{H}_{0_{SSI}}^{StatAllele}$ (see Eq. (2)), $\mathcal{H}_{0_{SSI}}^{StatGeno}$ (see Eq. (3)), $\mathcal{H}_{0_{SSI}}^{BioAllele}$ (see Eq. (4)) and $\mathcal{H}_{0_{SSI}}^{BioGeno}$ (see Eq. (5)). Statistical hypotheses refers to a linearity in the effects that is measured by odds-ratio as proposed in Equations (2) and (3). However biological hypotheses are based on biologically-driven measures of association that are assumed to be equal in cases and controls, as displayed in Equations (4) and (5).

In each of the four situations, the null hypothesis can be formalized via the joint distribution at the genotype level:

$$\pi_{k,\ell}^y = \mathbb{P}[Y = y, X_i = k, X_2 = \ell] \text{ for } y \in \{0,1\}, \;\; k \in \{AA, Aa, aa\} \text{ and } \ell \in \{BB, Bb, bb\}.$$

If association is investigated at the allelic level, the joint probability distribution is denoted by $p_{k,\ell}^y$, where $p_{k,\ell}^y$ is obtained according to $\pi_{k,\ell}^y$'s as follows:

$$p_{a,b}^y = \pi_{aa,bb}^y + \frac{\pi_{aa,Bb}^y + \pi_{Aa,bb}^y}{2} + \frac{\pi_{Aa,Bb}^y}{4} \quad ; \quad p_{a,B}^y = \pi_{aa,BB}^y + \frac{\pi_{aa,Bb}^y + \pi_{Aa,BB}^y}{2} + \frac{\pi_{Aa,Bb}^y}{4}$$
$$p_{A,b}^y = \pi_{AA,bb}^y + \frac{\pi_{Aa,bb}^y + \pi_{AA,Bb}^y}{2} + \frac{\pi_{Aa,Bb}^y}{4} \quad ; \quad p_{A,B}^y = \pi_{AA,BB}^y + \frac{\pi_{aA,BB}^y + \pi_{AA,Bb}^y}{2} + \frac{\pi_{Aa,Bb}^y}{4}$$

The formal notations for the joint distributions considered in SNPxSNP association tests are summarized in Table 2.

The four null hypotheses can then be formulated as follows:

 ISSN: 2102-6238

TABLE 2. *Joint distributions of the triplet (Y, X_1, X_2) when X_1 and X_2 have three categories (genotypic data) or two categories (allelic data).*

	Joint distribution in controls	Joint distributions in cases
Genotypic data	$\begin{bmatrix} \pi^0_{AA,BB} & \pi^0_{AA,Bb} & \pi^0_{AA,bb} \\ \pi^0_{Aa,BB} & \pi^0_{Aa,Bb} & \pi^0_{Aa,bb} \\ \pi^0_{aa,BB} & \pi^0_{aa,Bb} & \pi^0_{aa,bb} \end{bmatrix}$	$\begin{bmatrix} \pi^1_{AA,BB} & \pi^1_{AA,Bb} & \pi^1_{AA,bb} \\ \pi^1_{Aa,BB} & \pi^1_{Aa,Bb} & \pi^1_{Aa,bb} \\ \pi^1_{aa,BB} & \pi^1_{aa,Bb} & \pi^1_{aa,bb} \end{bmatrix}$
Allelic data	$\begin{bmatrix} p^0_{A,B} & p^0_{A,b} \\ p^0_{a,B} & p^0_{a,b} \end{bmatrix}$	$\begin{bmatrix} p^1_{A,B} & p^1_{A,b} \\ p^1_{a,B} & p^1_{a,b} \end{bmatrix}$

$$\mathcal{H}_{0SSI}^{StatAllele} : \frac{\left(\frac{p^1_{a,b}}{p^0_{a,b}}\frac{p^1_{A,B}}{p^0_{A,B}}\right)}{\left(\frac{p^1_{a,B}}{p^0_{a,B}}\frac{p^1_{A,b}}{p^0_{A,b}}\right)} = 1 \tag{2}$$

$$\mathcal{H}_{0SSI}^{StatGeno} : \frac{\left(\frac{\pi^1_{k,\ell}}{\pi^0_{k,\ell}}\frac{\pi^1_{AA,BB}}{\pi^0_{AA,BB}}\right)}{\left(\frac{\pi^1_{k,BB}}{\pi^0_{k,BB}}\frac{\pi^1_{AA,\ell}}{\pi^0_{AA,\ell}}\right)} = 1 \quad \left(\forall (k,\ell) \in [Aa, aa] \times [Bb, bb]\right) \tag{3}$$

$$\mathcal{H}_{0SSI}^{BioAllele} : \frac{\left(\frac{p^1_{ab} - (p^1_{ab}+p^1_{aB})(p^1_{ab}+p^1_{Ab})}{p^0_{ab} - (p^0_{ab}+p^0_{aB})(p^0_{ab}+p^0_{Ab})}\right)\left(\frac{p^1_{AB}+p^1_{Ab}+p^1_{aB}+p^1_{ab}}{p^0_{AB}+p^0_{Ab}+p^0_{aB}+p^0_{ab}}\right)}{\sqrt{\frac{(p^1_{AB}+p^1_{Ab})(p^1_{aB}+p^1_{ab})(p^1_{AB}+p^1_{aB})(p^1_{Ab}+p^1_{ab})}{(p^0_{AB}+p^0_{Ab})(p^0_{aB}+p^0_{ab})(p^0_{AB}+p^0_{aB})(p^0_{Ab}+p^0_{ab})}}} = 1 \tag{4}$$

$$\mathcal{H}_{0SSI}^{BioGeno} : \frac{\left(\frac{\pi^1_{k,\ell}}{\pi^0_{k,\ell}}\frac{\pi^0_{AA,BB}}{\pi^1_{AA,BB}}\right)}{\left(\frac{\pi^1_{k,BB}+\pi^1_{k,Bb}+\pi^1_{k,bb}}{\pi^0_{k,BB}+\pi^0_{k,Bb}+\pi^0_{k,bb}}\frac{\pi^0_{AA,BB}+\pi^0_{AA,Bb}+\pi^0_{AA,bb}}{\pi^1_{AA,BB}+\pi^1_{AA,Bb}+\pi^1_{AA,bb}}\right)\left(\frac{\pi^1_{AA,\ell}+\pi^1_{Aa,\ell}+\pi^1_{aa,\ell}}{\pi^0_{AA,\ell}+\pi^0_{Aa,\ell}+\pi^0_{aa,\ell}}\frac{\pi^0_{AA,BB}+\pi^0_{Aa,BB}+\pi^0_{aa,BB}}{\pi^1_{AA,BB}+\pi^1_{Aa,BB}+\pi^1_{aa,BB}}\right)} = 1 \tag{5}$$
$$\left(\forall (k,\ell) \in [Aa, aa] \times [Bb, bb]\right)$$

Table 3 indicates the underlying null hypothesis (among $\mathcal{H}_{0SSI}^{StatAllele}$, $\mathcal{H}_{0SSI}^{StatGeno}$, $\mathcal{H}_{0SSI}^{BioAllele}$, $\mathcal{H}_{0SSI}^{BioGeno}$) addressed by each of method considered in this review and Section 3 aims at introducing each method in details.

2.2. *Gene-gene interaction statistical hypotheses*

In a second part, we focus on gene-level testing approaches. In such approaches, we consider two SNP-sets (called gene for ease of reading) where each gene is a collection of respectively m_1 and m_2 SNPs. The observed genotypes for gene X_1 can be represented by a $n \times m_1$ matrix: $\mathbf{X_1} = [x^1_{ij}]_{i\in 1\ldots n; j\in 1\ldots m_1}$ where $x^1_{ij} \in \{0;1;2\}$ is the number of copies of the minor allele for SNP j carried by individual i. A similar representation is used for gene X_2 where $\mathbf{X_2}$ is a $n \times m_2$ matrix. Let us further introduce $\mathbf{X_1^c}$ and $\mathbf{X_2^c}$ the matrices of observed genotypes among controls for gene X_1 and X_2 and $\mathbf{X_1^d}$ and $\mathbf{X_2^d}$ among cases for both genes. Thus $\mathbf{X_1^c}$ is a $n_c \times m_1$ matrix, $\mathbf{X_1^d}$ a $n_d \times m_1$ matrix, $\mathbf{X_2^c}$ a $n_c \times m_2$ matrix and $\mathbf{X_2^d}$ a $n_d \times m_2$ matrix. A general setup of the observed values can be presented as proposed in Equation (6):

 ISSN: 2102-6238

$$\mathbf{Y} = \begin{bmatrix} y_1 \\ \vdots \\ y_{n_c} \\ y_{n_c+1} \\ \vdots \\ y_{n_c+n_d} \end{bmatrix}; \mathbf{X_1} = \begin{bmatrix} \mathbf{X_1^c} \\ \\ \mathbf{X_1^d} \end{bmatrix} = \begin{bmatrix} x^1_{11} & \cdots & x^1_{1m_1} \\ \vdots & \ddots & \vdots \\ x^1_{n_c1} & \cdots & x^1_{n_cm_1} \\ x^1_{(n_c+1)1} & \cdots & x^1_{(n_c+1)m_1} \\ \vdots & \ddots & \vdots \\ x^1_{(n_c+n_d)1} & \cdots & x^1_{(n_c+n_d)m_1} \end{bmatrix}; \mathbf{X_2} = \begin{bmatrix} \mathbf{X_2^c} \\ \\ \mathbf{X_2^d} \end{bmatrix} = \begin{bmatrix} x^2_{11} & \cdots & x^2_{1m_2} \\ \vdots & \ddots & \vdots \\ x^2_{n_c1} & \cdots & x^2_{n_cm_2} \\ x^2_{(n_c+1)1} & \cdots & x^2_{(n_c+1)m_2} \\ \vdots & \ddots & \vdots \\ x^2_{(n_c+n_d)1} & \cdots & x^2_{(n_c+n_d)m_2} \end{bmatrix} \tag{6}$$

The general question raised by the detection of gene-gene interaction is to determine whether accounting for the joint information of $\mathbf{X_1}$ and $\mathbf{X_2}$ improves the explanation of $\mathbf{Y}$ compared to only considering both marginal informations from $\mathbf{X_1}$ and $\mathbf{X_2}$. By considering the dimensionality of each gene and the correlation within and between genes, such a question can be formalized in many different ways, thus leading to a large number of potential null hypothesis. In this review we focus on four main statistical approaches introduced in the literature: a signal detection approach, a dimension reduction approach, an approach based on comparing covariance structures and an entropy-based approach. Section 4 aims at explaining the methods in the literature by focusing on the design of statistics with respect to the statistical approaches as summarized in Table 5.

3. SNP-SNP interaction

In this section we focus our attention to statistical methods that aim at detecting an association between Y and a pair (X_i, X_j) where $i \neq j$ and $i = 1,\ldots,p$, $j = 1,\ldots,p$. Let A and a (resp. B and b) be the two alleles for X_i (resp. X_j). When considering the raw representation for X_i and X_j, analyzed data can be summarized into a $3\times3\times2$ three-way contingency table. In such a table, cell (i,j,k) is denoted by $n^k_{i,j}$ and corresponds to the number of individual with genotype $i \in \{AA, Aa, aa\}$ for X_i, genotype $j \in \{BB, Bb, bb\}$ for X_j and disease status $k \in \{0,1\}$ (see Figure 1).

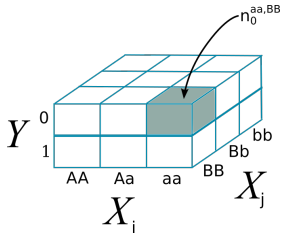

FIGURE 1. *Raw representation for SNP-SNP interaction using a* $3\times3\times2$ *three-way contingency table.*

A large number of methods have been proposed in the literature to detect interaction at the SNP level, *i.e.* SNP-SNP interaction (Shang et al., 2011). Based on statistical modeling hypothesis, these methods can be grouped into 3 main classes: regression-based methods, methods based on Wald-like tests and entropy-based methods. The purpose of regression-based methods is to model the relationship between Y, X_1 and X_2 using either a logistic regression (Cordell, 2002; Purcell et al., 2007) or a Poisson regression (Wan et al., 2010). In Wald-like tests, proposed methods

 ISSN: 2102-6238

introduced Wald statistics to either compare the correlation between X_1 and X_2 in cases and controls (Zhao et al., 2006), or to evaluate differences in odds-ratio between cases and controls (Wu et al., 2010b; Ueki and Cordell, 2012; Emily, 2012). In the third class, several measures of interaction based on information theory are introduced through the derivation of information gain statistics (Fan et al., 2011; Chen et al., 2011; Kwon et al., 2014), relative information gain statistics (Yee et al., 2013; Dong et al., 2008; Kwon et al., 2014; Chattopadhyay et al., 2014) and interaction information statistics (Mielniczuk and Rdzanowski, 2017).

TABLE 3. *Summary of the SNP-SNP interaction methods and their respective null hypothesis presented in Equations 2, 3, 4 and 5.*

Statistical model	Popular name	Null hypothesis	Reference
1dof logistic regression	PLINK	$\mathcal{H}0_{SSI}^{StatAllele}$	Purcell et al. (2007)
4dof logistic regression		$\mathcal{H}0_{SSI}^{StatGeno}$	Cordell (2002)
Poisson regression	BOOST	$\mathcal{H}0_{SSI}^{StatGeno}$	Wan et al. (2010)
Correlation	W_{Zhao}	$\mathcal{H}0_{SSI}^{BioAllele}$	Zhao et al. (2006)
Odds-ratio	W_{Wu}	$\mathcal{H}0_{SSI}^{StatAllele}$	Wu et al. (2010b); Ueki and Cordell (2012)
Odds-ratio	IndOR	$\mathcal{H}0_{SSI}^{BioGeno}$	Emily (2012)
Entropy	IG, RIG, ES, II	$\mathcal{H}0_{SSI}^{BioGeno}$	Fan et al. (2011); Chen et al. (2011) Kwon et al. (2014); Yee et al. (2013) Dong et al. (2008); Chattopadhyay et al. (2014) Mielniczuk and Rdzanowski (2017)

In the remainder of this section details regarding the most popular methods in each of the 3 classes are given. These methods relies on several statistical hypothesis as described in Equations 2, 3, 4 and 5. Table 3 provides a summary of the underlying null hypothesis addressed by each method. A qualitative comparative summary of the developed methods is proposed in Table 4. Our evaluation focus on four main features known to play a role in the statistical power for detecting interaction at the SNP level and in the interpretation of the results: (1) the dependence to an underlying modeling of interaction, (2) the ability to detect interaction at the genotype level, (3) the closed formulation of the statistics and (4) the control of the type-I error rate.

3.1. Regression-based methods

In regression-based models, proposed procedures are based on the statistical definition of interaction that corresponds to a deviation from the additivity of effects. A first class of methods relies on the regression of the binary outcome response using appropriate logistic regression (Cordell, 2002; Purcell et al., 2007). In another class, a regression model is based on the observed counts in contingency table (see Figure 1) using a dedicated Poisson regression (Wan et al., 2010).

Logistic regression

The regression-based procedures, proposed in Purcell et al. (2007) and Cordell (2002), aim at explaining the disease status Y knowing the values of X_i and X_j using a logistic model. Statistical test for interaction is then performed by testing the nullity of the interaction coefficients. In more details, the implementation of the software PLINK (Purcell et al., 2007) focus on an allelic representation of the X_i and is based on the model defined in Equation (7). PLINK's statistical test

 ISSN: 2102-6238

is defined through the following statistical hypothesis : $\mathcal{H}_0 : \delta = 0$ vs $\mathcal{H}_1 : \delta \neq 0$, corresponding to $\mathcal{H}_{0_{SSI}}^{StatAllele}$ in Equation (2), and can be performed using standard techniques based on maximum likelihood estimation (McCullagh and Nelder, 1989). It can be remarked that under $\mathcal{H}_0$, classical association statistics follow a χ^2 distribution with 1 degree-of-freedom.

$$\text{logit}\Big(\mathbb{P}[Y = 1|(X_i, X_j) = (x_i, x_j)]\Big) = \alpha + \beta\mathbb{I}_{x_i=A} + \gamma\mathbb{I}_{x_j=B} + \delta\mathbb{I}_{(x_i,x_j)=(A,B)} \tag{7}$$

A genotypic version of PLINK's test that relies on a categorical modeling of the X_i's has been proposed by Cordell (2002). Such a statistical procedure is performed by testing the nullity of the 4-dimensional vector $[\delta_{Aa,Bb}, \delta_{aa,Bb}, \delta_{Aa,bb}, \delta_{aa,bb}]$, where $\delta_{k,\ell}$ are the interaction coefficients in Equation (8). A likelihood ratio test (LRT) can then be performed to test for $\mathcal{H}_0 : [\delta_{Aa,Bb}, \delta_{aa,Bb}, \delta_{Aa,bb}, \delta_{aa,bb}] = [0,0,0,0]$ vs. $\mathcal{H}_1 : [\delta_{Aa,Bb}, \delta_{aa,Bb}, \delta_{Aa,bb}, \delta_{aa,bb}] \neq [0,0,0,0]$, corresponding to $\mathcal{H}_{0_{SSI}}^{StatGeno}$ in Equation (3). Under $\mathcal{H}_0$, LRT is expected to follow a χ^2 distribution with 4 degrees-of-freedom.

$$\begin{aligned}\text{logit}\Big(\mathbb{P}[Y = 1|(X_i, X_j) = (x_i, x_j)]\Big) &= \alpha + \sum_{k\in\{Aa,aa\}} \beta_k\mathbb{I}_k(x_i) + \sum_{\ell\in\{Bb,bb\}} \gamma_\ell\mathbb{I}_\ell(x_j) \\ &+ \sum_{(k,\ell)\in\{Aa,aa\}\times\{Bb,bb\}} \delta_{k,\ell}\mathbb{I}_{(k,\ell)}(x_i, x_j)\end{aligned} \tag{8}$$

Poisson regression (BOOST)

An alternative method based on a log-linear model that focus on the conditional modeling of the counts, with respect to Y, X_i and X_j, has further been proposed by Wan et al. (2010). Let N be the random variable coding for the number of sampled individuals; the log-linear model proposed by Wan et al. (2010) can then be defined as in Equation (9). Similar to the logistic regression procedure, testing for SNP-SNP interaction can be performed with a likelihood ratio test where: $\mathcal{H}_0 : [\lambda_{1,Aa,Bb}, \lambda_{1,aa,Bb}, \lambda_{1,Aa,bb}, \lambda_{1,aa,bb}] = [0,0,0,0]$ and $\mathcal{H}_1 : [\lambda_{1,Aa,Bb}, \lambda_{1,aa,Bb}, \lambda_{1,Aa,bb}, \lambda_{1,aa,bb}] \neq [0,0,0,0]$, equivalent to $\mathcal{H}_{0_{SSI}}^{StatGeno}$ in Equation (3). Under $\mathcal{H}_0$, such a statistic is expected to follow a χ^2 distribution with 4 degrees-of-freedom.

$$\begin{aligned}\mathbb{E}[N|(Y = y, X_i = x_i, X_j = x_j)] &= \lambda + \lambda_Y\mathbb{I}_1(y) + \sum_{k\in\{Aa,aa\}} \lambda_{i,k}\mathbb{I}_k(x_i) + \sum_{\ell\in\{Bb,bb\}} \lambda_{j,\ell}\mathbb{I}_\ell(x_j) \\ &+ \sum_{(k,\ell)\in\{Aa,aa\}\times\{Bb,bb\}} \lambda_{X_i,X_j,k,\ell}\mathbb{I}_{(k,\ell)}(x_i, x_j) \\ &+ \sum_{(k,\ell)\in\{Aa,aa\}\times\{Bb,bb\}} \lambda_{Y,X_i,X_j,1,k,\ell}\mathbb{I}_{(1,k,\ell)}(y, x_i, x_j)\end{aligned} \tag{9}$$

The Poisson regression-based method has been implemented in a software called BOOST (Wan et al., 2010). Improvements have been proposed to speed up the computation of the test, such as a GPU implementation (Yung et al., 2011).

Strengths for regression-based methods. Regression-based methods rely on a well-established statistical modeling and therefore offer strong guarantees in terms of control of the Type-I error. Furthermore, since Equations (8) and (9) are based on genotype data, the 4 dof logisitic regression and the BOOST method are able to detect interaction at the genotype level.

 ISSN: 2102-6238

Weaknesses for regression-based methods. However compared to linear regression, logistic and Poisson regression do not show a closed formulation for the estimation of the regression coefficients. Statistical tests of the significance of regression coefficients, as proposed in Equations (7), (8) and (9), resort to a computational optimization of the likelihood that may introduce numerical instabilities. However, the major limitation of regression-based methods is the assumption that an interaction is a deviation to linearity. Finally, the assumption made by the 1dof logistic regression test, implemented in PLINK, only consider interaction at the allele level thus preventing from detecting interaction at the genotype level.

3.2. *Wald-like tests*

Wald statistics are widely used in statistics to test whether a vector of parameters θ is equal to a targeted vector θ_0. Given that θ and θ_0 are q-dimensions vectors and that $\widehat{\theta}$ is an unbiased estimator for θ, the statistic W (see Equation (10)) follows a χ^2 distribution with q degrees-of-freedom under $\mathcal{H}_0 : \theta = \theta_0$.

$$W = (\widehat{\theta} - \theta_0)^t \left(\mathbb{V}\left[\widehat{\theta}\right]\right)^{-1} (\widehat{\theta} - \theta_0) \sim_{\mathcal{H}_0} \chi^2_q \tag{10}$$

In the search for SNP-SNP interactions, several procedures have been proposed in the literature that differ in their definition of the vector θ. In the remainder of this section, three Wald statistics, W_{Zhao} (Zhao et al., 2006), W_{Wu} (Wu et al., 2010b) and W_{IndOR} (Emily, 2012) are introduced, corresponding to different modeling of the interaction.

One-dimensional W_{Zhao}

The procedure proposed by Zhao et al. (2006) relies on the allelic level of interaction and aims at comparing the correlations between X_1 and X_2 conditional to the disease status, respectively $Y = 0$ and $Y = 1$. Correlation refers to the statistical definition of correlation as it is related to the biological concept of Linkage Disequilibrium (Hill and Robertson, 1968). More precisely, Zhao et al. (2006) proposed the following W_{Zhao} statistic:

$$W_{Zhao} = \frac{(r_1 - r_0)^2}{\mathbb{V}(r_1) + \mathbb{V}(r_0)} \tag{11}$$

where for $k \in [0, 1]$:

$$r_k = \frac{\mathbb{P}[(X_i, X_j) = (A, B)|Y = k] - \mathbb{P}[X_i = A|Y = k]\mathbb{P}[X_i = B|Y = k]}{\sqrt{\mathbb{P}[X_i = A|Y = k](1 - \mathbb{P}[X_i = A|Y = k])\mathbb{P}[X_j = B|Y = k](1 - \mathbb{P}[X_j = B|Y = k])}}$$

Details regarding the maximum-likelihood estimation procedure for W_{Zhao} can be found in Zhao et al. (2006). Testing for interaction is based on the property that under the null hypothesis $\mathcal{H}_0 : r_1 = r_0$, $W_{Zhao} \sim_{\mathcal{H}_0} \chi^2_{1dof}$. It is noteworthy that the null hypothesis corresponds to $\mathcal{H}_{0\,SSI}^{BioAllele}$ in Equation (4).

One-dimensional W_{Wu}

The procedure proposed in Wu et al. (2010b) is also based on the allelic data but aims to compare the joint allelic odds-ratio conditional to the disease status ($Y = 0$ and $Y = 1$). The following Wald statistic, W_{Wu}, is then proposed:

 ISSN: 2102-6238

$$W_{Wu} = \frac{(\lambda_1 - \lambda_0)^2}{\mathbb{V}(\lambda_1) + \mathbb{V}(\lambda_0)} \tag{12}$$

where λ_i is the allele log odds-ratio conditional to $Y = k$ where $k \in \{0;1\}$. More precisely for $k \in [0,1]$:

$$\lambda_k = \log\left(\frac{\left(\frac{\mathbb{P}[(X_i,X_j)=(a,b)|Y=k]}{\mathbb{P}[(X_i,X_j)=(a,B)|Y=k]}\right)}{\left(\frac{\mathbb{P}[(X_i,X_j)=(A,b)|Y=k]}{\mathbb{P}[(X_i,X_j)=(A,B)|Y=k]}\right)}\right)$$

An estimation procedure for λ_k and $\mathbb{V}(\lambda_k)$ is proposed in Wu et al. (2010b) under the assumptions that (1) the two populations $Y = 0$ and $Y = 1$ are independent and (2) the phase, *i.e.* the knowledge of the alleles given the genotypes, is known. In Ueki and Cordell (2012), improvement regarding the estimation of $\mathbb{V}(\lambda_k)$ has been proposed in order to account for the unknown phase. Similar to W_{Zhao}, testing for interaction is based on the following distribution $W_{Wu} \sim_{\mathcal{H}_0} \chi^2_{1dof}$ where $\mathcal{H}_0 : \lambda_1 = \lambda_0$, however the null hypothesis is equivalent to $\mathcal{H}_{0_{SSI}}^{StatAllele}$ in Equation (2).

Multidimensional IndOR

By dealing with data at the genotype level, Emily (2012) proposed a method to search for interaction based on a biological definition of interaction. Indeed, assuming that conditional to $Y = 0$, X_i and X_j are independent, biological interaction is defined as a departure from independence. Therefore, testing for biological interaction corresponds to testing for the following statistical hypothesis: $\mathcal{H}_0 : X_i \perp\!\!\!\perp X_j | Y = 1$. Such a test refers to a case-only test where hypothesis is made in the control population ($Y = 0$). However, in practice, assuming independence between X_i and X_j is hardly realistic since many biological mechanisms may induce dependence between X_i's, such as Linkage Disequilibrium for instance. In Emily (2012), no assumption is assumed in the control population but focus is made on capturing a variation in the amount of dependency between the two populations $Y = 0$ and $Y = 1$. By using the ratio between the joint distribution and the product of marginal distributions, the statistical hypothesis tested in Emily (2012) can be written as $\forall (x_i, x_j) \in [AA, Aa, aa] \times [BB, Bb, bb]$:

$$\begin{aligned}
\mathcal{H}_0 \quad &: \quad \frac{\mathbb{P}\Big[(X_i,X_j) = (x_i,x_j)|Y=1\Big]}{\mathbb{P}[X_i = x_i|Y=1]\mathbb{P}[X_j = x_j|Y=1]} = \frac{\mathbb{P}[(X_i,X_j) = (x_i,x_j)|Y=0]}{\mathbb{P}[X_i = x_i|Y=0]\mathbb{P}[X_j = x_j|Y=0]} \\
\mathcal{H}_1 \quad &: \quad \frac{\mathbb{P}\Big[(X_i,X_j) = (x_i,x_j)|Y=1\Big]}{\mathbb{P}[X_i = x_i|Y=1]\mathbb{P}[X_j = x_j|Y=1]} \neq \frac{\mathbb{P}[(X_i,X_j) = (x_i,x_j)|Y=0]}{\mathbb{P}[X_i = x_i|Y=0]\mathbb{P}[X_j = x_j|Y=0]}
\end{aligned}$$

It is straightforward to show that such a null hypothesis is equivalent to $\mathcal{H}_{0_{SSI}}^{BioGeno}$ in Equation (5). Using bayes formula, the above hypothesis are equivalent to:

$$\mathcal{H}_0 : \Phi = [0;0;0;0] \text{ and } \mathcal{H}_1 : \Phi \neq [0;0;0;0]$$

where $\Phi = (\varphi_1, \varphi_2, \varphi_3, \varphi_4)$,

 ISSN: 2102-6238

$$\varphi_1 = \log\left(\frac{OR(Aa,Bb)}{OR(Aa)OR(Bb)}\right) \quad ; \quad \varphi_2 = \log\left(\frac{OR(aa,Bb)}{OR(aa)OR(Bb)}\right);$$
$$\varphi_3 = \log\left(\frac{OR(Aa,bb)}{OR(Aa)OR(bb)}\right) \quad ; \quad \varphi_4 = \log\left(\frac{OR(aa,bb)}{OR(aa)OR(bb)}\right);$$

and, by considering AA, BB as the baseline genotype, odds-ratios (OR) can be defined as following:

$$OR(x_1,x_2) = \frac{odds(x_1,x_2)}{odds(AA,BB)} = \frac{\frac{\mathbb{P}[Y=1|X_1=x_1,X_2=x_2]}{\mathbb{P}[Y=0|X_1=x_1,X_2=x_2]}}{\frac{\mathbb{P}[Y=1|X_1=AA,X_2=BB]}{\mathbb{P}[Y=0|X_1=AA,X_2=BB]}}$$

$$OR(x_1) = \frac{odds(x_1)}{odds(AA)} = \frac{\frac{\mathbb{P}[Y=1|X_1=x_1]}{\mathbb{P}[Y=0|X_1=x_1]}}{\frac{\mathbb{P}[Y=1|X_1=AA]}{\mathbb{P}[Y=0|X_1=AA]}} \quad \text{and} \quad OR(x_1) = \frac{odds(x_2)}{odds(BB)} = \frac{\frac{\mathbb{P}[Y=1|X_2=x_b]}{\mathbb{P}[Y=0|X_2=x_b]}}{\frac{\mathbb{P}[Y=1|X_2=BB]}{\mathbb{P}[Y=0|X_2=BB]}}$$

In Emily (2012), to test for $\mathcal{H}_0$, a Wald statistic, IndOR, has been defined as follows:

$$\text{IndOR} = \Phi V_\Phi^{-1} \Phi^t \tag{13}$$

where V_Φ^{-1} is the inverse of variance-covariance matrix for Φ and Φ^t is the transposed vector of Φ. Under the null hypothesis of the same amount of dependence between cases and controls, the score IndOR follows a central χ^2 distribution with four degrees of freedom.

Estimation of the multidimensional vector Φ has been performed using Maximum Likelihood Estimator (MLE) (Thomas, 2004), while an estimation of the covariance matrix V_Φ has been proposed using δ approximation of the counts.

Strengths for Wald-like tests. Compared to regression-based methods, interaction in odds-ratio is treated as a residual term and can implicitly consider nonlinear interaction between two unlinked loci (Ueki and Cordell, 2012). Furthermore, thanks to the large amount of literature regarding odds-ratio, especially in the epidemiology community, asymptotic approximations allow for proposing closed formulations of the statistic thus fastening the computation of the tests and controlling the type-I error rate. As based on genotype data, IndOR is the only Wald-like statistic to have power to detect genotype interaction (Emily, 2012).

Weaknesses for Wald-like tests. The use of one-dimensional statistic in W_{Zhao} and W_{Wu} prevents those methods to detect genotype interaction.

3.3. *Entropy-based methods*

Since 2006, it has been argued that entropy-based methods, that rely on information theory (Shannon, 2001) are particularly powerful and adapted to capture nonlinear relationships between variables (Ferrario and Konig, 2016). More precisely, entropy-based methods rely on a qualitative definition of SNP-SNP interaction: X_i and X_j are said to interact when the strength of the joint prediction ability of X_i and X_j in explaining Y is larger than the sum of the individual prediction abilities of X_i and X_j. Therefore, such methods are based on a measure of the strength of prediction

 ISSN: 2102-6238

ability of a single X_i or a pair (X_i, X_j) in explaining Y defined as the mutual information (Cover and Thomas, 2006).

The mutual information I between two random variables X and Y, with supports denoted respectively by $\mathcal{S}_X$ and $\mathcal{S}_Y$, is defined as follows:

$$I(X,Y) = \sum_{x\in\mathcal{S}_X}\sum_{y\in\mathcal{S}_Y} \mathbb{P}[X=x, Y=y]\log\left(\frac{\mathbb{P}[X=x, Y=y]}{\mathbb{P}[X=x]\mathbb{P}[Y=y]}\right) \tag{14}$$

As detailed in Mielniczuk and Rdzanowski (2017), the mutual information between X and Y is related to the Kullback-Leibler distance between the joint distribution (X,Y) and the product of the marginal distributions in X and Y. Alternatively, the mutual information I can be defined as follows (see Ferrario and Konig (2016) for example):

$$I(X,Y) = H(X)H(X|Y)\ (= H(Y)H(Y|X))$$

where $H(X)$ is the entropy of X:

$$H(X) = \sum_{x\in\mathcal{S}_X} \mathbb{P}[X=x]\log\left(\mathbb{P}[X=x]\right)$$

and $H(X|Y)$ the conditional entropy of X given Y:

$$H(X|Y) = H(X,Y) - H(X)$$

where $H(X,Y)$ is the entropy of the joint variable (X,Y).

Methods based on the mutual information measure compare the strength of the joint prediction ability of X_i and X_j the sum of their individual prediction abilities in explaining Y can be divided into three main classes: (1) information gain methods, (2) relative information gain methods and (3) interaction information methods. All these methods aim at testing the null hypothesis $\mathcal{H}_{0_{SSI}}^{BioGeno}$ in Equation (5).

Information gain methods

Information gain methods are a class of methods that aim at quantifying the gain in mutual information between X_i and X_j by conditioning on the knowledge of Y.

In Fan et al. (2011), the authors proposed to measure gain in information by comparing the mutual information between X_i and X_j conditional to $Y=1$ to the unconditional mutual information. The following statistic has therefore been proposed:

$$IG_{Fan} = I(X_i, X_j|Y=1) - I(X_i, X_j) \tag{15}$$

where:

$$I(X_i, X_j|Y=1) = \sum_{x_i\in\{AA,Aa,aa\}}\sum_{x_j\in\{BB,Bb,bb\}} \mathbb{P}[X_i=x_i, X_j=x_j|Y=1]\log\left(\frac{\mathbb{P}[X_i=x_i, X_j=x_j|Y=1]}{\mathbb{P}[X_i=x_i|Y=1]\mathbb{P}[X_j=x_j|Y=1]}\right)$$

In Chen et al. (2011) and more recently in Su et al. (2015), a similar approach is proposed where the Cumulative Mutual Information $CMI(X_i, X_j, Y)$ is used instead of $I(X_i, X_j|Y=1)$ in Equation (15):

$$IG_{Chen} = CMI(X_i, X_j, Y) - I(X_i, X_j) \tag{16}$$

 ISSN: 2102-6238

where:

$$CMI(X_i, X_j, Y) = \sum_{x_i \in \{AA,Aa,aa\}} \sum_{x_j \in \{BB,Bb,bb\}} \sum_{Y \in \{0,1\}} \mathbb{P}[X_i = x_i, X_j = x_j, Y = y] \log\left(\frac{\mathbb{P}[X_i = x_i, X_j = x_j | Y = y]}{\mathbb{P}[X_i = x_i | Y = y]\mathbb{P}[X_j = x_j | Y = y]}\right)$$

Alternatively, in Kwon et al. (2014), the information gain is measured by comparing the unconditional entropy for Y to its entropy conditional to (X_i, X_j). The following measure has therefore been proposed:

$$IG_{Igent} = H(Y) - H(Y|X_i, X_j) \tag{17}$$

Relative information gain methods

The class of methods based on relative information gain aims at normalizing the information gain and refers to normalized mutual information measures (Yee et al., 2013). For example, several papers proposed a normalized version of IG_{Igent} in Equation (17) (see Dong et al. (2008); Yee et al. (2013); Kwon et al. (2014)) by studying the following measure:

$$RIG_{Igent} = \frac{H(Y) - H(Y|X_i, X_j)}{H(Y)} \tag{18}$$

A slightly different approach is taken by Chattopadhyay et al. (2014), where the normalized entropy score ES is provided as a measure of relative information gain:

$$ES = \frac{\min(H(X_i), H(X_j)) - H(X_i, X_j))}{\min(H(X_i, H(X_j))} \tag{19}$$

Interaction information methods

In Mielniczuk and Rdzanowski (2017) an interaction information score, II is introduced to compared the joint prediction capacity of (X_i, X_j) to the sum of the marginal prediction capacities:

$$II(X_i, X_j, Y) = I((X_i, X_j), Y) - I(X_i, Y) - I(X_j, Y) \tag{20}$$

It is argued that X_i and X_j interact predictively in explaining Y when $II(X_i, X_j, Y)$ is positive.

Strengths for entropy-based methods. The main advantage of entropy-based methods is that they do not rely on any model assumption and have the ability to catch any type of interaction signal, such as for example interaction at the genotype level.

Weaknesses for entropy-based methods. However the practical use of entropy-based methods suffers from several limitations that are all related to the lack of known distribution under the null hypothesis. Significance testing therefore relies on resampling methods, such as permutations, which dramatically increase the computational cost and is likely to generate overfitting.

4. Gene-Gene interaction

In contrast to SNP-level approaches, gene-level testing can help characterizing functional, compositional and statistical interactions (Phillips, 2008). Such tests allow for all the SNPs within the

 ISSN: 2102-6238

TABLE 4. *Summary of the four features, related to the statistical power, of SNP-based statistical methods (Model free, capacity to detect genotype interaction, closed formulation and control of type-I error rate). Each feature is evaluated according to three scales of ability: + for a good ability, ∘ for a moderate ability and - for a poor ability.*

	Model free	Detection of genotype interaction	Closed formulation	Control of the type-I error
Regression-based methods				
1 dof logistic (PLINK)	-	-	∘	+
4 dof logistic, Poisson (BOOST)	-	+	∘	+
Methods based on Wald-like tests				
W_{Zhao}, W_{Wu}	+	-	+	+
IndOR	+	+	+	+
Entropy-based methods				
IG_{Fan}, IG_{Chen}, IG_{Igent}, RIG_{Igent}, ES, II	+	+	-	-

region of a gene to be jointly modeled as a set and can take into account the LD structure within a gene (Huang et al., 2011). Thus, by aggregating signals across variants in a gene, statistical power is likely to be increased in situations when multiple causal interactions influence the phenotype of interest (Wu et al., 2010a). Furthermore, the use of the gene as the statistical unit can greatly facilitate the biological interpretation of findings (Jorgenson and Witte, 2006; Neale and Sham, 2004). As detailed in Equation (6), each gene is a collection of respectively m_1 and m_2 SNPs. The observed genotypes for gene X_1 can be represented by a $n \times m_1$ matrix: $\mathbf{X_1}$ that is further decomposed into a $n_c \times m_1$ matrix $\mathbf{X_1^c}$ and a $n_d \times m_1$ matrix $\mathbf{X_1^d}$ to distinguish between controls ($Y = 0$) and cases ($Y = 1$) observations. Similar notations are used for gene X_2.

From a statistical point-of-view, detecting interaction at the SNP-set level is challenging. Tackling the two issues of SNP-set association and interaction indeed requires the simultaneous modeling of the correlation within and between the two SNP-sets. Nevertheless, the very recent years have seen the development of statistical methods dedicated to the detection of interaction at the SNP-set level. First the issue of detecting SNP-sets interaction can be seen as a **signal detection problem** where interaction is tested for each pair of SNPs within genes. By doing so, a total of $m_1 \times m_2$ tests are performed and those tests are correlated. Therefore testing for gene-gene interaction is considered by testing whether at least one SNP pair is significant by the minimum p-value (*minP*, Emily, 2016a), a Gene-Based Association Test Using Extended Simes (*GATES*, Li et al., 2011), and two truncated tests, the truncated tail strength (*tTS*, Jiang et al., 2011) and the truncated product p-values (*tProd*, Zaykin et al., 2002). Next, instead of combining single SNP-pairs, another approach consists in **reducing the dimensionality** of each gene $\mathbf{X_1}$ and $\mathbf{X_2}$ using a Principal Component Analysis. Interaction is then tested between the retrieved components in each gene in a multivariate logistic model Li et al., 2009). A third class of methods is based on the comparison of the **conditional joint covariance structures** between $\mathbf{X_1}$ and $\mathbf{X_2}$ in controls and cases. Proposed methods aim at modeling the joint distribution of SNPs within and between two genes through Composite Linkage Disequilibrium (*CLD*, Rajapakse et al., 2012), Canonical Correspondance Analysis (*CCA*, Peng et al., 2010), Kernel Canonical Correspondance Analysis (*KCCA*, Larson et al., 2014), Partial Least Square Path Modeling (*PLSPM*, Zhang et al., 2013). A final statistical procedure relies on the Shannon entropy and aims at proposing a Gene-Based Information Gain Method (*GBIGM*, Li et al., 2015). Table 5 provides a summary of

 ISSN: 2102-6238

the underlying null hypothesis addressed by each Gene-Gene method considered in this review.

TABLE 5. *Summary of the Gene-Gene interaction methods and their respective class of underlying null hypothesis.*

Statistical model	Underlying null hypothesis	Reference
minP, GATES, tTS and tProd	Signal detection	Emily (2016a)
PCA	Dimension reduction	Li et al. (2009)
CCA, KCCA, PLSPM, CLD	Equality of Y-conditional covariances structures	Peng et al. (2010); Yuan et al. (2012) Larson et al. (2014); Zhang et al. (2013) Rajapakse et al. (2012)
Entropy	Nullity of the gain in Shannon information	Li et al. (2015) Li et al. (2015)

The remainder of this section is devoted to the detailed presentation of these methods. Furthermore, for each method, a qualitative evaluation is proposed with respect to four main characteristics: (1) the ability to detect non-linear interaction (2) the robustness to the structure of the data, (3) the need for manual parameterization and (4) the computational efficiency. A summary of the respective advantages and drawbacks is proposed in Table 6.

4.1. Signal detection: combination of statistics

Testing for interaction between two sets of SNPs can be addressed by considering such combination as an extension of SNP-SNP interaction tests. Gene-gene interaction can indeed be performed by applying SNP-SNP interaction tests to all possible SNP pairs between two genes, thus resulting in a total of $m_1 \times m_2$ tests. Let denote W_{ij} the statistic used to test the interaction between X_i^1, the i^{th} SNP of $\mathbf{X_1}$, and X_j^2, the j^{th} SNP of $\mathbf{X_2}$. All pairwise tests are summarized into a $m_1 \times m_2$ vector of statistics $\mathbb{W} = [W_{11}, \ldots, W_{m_1 m_2}]$ (see Figure 2). Testing for the significance of the vector $\mathbb{W}$ therefore consists in combining a set of statistics and can be formalized as a signal detection problem as defined by Donoho and Jin (2004). If we consider, without loss of generality, that under $\mathcal{H}_0$: $W_{ij} = 0$, then testing the significance of $\mathbb{W}$ can be performed by testing the following statistical hypothesis:

$$\begin{aligned} \mathcal{H}_0 : &\quad \forall 1 \leq i \leq m_1 \text{ and } \forall 1 \leq j \leq m_2, \;\; W_{i,j} = 0, \\ \mathcal{H}_1 : &\quad \exists (i,j) \text{ where } 1 \leq i \leq m_1 \text{ and } 1 \leq j \leq m_2, \;\; W_{i,j} \neq 0. \end{aligned} \tag{21}$$

Since the X_i^1's (resp. X_i^2's) are expected to be correlated within $\mathbf{X_1}$ (resp. $\mathbf{X_2}$) and each X_i is used in several pairs, the elements of $\mathbb{W}$ are not independent. Therefore, the covariance matrix for $\mathbb{W}$, denoted by Σ and displayed in Figure 2, is not diagonal and the issue of detecting interaction between two SNP sets falls into the paradigm of signal detection under dependence (Hall and Jin, 2010).

In Emily (2016a), the author introduced a framework that aims at aggregating p-values obtained at the SNP level into a test at the gene level. Let consider a standard logistic regression to model the association between the two SNPs, X_i^1 and X_j^2, and the phenotype Y:

$$\log\left(\frac{\mathbb{P}[Y=1|X_i^1=x_1, X_j^2=x_2]}{1-\mathbb{P}[Y=1|X_i^1=x_1, X_j^2=x_2]}\right) = \beta_0^{i,j} + \beta_1^{i,j} x_1 + \beta_2^{i,j} x_2 + \beta_3^{i,j} x_1 x_2 \tag{22}$$

 ISSN: 2102-6238

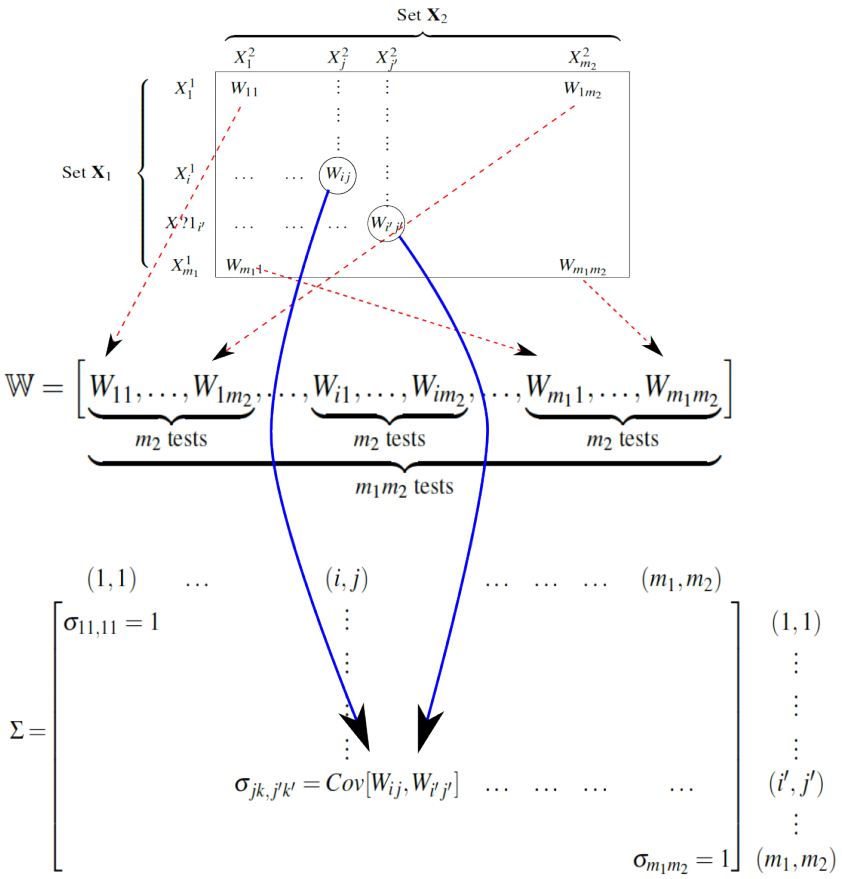

FIGURE 2. *Flowchart of the signal detection approach in gene-gene testing. All pairwise tests between a SNP from the set* $\mathbf{X_1}$ *and a SNP from the set* $\mathbf{X_2}$ *are stored in* $m_1 \times m_2$ *vector* $\mathbb{W}$*. The covariance matrix of* $\mathbb{W}$ *is given by* Σ*.*

where $\beta_3^{i,j}$ is interpreted as the regression coefficient that weight of the interaction between the two SNPs. The interaction between the two SNPs is then tested by means of the following statistical null and alternative hypothesis: $\mathcal{H}_0 : \beta_3^{i,j} = 0$ and $\mathcal{H}_1 : \beta_3^{i,j} \neq 0$. To test $\mathcal{H}_0$ against $\mathcal{H}_1$, we used the following Wald statistic:

$$W_{ij} = \frac{\beta_3^{i,j}}{\sigma\left(\beta_3^{i,j}\right)} \tag{23}$$

Therefore under $\mathcal{H}_0$ in Equation (21), we have:

$$\mathbb{W} = [W_{11}, \ldots, W_{m_1 m_2}] \sim \mathcal{N}(\mathbf{0}, \Sigma),$$

where $\mathcal{N}(\mathbf{0}, \Sigma)$ is the multivariate normal density with mean $\mathbf{0}$, the $m_1 \times m_2$ null vector, and covariance matrix Σ for which an estimation is proposed in Emily (2016a). Four main methods developed to aggregate p-values are proposed: the minimum p-value minP (Liu et al., 2010), the

 ISSN: 2102-6238

GATES procedure (Li et al., 2011) and two truncated tests, the truncated product p-values (Zaykin et al., 2002) and the truncated tail strength (Jiang et al., 2011).

minP

In the minP procedure, to combine a set of p-values, the maximum of the absolute values for the observed Wald statistics is compared to the asymptotic distribution expected under $\mathcal{H}_0$. More precisely the probability, minP, that at least one absolute value for Wald statistics is as large as the maximum of the observed absolute values under the null hypothesis is computed. Let $\mathbb{Z} = [Z_1, \ldots, Z_{m_1 m_2}]$ be a multivariate Gaussian random vector with the following distribution $\mathbb{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma)$ and $W_{\max} = \max\{|W_{11}|, \ldots, |W_{m_1 m_2}|\}$ be the maximum of the absolute values for the observed Wald statistics. Thus, the minP probability is obtained by the following formula:

$$\text{minP} = 1 - \mathbb{P}\Big[\max(|Z_1|, |Z_2|, \ldots, |Z_{m_1 m_2}|) < W_{\max}\Big]. \tag{24}$$

Since the SNP-SNP interaction test is two-sided, one can remark that $W_{\max} = \Phi^{-1}(1 - P_{\min}/2)$, where Φ is the standard normal distribution function and $P_{\min}$ the minimum of the observed p-values. Equation (24) is then equivalent to the one proposed by Conneely and Boehnke in (Conneely and Boehnke, 2007).

Strengths. As shown by Emily (2016a), minP is highly robust to a wide range of data structure as well as nature of signal. It is also easy to run since no parameters has to be set by the user.

Weaknesses. However, based on the definition of W_{ij} in Equation (23), the minP method is not designed to catch non-linear interaction. Furthermore, the computation of Equation (24) requires the calculation of the probability distribution of a multivariate normal random variable. Such calculation can be approximated by the `pmvnorm` function from the R package `mvtnorm` (Genz and Bretz, 2009). However, the function `pmvnorm` is applicable to arbitrary covariance structures and dimensions up to 1000. Some heuristics are necessary to scale up to more than 1000 pairs of SNPs.

GATES

The GATES procedure, proposed by (Li et al., 2011), is an extension of the Simes procedure used to assess the gene level association significance. Let $p_{(1)}, \ldots, p_{(m_1 m_2)}$ be the ascending SNPxSNP interaction $m_1 \times m_2$ p-values, GATES p-value is then defined in Equation (25):

$$\text{p}_{GATES} = \min\left(\frac{mep_{(1)}}{me_{(1)}}, \frac{mep_{(2)}}{me_{(2)}}, \ldots, \frac{mep_{(m_1 m_2)}}{me_{(m_1 m_2)}}\right) \tag{25}$$

where me is the number of effective tests among the $m_1 \times m_2$ tests and $me_{(i)}$ the number of effective tests among the i most significative tests associated with the lowest order p-values $p_{(1)}, \ldots, p_{(i)}$. The number of effective tests ought to characterize the number of independent tests equivalent to the correlated tests that are really performed and is often used to account for dependence in a multiple testing correction.

Although no formal definition of the number of effective tests has been formulated in the literature, several procedures have been proposed to estimate such number. All methods are based on a transformation of the set of eigenvalues of the SNP covariance matrix assuming that (1) if the

 ISSN: 2102-6238

SNPs are independent, the number of effective tests is the total number of tests, (2) if the absolute value of the correlation between any pair of SNPs is equal to 1, the number of effective tests is 1. The most popular calculation of the number of effective tests are: Cheverud-Nyholt method (Cheverud, 2001; Nyholt, 2004), Keff method (Moskvina and Schmidt, 2008), Li and Ji method (Li and Ji, 2005) and Galwey (Galwey, 2009).

Strengths. In principle the GATES method is an attractive method to cope with correlated data through the use of the effective number of tests.

Weaknesses. However, in practice the estimation of the effective number of tests suffers from major limitations. It has been shown that the validity of the estimation of the number of effective tests is very sensitive to SNPs set characteristics Hendricks et al. (2014). Furthermore, the computation of the effective number of tests requires the computation of eigen decomposition which is known to raise issue in a high-dimensional setting. Finally, similar to minP, GATES is not designed to detect non-linear interaction.

tTS and tProd

tTS and tProd procedures are two truncated tail strength methods that aim at combining signals from all single-SNP p-values less than a predefined cutoff value (Jiang et al., 2011). Denoting by τ the cutoff value, the two truncated p-values are defined as follows (Zaykin et al., 2002):

$$tTS = \frac{1}{m_1 m_2} \sum_{i=1}^{m_1 m_2} \mathbb{I}(p_{(i)} < \tau)\left(1 - p_{(i)} \frac{m_1 m_2 + 1}{i}\right) \tag{26}$$

$$tProd = \prod_{i=1}^{m_1 m_2} p_i^{\mathbb{I}(p_i < \tau)} \tag{27}$$

where $\mathbb{I}$ is the indicator function. When p-values are correlated, the null distributions of tTS and $tProd$ are unknown. Following the approach proposed by (Zaykin et al., 2002), a p-value is obtained by computing an empirical null distribution using Monte-Carlo (MC) simulations. For each MC iteration, an empirical value for tTS (or $tProd$) is obtained by simulating a vector of W_{jk} with respect to a multivariate normal distribution with a vector of 0 means and $\widehat{\Sigma}$ as covariance matrix. The empirical p-value is calculated as the proportion of simulated statistics larger than the observed statistic on the "true" set of W_{jk}.

Strengths. The computation of tTS and tProd is rather simple and fast which allows their use for large SNP sets.

Weaknesses. However, in practice tTS and tProd both require the tuning of a cutoff value τ. The choice of τ is critical to warrant a control of the type-I error and to improve the statistical power (Zaykin et al., 2002). An optimal choice for τ is very sensitive to the number of p-values to be combined and to the correlation structure among those p-values. Furthermore, as based on Equation (23), tTS and tProd are not designed to detect non-linear interaction.

 ISSN: 2102-6238

4.2. *Multidimensional methods*

Instead of combining tests at the SNP level, a class of methods aim at modeling the joint distribution of SNPs within and between two genes in a multidimensional settings. Several multidimensional methods have been proposed in the literature to tackle the issue of gene-gene interaction using either Principal Components Analysis, U-like statistics or Modeling of covariance structures or entropy measures.

Principal Component Analysis - PCA

In Li et al. (2009), the authors proposed to test the interaction between the two sets $\mathbf{X_1}$ and $\mathbf{X_2}$ by comparing their respective decomposition in principal components. More precisely, a likelihood ratio test is performed to compare the model $\mathcal{M}_{Inter}$ to the model $\mathcal{M}_{No}$, where $\mathcal{M}_{Inter}$ refers to the logistic model including interaction effects while $\mathcal{M}_{No}$ does not consider interaction terms. Formally, $\mathcal{M}_{Inter}$ is defined in Equation (28):

$$\text{logit}\left(\mathbb{P}\left[Y=1|PC^1_{\mathbf{X_1}}\dots PC^{n_1}_{\mathbf{X_1}}, PC^1_{\mathbf{X_2}}\dots PC^{n_2}_{\mathbf{X_2}}\right]\right) = \beta_0 + \sum_{i=1}^{n_1} PC^i_{\mathbf{X_1}} + \sum_{j=1}^{n_2} PC^j_{\mathbf{X_2}} + \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} PC^i_{\mathbf{X_1}} PC^j_{\mathbf{X_2}} \tag{28}$$

and $\mathcal{M}_{No}$ in Equation (29):

$$\text{logit}\left(\mathbb{P}\left[Y=1|PC^1_{\mathbf{X_1}}\dots PC^{n_1}_{\mathbf{X_1}}, PC^1_{X_2}\dots PC^{n_2}_{X_2}\right]\right) = \beta_0 + \sum_{i=1}^{n_1} PC^i_{\mathbf{X_1}} + \sum_{j=1}^{n_2} PC^j_{X_2} \tag{29}$$

In models $\mathcal{M}_{Inter}$ and $\mathcal{M}_{No}$, $PC^i_{\mathbf{X_1}}$ and $PC^j_{\mathbf{X_2}}$ are the i^{th} principal component of $\mathbf{X_1}$ and the j^{th} principal component of $\mathbf{X_2}$. The number of principal components, n_1 and n_2, kept in the interaction test is determined by the percentage of inertia retrieved by the PCA. Such a percentage is fixed beforehand and Li et al. (2009) suggested to retrieve 80% of inertia for both $\mathbf{X_1}$ and $\mathbf{X_2}$.

Strengths. The computation of the PCA method is very efficient and allows for large scale testing. Furthermore it is quiet robust to the structure of the data since PCA aims at reducing the dimensionality of the data.

Weaknesses. However, the practical use of the PCA method require the choice of the amount of variability the PCA should retrieve. Such a choice is likely to impact the stability of the result as well as the computational performance. Indeed, the higher the percentage of inertia is, the higher the number of components is kept in the regression models (28) and (29).

U-like statistics

Several studies have proposed statistical tests based on a U-like statistic that can be defined as follows:

$$U = \frac{z^d - z^c}{\sqrt{\mathbb{V}(z^d - z^c)}}. \tag{30}$$

The main idea behind Equation (30) is to measure relationship between $\mathbf{X_1}$ and $\mathbf{X_2}$ in the two subpopulations of cases ($Y = 1$) and controls ($Y = 0$) separately. Two measures of interaction are

 ISSN: 2102-6238

then calculated, z^d in cases and z^c in controls, and compared by normalizing the difference $z^d - z^c$ in Equation (30). Therefore, if the absolute value of the U statistic is high enough then it means that the amount of interaction of $\mathbf{X_1}$ and $\mathbf{X_2}$ is different between $Y = 1$ and $Y = 0$, thus indicating that Y is associated with the interaction between $\mathbf{X_1}$ and $\mathbf{X_2}$.

In Peng et al. (2010), the authors proposed to used the following U-statistic:

$$U_{CCA} = \frac{z^d_{CCA} - z^c_{CCA}}{\sqrt{\mathbb{V}\left(z^d_{CCA} - z^c_{CCA}\right)}} \tag{31}$$

where $z^d_{CCA} = \frac{1}{2}\left(\log(1 + r^d) - \log(1 - r^d)\right)$ is the Fisher transformation of r^d defined as the maximal canonical correlation coefficient between $\mathbf{X_1^d}$ and $\mathbf{X_2^d}$. z^c_{CCA} is defined similarly by considering the canonical correlation is the control population. The use of the Fisher transformation allows for U_{CCA} to follow a standard normal distribution under $\mathcal{H}_0$ of no association: $U_{CCA} \sim_{\mathcal{H}_0} \mathcal{N}(0, 1)$. However $\mathbb{V}\left(z^d_{CCA} - z^c_{CCA}\right)$ in Equation (31) is not known and Peng et al. (2010) proposed a bootstrap estimation for $\mathbb{V}\left(z^d_{CCA} - z^c_{CCA}\right)$.

By definition of the canonical correlation analysis, U_{CCA} is limited to detect linear interaction between $\mathbf{X_1}$ and $\mathbf{X_2}$. To overcome such limitation, U_{CCA} has been extended to U_{KCCA} in Yuan et al. (2012) and Larson et al. (2014) by using a kernelized version of the canonical correlation as follows:

$$U_{KCCA} = \frac{z^d_{KCCA} - z^c_{KCCA}}{\sqrt{\mathbb{V}\left(z^d_{KCCA} - z^c_{KCCA}\right)}} \tag{32}$$

Compared to z^d_{CCA} in U_{CCA}, z^d_{KCCA} is defined as the Fisher transformation of the maximum kernel canonical coefficient. In Larson et al. (2014), the authors used a Gaussian mapping of the original data by applying a Radial Basis kernel function. As for U_{CCA}, based on the Fisher transformation the significance U_{KCCA} can be tested by comparing the observed value with the standard gaussian distribution since $U_{KCCA} \sim_{\mathcal{H}_0} \mathcal{N}(0, 1)$. Similar to U_{CCA}, the variance in Equation (32) is unknown and can be estimated using resampling techniques, such as bootstrap as proposed in Yuan et al. (2012) and Larson et al. (2014).

A third U-statistic, called U_{PLSPM} and based on Partial Least Square Path Modelling, has been introduced in the litterature by Zhang et al. (2013). More precisely, U_{PLSPM} is defined as follows:

$$U_{PLSPM} = \frac{z^d_{PLSPM} - z^c_{PLSPM}}{\sqrt{\mathbb{V}\left(z^d_{PLSPM} - z^c_{PLSPM}\right)}} \tag{33}$$

where z^d_{PLSPM} (resp. z^c_{PLSPM}) is defined as the path coefficient between $\mathbf{X_1^d}$ and $\mathbf{X_2^d}$ (resp. $\mathbf{X_1^c}$ and $\mathbf{X_2^c}$). Since the distribution of the path coefficient is not known, the distribution of U_{PLSPM} under $\mathcal{H}_0$ is not known. To test the significance of U_{PLSPM}, Zhang et al. (2013) therefore proposed a permutation procedure where the distribution of U_{PLSPM} under $\mathcal{H}_0$ is estimated by permuting the observed value for Y.

Strengths. U-like statistics consists in an attractive way to summarize the multidimensionality of the interaction into a one-dimensional statistic. It can be designed to catch non-liner interactions,

 ISSN: 2102-6238

through the use of Kernel Correspondence Analysis or Partial Least Square Path Modeling. Furthermore, it can be computationally efficient with the Canonical Correlation Analysis where no additional parameter needs to be set.

Weaknesses. However, with the KCCA, the choice of the kernel is crucial in terms of power and drastically decreases the computational efficiency. Finally, the main limitation of U-like statistics is the lack of robustness to data structure. Such a characteristic is mainly due to the underlying gaussian assumption of the compared coefficients, namely z^d and z^c, in Equation (30).

Covariance structures based method

In Rajapakse et al. (2012), the authors proposed to test the association between Y and the interaction between two sets $\mathbf{X_1}$ and $\mathbf{X_2}$ by comparing the covariance structures in cases ($Y = 1$) and controls ($Y = 0$). More precisely, Rajapakse et al. (2012) introduced the Composite Linkage Disequilibrium (CLD) method that is based on the normalized quadratic distance (NQD) and is defined in Equation (34):

$$\delta^2 = \text{tr.}\left((\tilde{D} - \tilde{C})W^{-1}(\tilde{D} - \tilde{C})W^{-1}\right) \tag{34}$$

where $\tilde{D}$, $\tilde{C}$ and W are three $(m_1 + m_2) \times (m_1 + m_2)$ matrices of the covariance between the whole set of SNPs that combines SNPs from both genes. More precisely, $\tilde{D}$ and $\tilde{C}$ are defined as follows:

$$\tilde{D} = \begin{bmatrix} W_{11} & D_{12} \\ D_{21} & W_{22} \end{bmatrix} \qquad \tilde{C} = \begin{bmatrix} W_{11} & C_{12} \\ C_{21} & W_{22} \end{bmatrix} \tag{35}$$

where W_{11} (resp. W_{22}) is the pooled estimate of the covariance matrix for $\mathbf{X_1}$ (resp. $\mathbf{X_2}$, $D_{12}(= D'_{21})$ and $C_{12}(= C'_{21})$ are the sample covariance matrix between the two genes estimated from $\left(\mathbf{X_1^d}, \mathbf{X_2^d}\right)$ and $\left(\mathbf{X_1^c}, \mathbf{X_2^c}\right)$ respectively. In more details, the sample covariance matrices in cases, denoted by D, and in controls, denoted by C, can be partitioned in 4 blocks as follows in Equation (36):

$$D = \text{Cov}\left(\mathbf{X_1^d}, \mathbf{X_2^d}\right) = \begin{bmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{bmatrix} \qquad C = \text{Cov}\left(\mathbf{X_1^c}, \mathbf{X_2^c}\right) = \begin{bmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{bmatrix} \tag{36}$$

The pooled estimate of the covariance matrix, W, can thus be obtained by Equation (37):

$$W = \frac{n_c C + n_d D}{n_c + n_d} = \begin{bmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{bmatrix} \tag{37}$$

Since the distribution of δ^2 is not known under the null hypothesis, significance testing is performed using permutation tests, as proposed by (Rajapakse et al., 2012).

Strengths. The main advantage of the CLD method is its flexibility to detect a large panel of interaction signal including non-linear interaction. By including genetical knowledge in the decomposition of covariance, CLD is robust to various data structure.

Weaknesses. Conversely, the use of arguments from genetics is a limitation of the use of CLD outside association genetics. Nevertheless, the main limitation of CLD is the use of permutation to test for significance. Although the computation of δ^2 in Equation (34) is relatively fast, a large of permutation is required to reach significance level, thus substantially reducing the computational cost.

 ISSN: 2102-6238

Entropy-based methods

As for the investigation of interaction at the individual level of the X_i's, entropy-based methods are appropriate to detect nonlinear interactions between two sets of X_i (Shannon, 2001). Therefore, methods based on information gain have been extended to detect interaction at the level of the gene. For example, the Gene based Information Gain Method (GBIGM) method, introduced by (Li et al., 2015), is based on the information gain rate $\Delta R_{1,2}$. $\Delta R_{1,2}$ is defined as in Equation (38):

$$\Delta R_{1,2} = \frac{\min(H_1, H_2) - H_{1,2}}{\min(H_1, H_2)} \tag{38}$$

where H_1, H_2, $H_{1,2}$ are the conditional entropies (given $\mathbf{Y}$) of $\mathbf{X_1}$, $\mathbf{X_2}$ and the pooled SNP set $(\mathbf{X_1}, \mathbf{X_2})$ respectively. Assuming that $H(.)$ is the classical entropy function, we have:

$$H_1 = H(\mathbf{Y}, \mathbf{X_1}) - H(\mathbf{X_1}) \tag{39}$$
$$H_2 = H(\mathbf{Y}, \mathbf{X_2}) - H(\mathbf{X_2}) \tag{40}$$
$$H_{1,2} = H(\mathbf{Y}, \mathbf{X_1}, \mathbf{X_2}) - H(\mathbf{X_1}, \mathbf{X_2}) \tag{41}$$

Since the distribution of $\Delta R_{1,2}$ is unknown, the significance testing is performed by permutations as suggested by (Li et al., 2015).

Strengths. The main advantage of the GIBGM method is its ability to detect non-linear interaction. As being based on measures from the information theory, no additional parameter needs to be set prior to the analysis.

Weaknesses. However, in practice GBIGM suffers from several limitations. First, as shown in Emily (2016a), GBIGM is very sensitive to the data structure and control of the type-I error is not guarantee. Next, GBIGM is highly time consuming because of the permutation based significance testing and since the computation of $\Delta R_{1,2}$ in Equation (38) can substantially increase with the size of the SNP set.

TABLE 6. *Summary of the four features, related to the statistical power, of gene-based statistical methods (Capacity to detect non-linear interaction, robustness to data structure, robustness to parameter settings and computational efficiency). Each feature is evaluated according to five scales of ability: ++ for a very good ability, + for a good ability, ∘ for an average ability, - for a poor ability and – for a very poor ability.*

	Detection of non-linear interaction	Robustness to data structure	Parameters free	Computational efficiency
Combination of tests				
minP	- -	+ +	+ +	-
GATES	- -	∘	-	∘
tTS and tProd	- -	-	- -	+
Multidimensional methods				
PCA	- -	+	∘	+ +
U_{CCA}	-	-	+ +	+ +
U_{KCCA}	+ +	-	-	- -
$U_{PLS\,PM}$	+	- -	+ +	- -
Covariance modeling (CLD)	+	+	+ +	∘
Entropy (GBIGM)	+ +	- -	+ +	- -

 ISSN: 2102-6238

5. Large-scale testing for interaction

The ultimate goal of statistical methods for detecting interaction in GWAS is to analyze the whole genome at a time, which corresponds to deal with interaction between hundred of thousand of variables. Such a deep modification in the scale at which interaction can be investigated not only raises the question of the scalability of the methods introduced in Sections 3 and 4, but also allows the use of multidimensional regression models or machine learning approaches. Those two latter classes of methods do not restrict themselves to pairwise testing but have the potential to integrate higher order of interactions. Because of the enormous number of potential interacting signals, a large number of statistical strategies have been proposed to test for pairwise interaction at the genome level. From a statistical point-of-view, these strategies can be grouped into 3 main classes. First, the genome can be seen as a large set of SNPs or genes. Therefore a first strategy consists in extending the methods proposed at the SNP level (Section 3) or at the Gene level (Section 4) to more than one SNP pair or one pair of SNP sets. In a second modeling strategy, rather than aggregating pairwise tests, multidimensional regression can be performed to simultaneously account for all pairs. In this context, to cope with the curse of dimensionality, penalized regressions are encouraged. However, regression-based methods are often criticized for their inability to deal with nonlinear models and with high-dimensional data that contain many potentially interacting predictor variables (McKinney et al., 2006; Koo et al., 2013; Mackay and Moore, 2014). In this context, a third class, refers to as machine-learning methods, is often cited as an attractive alternative. These methods do not fit a single prespecified model, nor do they attempt an exhaustive search, but rather they attempt to step through the space of possible models, including potentially large numbers of main effects and multiway interactions, in a computationally efficient way (Cordell, 2009).

For the remainder of this section, we focus on representative methods for each of the three classes: (1) exhaustive search method, (2) regression-based methods and (3) machine-learning methods. Each method is evaluated through six main features known to impact their respective computational and statistical efficiencies: the ability to detect pure epistasis, to identify a causal pair and to scale up to genome level, the sensitivity to an underlying model assumption and to the hand-tuning of parameters as well as the robustness to the data structure. Table 7 provides a qualitative summary of these characteristics per method.

5.1. Exhaustive pairwise testing

A natural extension of the methods proposed in Sections 3 and 4 is to perform an exhaustive pairwise testing at the genome scale. When considering the SNP level, as displayed in the design (1), an exhaustive strategy consists in testing all pairs (X_i, X_j) with one of the SNP-SNP method presented in Section 3, thus resulting in set of $p(p-1)/2$ SNP-SNP tests. Although several hypothesis can be tested by combining these $p(p-1)/2$ tests, it is common to test whether none of the pair (X_i, X_j) is associated with Y. Such an hypothesis can be formalized as follows:

$$\mathcal{H}_0 = \left\{\forall (i,j) \in [1 \dots p]^2, Y \perp\!\!\!\perp (X_i : X_j)\right\} \text{ vs. } \mathcal{H}_1 = \left\{\exists (i,j) \in [1 \dots p]^2, Y \not\perp\!\!\!\perp (X_i : X_j)\right\} \tag{42}$$

When considering the gene level, we can assume that the genome is decomposed into L predefined blocks: $\mathbf{X}_1, \dots, \mathbf{X}_L$. An exhaustive strategy at the gene level thus consists in testing all

 ISSN: 2102-6238

pairs of blocks with one of the method presented in Section 4. It results in a total of $L(L-1)/2$ tests from which the following hypothesis can be tested:

$$\mathcal{H}_0 = \left\{\forall(\ell_1,\ell_2) \in [1\dots L]^2, Y \perp\!\!\!\perp (\mathbf{X}_{\ell_1} : \mathbb{X}_{\ell_2}) \text{ vs. } \mathcal{H}_1 = \left\{\exists(\ell_1,\ell_2) \in [1\dots L]^2, Y \not\perp\!\!\!\perp (\mathbb{X}_{\ell_1} : \mathbb{X}_{\ell_2})\right\}\right. \quad (43)$$

Strengths. One of the main advantage of exhaustive search is that such strategies are based on pairs testing and thus are appropriate to detect pure epistasis. Furthermore SNP-SNP exhaustive testing is dedicated to the identification of causal pairs thus proposing a direct interpretation of the findings. However for gene-gene interaction, the purpose is not to detect a causal pair of SNPs but a causal pair of genes which, in some cases, can facilitate the functional interpretation of the results. Furthermore, since SNP-SNP and Gene-Gene methods are robust to the design of experiments, exhaustive tests are robust to the structure of the data set. However, it is noteworthy that the correlation structure of the variables has to be accounted for in the hypothesis testing (42) and (43). Considering that the structure of the genome is complex and multilevel, an appropriate correction for multiple testing is not straightforward, especially for SNP-SNP interaction. Finally, due to the exhaustivity of the approach no parameter has to be set before running the analysis.

Weaknesses. Conversely, the exhaustive search is computationaly intensive. For SNP-SNP exhaustive testing and using a DNA chip with 500 000 SNPs, the total number of interaction tests reach 125 billions. Although computational cost can be reduced using parallel coding and high-performance grid computing, performing an exhaustive search at the genome scale is still challenging (Prabhu and Pe'er, 2012). Another main drawback of the exhaustive approach is the lack of flexibility of such methods to detect a wide range of interaction models.

5.2. *Regression-based methods*

Logic regression

Logic regression has been introduced to reduce the dimensionality induced by the amount of variable combinations (Kooperberg and Ruczinski, 2005). This method uses Boolean logic to select a subset of categorical variables that are associated with the disease outcome. The categories of the selected variables are converted into binary variables, and using logic models (*i.e.* logic expressions involving binary variables), a logistic model is fitted as follows:

$$\log\left(\frac{\mathbb{P}[Y=1|X=x]}{1-\mathbb{P}[Y=1|X=x=(x_1,\dots,x_p)]}\right) = \beta_0 + \sum_{i=1}^{K}\beta_i L_i(x)$$

where Y is the disease status and L_i is the Boolean expression of the p predictors $(x_1,\dots,x_p)$ such as $L_i = (X_1 \text{ OR } X_3) \text{ AND } X_2$. L_i is referred as a logic tree as the L_i are organized in a tree form (see Figure 3). Using this logic tree representation it is possible to obtain any other logic tree by a finite number of operations such as growing of branches, pruning of branches, and changing of leaves.

As proposed by Kooperberg and Ruczinski (2005), Markov Chain Monte Carlo approaches, such as simulated annealing, can be used to select logic trees. Extensions of the original method have been proposed to measure the importance of identified interactions (Schwender and Ickstadt, 2008) and to propose a test of the importance of the variables involved in detected combinations (Schwender et al., 2011).

 ISSN: 2102-6238

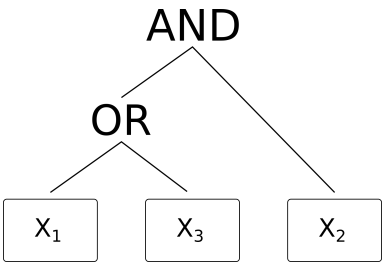

FIGURE 3. *Logic tree representing the logic expression* $L_i = (X_1 OR X_3) AND X_2$

Strengths. The main advantages of logic regression is its ability to catch interaction and to identify causal pairs of variables. Furthermore, logic regression is robust to the design of the dataset and correlation structure between variables can be accounted for in the model.

Weaknesses. Conversely, logic regression is based on the Boolean transformation of the data that need to be set by the user. Although model selection techniques (such as stepwise selection) can be performed to select the most appropriate Boolean transformations, it has to be balanced with an increase of the overfitting. Finally, despite the efforts put to propose efficient solution of the model, its use at the genome scale is still an issue.

Penalized logistic regression

When considering the multivariate set of X_i to explain the disease status Y, it can be natural to consider a multivariate logistic regression. Moreover in the context of the detection of pairwise interaction, the logistic model can be written as follows:

$$\log\left(\frac{\mathbb{P}[Y=1|X=x]}{\mathbb{P}[Y=0|X=x]}\right) = \alpha + \sum_{i=1}^{p}\beta_i x_i + \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\gamma_{ij}x_i x_j \tag{44}$$

Since GWAS fall into the paradigm of high-dimensional data, estimation of regression coefficients can be performed by penalizing the log-likelihood of model (47), denoted by $\ell(\alpha,\beta,\gamma)$, where $\beta = [\beta_1,\ldots,\beta_p]$ and $\gamma = [\gamma_{12},\ldots,\gamma_{p-1p}]$.

In Park and Hastie (2008), regression coefficients are estimated by maximizing the log-likelihood subject to a size constraint on L_2 norm of the coefficients that corresponds to minimizing the following equation:

$$\widehat{(\alpha,\beta,\gamma)}_{L_2} = \operatorname{argmin}\left(-\ell(\alpha,\beta,\gamma) + \frac{\lambda}{2}\|(\beta,\gamma)\|_2^2\right) \tag{45}$$

According to the authors, the use of the L_2 penalty has a number of attractive properties that overcome usual limitations, such as the stability of estimation, the multicollinearity and the presence of zero-cells in contingency tables Park and Hastie (2008).

However, in the context of variable (or pair of interacting variables) selection, it is preferable to perform an L_1 instead of an L_2 penalization. In principle the least absolute shrinkage and selection operator (LASSO) uses the L_1 penalty to perform both variable selection and shrinkage

 ISSN: 2102-6238

and estimates the coefficients of model (47) as by minimizing:

$$(\widehat{\alpha,\beta,\gamma})_{LASSO} = \text{argmin}\left(-\ell(\alpha,\beta,\gamma) + \lambda\left(\sum_{i=1}^{p}|\beta_i| + \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}|\gamma_{ij}|\right)\right) \tag{46}$$

In Wu et al. (2009), the authors proposed a two-step procedure to identify potential interaction at the genome scale. In a first step, a set of individual variables X_i is selected using a LASSO procedure based on logistic regression without interaction coefficient (*i.e.* model in Equation (47) with $\gamma_{ij} = 0$ for all i and j). In a second step, a second penalized logistic model is estimated based on the variables selected in the first step and by accounting for interaction between those variables. Such a strategy induces a hierarchy that restricts an interaction to be included in the model only if both variables are marginally important. In order to formalize this hierarchy, a set of convex constraints has been added to the LASSO to produce sparse interaction (Bien et al., 2013). An extension of the hierarchical interaction in the LASSO aims at proposing progressive penalization in order to allow a computationally fast penalization (Zhu et al., 2014).

Identification of interaction between set of SNPs can also be performed by considering a group-LASSO penalty as introduced by Meier et al. (2008). Such approach has been proposed in the context of genetic association studies in Yang et al. (2010) and more recently in Stanislas et al. (2017).

More recently, Gao et al. (2014) proposed the use of a LASSO penalization to estimate high order interaction in GWAS. The authors extended the model in Equation (47) by accounting for high-order interactions effects as follows:

$$\log\left(\frac{\mathbb{P}[Y=1|X=x]}{\mathbb{P}[Y=0|X=x]}\right) = \alpha + \sum_{i=1}^{p}\beta_i x_i + \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}\gamma_{ij}x_i x_j + \sum_{i=1}^{p-2}\sum_{j=i+1}^{p-1}\sum_{k=j+1}^{p}\delta_{ijk}x_i x_j x_k + \ldots \tag{47}$$

As an extension of the two steps procedure proposed by Wu et al. (2009), Gao et al. (2014) proposed a forward LASSO shrinkage estimator by first selecting variables with a marginal effects. Then k-order interaction terms are recursively included in the model using the following scheme: (1) form $k-$order interaction terms among the whole sets of variables and variables with non-zero $k-1$-way interaction effect, (2) shrink all k-order interactions, $k-1$-order interactions ... to zero in a LASSO regression by minimizing Equation 48:

$$(\widehat{\alpha,\beta,\gamma,\delta,\ldots})_{LASSO\_HO} = \text{argmin}\left(-\ell(\alpha,\beta,\gamma) + \lambda\left(\sum_{i=1}^{p}|\beta_i| + \sum_{i=1}^{p-1}\sum_{j=i+1}^{p}|\gamma_{ij}|\right) + \left(\sum_{i=1}^{p-2}\sum_{j=i+1}^{p-1}\sum_{k=j+1}^{p}|\delta_{ijk}|\right)\ldots\right) \tag{48}$$

Strengths. As for logic regression, the main advantage of penalized logistic regression is the ability to catch interaction and to identify causal pairs of variables.

Weaknesses. However, the practical application of a penalized logistic regression model at the genome scale raised several issues. First, the computation of the LASSO algorithm (and

 ISSN: 2102-6238

subsequent versions) at the genome scale require the estimation of several parameters, especially the regularized parameters. Such parameters are usually estimated using cross-validation which can hardly be performed genome-wide. Finally, LASSO methods are known to be unstable when dealing with correlated data. Therefore, in the context of GWAS, LASSO method are likely to lack in robustness to the data structure.

5.3. Machine learning methods

With the emergence of big data, the last few years have seen the development of numerous machine learning methods in many fields. Gene-gene interaction detection in GWAS is not exempt from this rule. Machine learning methods that can be adapted to gene-gene interaction fall into two main classes: tree-based methods and pattern recognition methods. In the remainder of this section, we focus on two tree-based methods (Classification and regression trees and random forests) and on three pattern recognition methods (Multifactor Dimensionality Reduction, Neural Networks and Support Vector Machine). The potential for those methods to detect gene-gene interaction have been identified in previous reviews (McKinney et al., 2006; Koo et al., 2013; Upstill-Goddard et al., 2013).

5.3.1. Tree-based methods

Classification and regression trees (CART)

Initially developed by Breiman et al. (1984), a classification tree consists in a set of nodes conferring a tree-like dichotomous structure. The building of the tree is decomposed into two main steps: a recursive partitioning step and a pruning step. The recursive partitioning of the tree consists in splitting each node into two offspring nodes according to the values of one variable. First, the top node, usually called the root of the tree, contains the entire training sample. The top node produces two child nodes defined to optimize the distribution homogeneity of the response variable, which is disease status Y in our context. Such a process is recursively repeated to each node to reach a terminal node containing a maximally homogeneous subsample. A popular splitting rule is to use the variable that maximizes the reduction in a quantity known as the Gini impurity at each node. By doing so it is hoped that the terminal nodes are pure, meaning that they only contain individuals with the same response. The second step consists in a bottom-up pruning process to remove some of the later splits or branches according to certain rules (Breiman et al., 1984) to avoid overfitting and to produce a final more parsimonious model.

To illustrate the use of CART in association studies, we used a publicly available dataset that contains the genotypes of more than 300 SNPs in a total of 429 patients (163 individuals affected by Rheumatoid Arthritis and 266 Health controls) (Chang et al., 2013; Emily et al., 2017). In Figure 4, the tree, obtained by restricting the analysis to 10 variables, show that after the pruning step only three variables have been used to build the splitting rules. From Figure 4, we can deduce that SNP `rs10184179`, `rs1006273` and `rs10400863` are interacting in susceptibility with the disease outcome.

Strengths. One of the main advantage of the CART method is its ability to deal with large scale dataset, thus making it applicable at the genome scale. Furthermore, CART is a model free method

 ISSN: 2102-6238

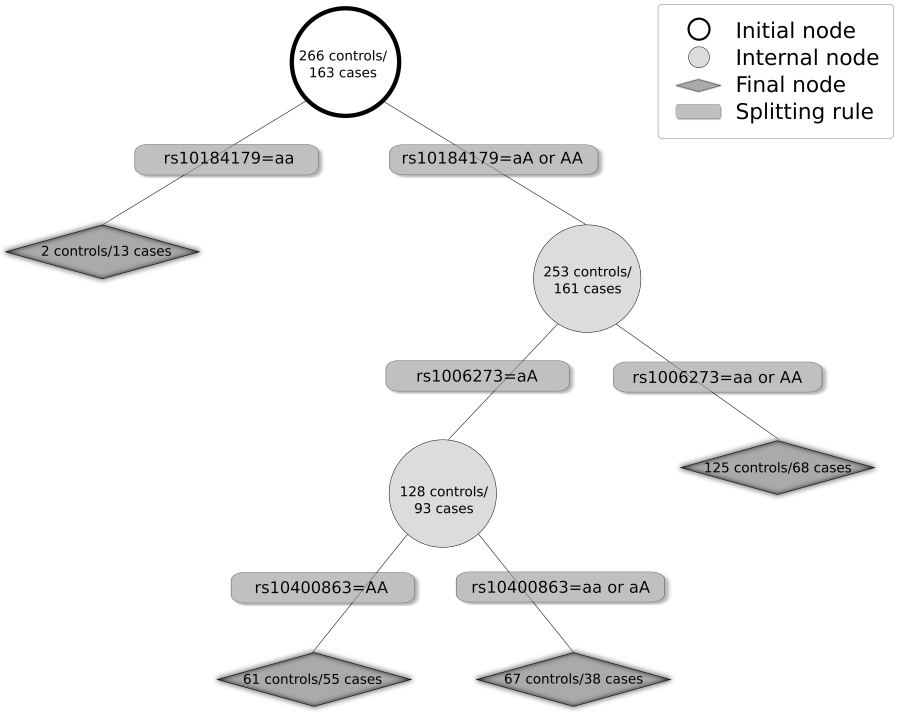

FIGURE 4. *Example of classification tree with 2 internal nodes and 4 final nodes. Splitting rule are based on the genotype of 3 SNPs* `rs10184179`*,* `rs1006273` *and* `rs10400863`*) that can be considered in interaction.*

so that any type of signal can potentially be detected. In particular, non-linear interaction are likely to be reported by CART.

Weaknesses. However, it is noteworthy that a classification tree do not include interaction variables *per se* in the model. Rather, the trees constructed allow for interaction in the sense that each path through a tree corresponds to a particular combination of values taken by certain predictor variables, thus including the potential interactions between them. Therefore, because it conditions on the main effects of variables at the first stage and on the main effects conditional on previously selected variables at subsequent stages, pure interactions in the absence of main effects can be missed (McKinney et al., 2009). By following paths in the trees, it is possible to identify causal pairs of variables. However, the pruning process has to be strong enough to cut edges so that low-level of interaction can be identified. To be computed, the CART required the tuning of some sensitive parameters such as the the measure of the splitting rule and the pruning method. Furthermore, the CART method is known to be unstable in the context of correlated data and to be sensitive to unbalanced design, thus making CART not robust to the structure of the studied dataset.

Random Forest

Rather than using a single tree, classification accuracy can be improved by growing an ensemble of trees. One of the most popular ensemble tree approach is the random forests approach introduced

 ISSN: 2102-6238

by Breiman (2001) and used in several genetic studies (for example Bureau et al. (2005)). Compared to the CART method, a random forest aims at generating a collection of n unpruned trees, where each unpruned tree is learnt on a bootstrap sample of the original individuals. Furthermore, at each node, rather than considering all possible predictor variables, only a random subset of the possible predictor variables is considered. This procedure thus results in a forest of trees, each of which will have been trained on a particular bootstrap sample of observations. The observations that were not used for growing a particular tree can be used as out-of-bag instances to estimate the prediction error. The out-of-bag observations can also be used to estimate variable importance in different ways including through use of a permutation procedure (McKinney et al., 2006, 2009).

Strengths. Similar to the CART method, random forest has good computational performances so that it can be performed at the genome scale.

Weaknesses. However, the efficiency of a random forest depends on a high number of features. First, two main parameters (the number of trees and how many splitting rules are applied to each node) have to fixed beforehand. Next, as for the CART method, random forest implicitly considers interactions, further work is required to separate main effects from interactions in random forests since variable importance measures reflect both main and interaction effects (Bureau et al., 2005; Winham et al., 2012). Unfortunately, previous work has shown that RF is not designed to explicitly test for SNP interactions (Winham et al., 2012). Therefore, modeling needs to be done carefully to detect interactions and new methodologies need to be designed to capture the pure interactions without main effects between SNPs when modeling with random forest. Very recently, a permutation based random forest approach has been however introduced to capture pure epistasis (Li et al., 2016).

5.3.2. Pattern recognition methods

Multifactor Dimensionality Reduction: MDR

In genetic epidemiology, a popular series of methods is based on the Multifactor Dimensionality Reduction (MDR) approach (Gola et al., 2016). The basis of the MDR method is an induction algorithm that converts two or more variables to a single variable. As firstly described by Ritchie et al. (2001), MDR aims at reducing the dimensionality of a set of categorical variables, X_i's, by pooling the sets of categories into high-risk and low-risk groups, thus reducing to a single binary variable. The identification of the two groups (high-risk and low risk) is performed according to the ratio of cases and controls in each class.

The original MDR procedure is presented in the algorithm 1. In more details, let consider $k \in [1,\ldots,K]$, the order of interaction set in the analysis, where K is the maximal interaction order ($K \leq p$). Let us also introduce $\ell_k \in \left[1,\ldots,\binom{k}{p}\right]$, where ℓ_k identifies the ℓ_k^{th} combination of k variables among the p predictors. For example, if $\ell_2 = 1$, the subset of $k = 2$ variables (X_1, X_2) is considered, if $\ell_3 = 3$, we considered the subset of 3 variables (X_1, X_2, X_5) or if $\ell_4 = \binom{4}{p}$ then $(X_{p-3}, X_{p-2}, X_{p-1},, X_p)$ is considered.

Finally, let us introduce $g_{k,\ell_k} : (x_1,\ldots,x_p) \to [1,\ldots,3^k]$ a function that maps the observed p genotypes into a set of values restricted to the k variables in the subset ℓ_k . For example,

 ISSN: 2102-6238

$g_{2,1}(0,0,x_3,\ldots,x_p) = 1\ (\forall x_3,\ldots,x_p)$, $g_{3,3}(0,0,x_3,x_4,2,x_6,\ldots,x_p) = 3\ (\forall x_3,x_4,x_6,\ldots,x_p)$, $g_{4,\binom{4}{p}} = (x_1,\ldots,x_{p-4},0,0,0,0) = 1\ (\forall x_1,\ldots,x_{p-4})$ or $g_{4,\binom{4}{p}}(x_1,\ldots,x_{p-4},2,2,2,2) = 3^4\ (\forall x_1,\ldots,x_{p-4})$. The main purpose is to find the optimal subset of k variables as follows:

$$(k^{opt},\ell_k^{opt}) = \arg\min_{(k,\ell_k)}\{err(k,\ell_k)\} \tag{49}$$

where

$$err(k,\ell_k) = \frac{1}{n.CV}\sum_{cv=1}^{n.CV} err(k,\ell_k,\mathcal{P}(cv)) \tag{50}$$

and

$$err(k,\ell_k,\mathcal{P}_{cv}) = \sum_{i\in\mathcal{I}} y_{\mathcal{P}_{cv}}(\widehat{i,k,\ell_k},\mathcal{I}) \neq y_i \tag{51}$$

where $y_{\mathcal{P}_{cv}}(\widehat{i,k,\ell_k},\mathcal{I}) = 1$ means that i is predicted to belong to the high-risk class. Individual i belong the high-risk class if the case-control ratio of individuals having the same k selected genotypes as i in the training set of individuals $\mathcal{I}$ is higher than a threshold τ. It can be formalized as follows where r is the case-control ratio function, estimated for each of the 3^k combination of k genotypes in the training set:

$$y(\widehat{i,k,\ell_k},\mathcal{I}) = 1 \text{ iif } r(g_{k,\ell_k}(x_{i,1},\ldots,x_{i,p}),k,\ell_k,\mathcal{I}) > \tau \tag{52}$$

$$r(g,k,\ell_k,\mathcal{I}) = \frac{\sum_{i\in\mathcal{I};y_i=1}\mathbb{I}_{\{g_{k,\ell_k}(x_{i,1},\ldots,x_{i,p})=g\}}}{\sum_{i\in\mathcal{I};y_i=0}\mathbb{I}_{\{g_{k,\ell_k}(x_{i,1},\ldots,x_{i,p})=g\}}} \tag{53}$$

Therefore, the final best model, the k_{opt}-variables combination model, is the model that minimizes the prediction error which estimated in the cross-validation procedure Ritchie et al. (2001). Therefore, MDR finds both the optimal interaction order k_{opt} and the corresponding k_{opt} factors that are significant in determining the disease status.

Strengths. The success of the MDR method in the genetic community is mainly due its computational efficiency that has allowed MDR to be applied at the genome scale. Furthermore, since MDR is model free, it has the ability to catch any type of interaction signal, such as non linear interaction and pure interaction signals. The flexibility of the method is further examplified by the vast amount of extensions and modifications of the original idea of MDR that has been proposed in the literature (see Gola et al. (2016) for a nice review of these evolution).

Weaknesses. Nevertheless, MDR-based methods suffer from main limitations. First, rather than testing for interaction, MDR seeks to identify combinations of variables that influence a disease outcome, possibly by interactions or by main effects. Such measure of heterogeneity might prevent from detecting pure epistasis and from identifying a true causal pair of variables. Furthermore, as illustrated by the numerous variations of the MDR method (Gola et al., 2016), the performances of the MDR depends on many parameters, such as K, τ, $n.CV$ in Algorithm **??**, that has to be set by the user. Such instability of MDR-based methods is further enhanced by their sensitivity to the design of dataset and especially to the case/control ratio. Finally, although resampling and cross-validation techniques are used throughout the MDR pipeline, MDR-based methods are known to suffer from overfitting.

 ISSN: 2102-6238

Algorithm 1 *MDR* algorithm

Require: nCV, $\{nCV$ is the number of cross-validations$\}$
Require: K $\{K$ is the maximal order of interaction$\}$
Require: τ $\{\tau$ is the threshold case-control ratio for determining high-risk group$\}$
  for $k \in [1,\ldots,K]$ **do**
    for $\ell_k \in \left[1,\ldots,\binom{k}{p}\right]$ **do**
      for $tCV \in [1,\ldots,nCV]$ **do**
        Split the whole set of individuals $\mathcal{I}$ in ten subsets $\mathcal{I}_s$ such as $\sharp\mathcal{I}_s = \sharp\mathcal{I}/10$ $\{\{I_1,\ldots,I_{10}\} = \mathcal{P}_{cv}$ in Equation (50)$\}$
        for $\mathcal{I}_{Test} \in \{\mathcal{I}_s\}$ **do**
          Set $\mathcal{I}_{Train} = \overline{\mathcal{I}_{Test}}$
          for $g \in \left[1,\ldots,3^k\right]$ **do**
            Compute $r(g,k,\ell_k,\mathcal{I}_{Train})$ as detailed in Equation (53)
          end for
        end for
        Compute the test error rate $err(k,\ell_k,\{I_1,\ldots,I_{10}\})$, based on the current partition $\mathcal{P}_{cv} = \{I_1,\ldots,I_{10}\}$, according to Equation (51)
      end for
      Compute the test error rate $err(k,\ell_k)$ as described in Equation (50)
    end for
  end for
  Select the optimal interaction order k^{opt} and the corresponding subset of variables ℓ_k^{opt} as defined in Equation 49
  return k^{opt}, ℓ_k^{opt}

Neural networks (NN)

NNs have been introduced to mimic the brain's ability to solve problems. An NN can be seen as an indirected graph composed of nodes that characterize the processing elements (or neurons) and directed edges that represent the connections of the nodes (or synaptic connections including the flow of information). Nodes are arranged in layers of three types: an input layer, a set of hidden (or internal) layers and an output layer (see Figure 5). Nodes from the input layer represents the p input variables $(X_1,\ldots,X_p)$ while, in the context of association studies, the output layer is composed of only one node characterizing the response variable Y. The K hidden layers are made by n_k $(k = 1\ldots K)$ nodes respectively, where K and n_k are the two main parameters of the architecture of the NN. Each hidden node, as well as the output node, can be represented as a weighted sum of its inputs as follows:

$$\forall j \in [1,n_1] : H_{1,j} = f\left(\sum_{i=1}^{p} v_{j,i}X_i\right)$$

$$\forall k \in [2,K], \forall j \in [1,n_k] : H_{k,j} = f\left(\sum_{i=1}^{n_{k-1}} w_{j,i}^{k-1} H_{k-1,i}\right)$$

$$Y = f\left(\sum_{i=1}^{n_K} z_i H_{K,i}\right)$$

where f is a nonlinear function and $v_{j,i}$, $w_{j,i}^{k-1}$, z_i are weights of connections between two nodes in two consecutive layers. More precisely, $v_{j,i}$ is the weight between X_i and $H_{1,j}$, $w_{j,i}^{k-1}$ the weight between $H_{k-1,i}$ and $H_{k,j}$ and z_i the weight between $H_{K,i}$ and Y. It is worth noting that f can be

 ISSN: 2102-6238

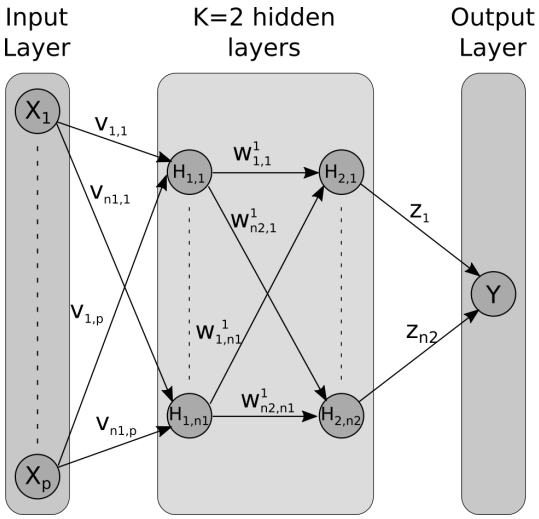

FIGURE 5. *Example of an architecture of a neural network with p variables in the input layer, $K = 2$ hidden layers with n_1 and n_2 nodes and an output layer with one node corresponding to a one-dimensional response variable.*

any nonlinear function, however, f is usually chosen to be sigmoid ($f(x) = 1/(1+e^{-x})$) or the Heaviside function ($f(x) = 1$ if $x > 0$ and 0 otherwise).

The architecture of neural networks is the key of success for detecting gene-gene interactions (Koo et al., 2013) and several strategies have been proposed to address this issue such as back propagation neural network (BPNN), genetic programming neural network (GPNN) and grammatical evolution neural networks (GENN). In BPNN, the optimization algorithm minimizes the error by changing the weights following each pass through the network. For that purpose, BPNN proposes small changes to the weights until it reaches a value to which any change makes the error higher, indicating the the error has been minimized (Skapura, 1995; Ritchie et al., 2003). GPNN was also proposed by Ritchie et al. (2003) to optimize the neural network. Genetic Programming was used to optimize the inputs from a larger pool of variables, the weights, and the connectivity of the network, including the number of hidden layers and the number of nodes in the hidden layer. Thus, the algorithm attempts to generate the appropriate network architecture for a given dataset. Finally, GENN is based on an evolutionary algorithm that uses genomes knowledge and grammars to define the observed populations Motsinger-Reif et al. (2008). Compared to GPNN, where only two connections between nodes are possible, in GENN the grammar allows for defining multiple connections between nodes selected by the algorithm. Variable numbers of connections allows for more complicated neural networks to be evolved and potentially makes GENN more powerful than GPNN.

Strengths. Neural networks are not based on a pre-specified model thus conferring to the method the flexibility to detect any type of signal (potentially pure epistatic signal). Since neural network is largely studied in the computer science community, its computational performances allow to handle with large scale data sets. However, neural network does not yet scale up to the genome level.

Weaknesses. One of the main drawback of neural network is the need to specify the neural architecture *a priori*. Therefore, neural network might fail at catching interaction because of a miss-specification of the architecture. Furthermore neural network is known to suffer from overfitting and therefore depends on the structure of the data. Finally, one of the major criticisms in association studies is their being black boxes, since no satisfactory explanation of their behavior has been offered. Output models are difficult in the interpretation and always need for comprehensive validation, thus preventing neural network to clearly identify causal pairs.

Support Vector Machine: SVM

Support Vector Machine or SVM is one of the most popular methods among machine learning algorithm to perform classification or regression. Considering the context of association studies with p markers $(X_1, \ldots, X_p)$, SVM basically aims at building a $p-1$ dimensional hyperplane (or set of hyperplanes in a high-dimensional space), which can be used to separate each individual (seen as a p-dimensional point) with response to response variable Y. A good separation between cases ($Y = 1$) and controls ($Y = 0$) is achieved by the hyperplane that has the largest distance to the nearest training-data point of any class (so-called functional margin) (Cortes and Vapnik, 1995). The application of SVM technique in the detection of gene-gene interaction has been proposed in Chen et al. (2008). The authors combined SVM with combinatorial optimization methods (such as genetic algorithm) and showed the good performance of SVM-based methods when dealing with unbalanced data. In Shen et al. (2010), it has been proposed to first use a L1 penalty to identify the most promising predicting interaction.

When the original data are not linearly separable, the X_i are mapped into higher dimensional space, also called feature space, with the hope of a linear separation of the data in that space. Mappings are designed to ensure that dot products may be computed easily by defining them in terms of a kernel function $K(x_i, x_j)$, which refers to the kernel trick. In Missiuro (2010), the authors used kernel SVM to predict which genes are genetically interacting with each other.

Strengths. SVM has been designed to avoid the specification of a model and therefore is a model free method and can detect complex interactions between variables.

Weaknesses. However, SVM has not been designed to identify variables associated with a response but rather to propose classifiers that combine variable. Therefore, SVM are designed neither to detect pure epistasis nor to identify causal pairs. Furthermore, its efficiency is based on the choice of several parameters, such as an appropriate kernel. Such a parameter dependency is one main reason why SVM is known to suffer from overfitting. Finally, the data transformation by means of a kernel function, can be costly in time of computation, thus preventing the use of SVM at the genome scale.

6. Statistical perspectives in gene-gene interaction

The investigation of gene-gene interaction has a long history in plant, animal and human genetics (Cordell, 2009) and has raised many statistical issues, thus allowing for significant statistical advances. In the context of GWAS, although a huge number of methods have been proposed and although the proposed methods cover the whole landscape of statistics, reported interaction

 ISSN: 2102-6238

TABLE 7. *Summary of the six features, related to the statistical power, of statistical methods at the genome scale (Capacity to detect pure interaction, capacity to identify the causal pair, the genome-wide scalability, robustness to a model specification, robustness to parameter settings and robustness to data structure). Each feature is evaluated according to five scales of ability: ++ for a very good ability, + for a good ability, ∘ for an average ability, - for a poor ability and – for a very poor ability.*

	Detection of pure epistasis	Identification causal pairs	Genome -wide scalability	Model free	Parameters free	Robustness to data structure
Exhaustive search						
SNP-SNP	+ +	+ +	- -	- -	+ +	+
Gene-Gene	+ +	+	- -	- -	+ +	+ +
Regression-based methods						
Logic regression	+ +	+ +	∘	∘	-	+
Penalized logistic regression	+ +	+	- -	-	∘	∘
Machine learning methods						
Tree-based methods						
CART	- -	∘	+ +	+	∘	-
Random Forest	-	-	+	-	-	-
Pattern recognition methods						
MDR	∘	-	+	+ +	- -	- -
NN	∘	- -	∘	+	-	- -
SVM	- -	-	∘	+ +	-	- -

remains very rare. However, genetic interaction is still considered to play a major role for tackling complex human disease genetics (Mackay and Moore, 2014). Furthermore, in the era of personalized medicine, there is growing evidence that detecting interaction is crucial to improve our understanding of major complex diseases.

The very low number of reported gene-gene interaction can therefore be explained by a lack of statistical power and biological interpretation. More efforts should be put in addressing several statistical issues. Among these issues, tackling the questions of computational burden alleviation, multiple testing correction and visualisation interpretation is central in future improvements.

6.1. Computational burden

One of the main limitations in performing genome-wide gene-gene interaction is the computational burden encountered in all strategies (exhaustive search, penalized regression or machine learning methods). To overcome such limitation, the incorporation of prior biological knowledge has been rapidly proposed (Emily et al., 2009). Many tools have been developed to incorporate biological knowledge in the analysis such as protein-protein interaction approaches, pathway approaches or comprehensive knowledge approaches (Ritchie, 2015). Although biological knowledge-based methods are seen as powerful strategies, they raise several issues. First, the uncertainty in biological knowledge is (almost) never accounting for when reducing the search space. However, ignoring uncertainty may result in a lack of power or in an uncontrolled type-I error rate, thus inflating the amount of false positive. Next, filtering the search space by conditioning on some biological prior generates threshold models from which genome-wide interpretation is not feasible. However, in the era of big data, we can hope that improvements in data management, data storage,

 ISSN: 2102-6238

computer performances and distributed computing could help reducing the computational cost.

6.2. *Multiple comparison issue*

Another important limitation of gene-gene interaction search in GWAS is the statistical confidence, related to the issue of multiple testing correction, in the obtained results. Multiple testing is a commonly encountered inference problem in large-scale genetic association studies as a result of simultaneous testing of multiple hypotheses. In the exhaustive pairwise testing approach, the number of tests performed, given by $L(L-1)/2$ where L can be either the number of SNPs or the number of genes, is expected to be very high. Furthermore, tests cannot be assumed to be independent and correlation between them arise from two main sources: (1) the pair (X_i, X_j) is obviously correlated with the pair (X_i, X_k) due to the common variable X_i and (2) because of the genome correlation structure if X_i and X_j are correlated (as being in Linkage Disequilibrium for example), each pair (X_i, X_k) is correlated with each pair (X_j, X_ℓ). Large numbers of tests and complex correlated tests lead to complicated error control in interaction analyses (Musani et al., 2007). Current solutions for controlling the family-wise error rate (FWER) or the false discovery rate (FDR) in gene-gene interaction at the genome scale are mainly based on the estimation of the effective number of tests (Hendricks et al., 2014). Such a strategy is limited by the lack of a proper definition of the effective number of tests and efforts from the statistical community should provide a better understanding of its behaviour in various correlation contexts.

To account for the complex correlation in the data, resampling strategies are commonly used to assess a significant level. This has been performed in single-testing, where the genome-wide significant threshold for p-value has been estimated to 5×10^{-8} using extensive permutations (Jannot et al., 2015). However, this can hardly be done for interaction because of the computational burden relative to the very low threshold to be reached. Furthermore, assessing very low significance levels with permutations is likely to suffer from a sample overfitting.

6.3. *Visualization and lack of interpretation*

A final major issue in gene-gene interaction in GWAS is the lack of interpretation of the results. The complexity of the search space as well as the complexity of interaction models make the interpretation of the results challenging. Furthermore, compared to single-testing approaches where results can easily be visualized with Manahattan plots and QQ plots (Clarke et al., 2011), visualisation is much more complex when dealing with interaction. Very few tools have been developed to address this issue and main focuses are made on heatmap and circos representation (Wu et al., 2013; Emily et al., 2017). However, thanks to the emergence of a new generation of data science, visualization is a rapidly growing area of research and it is certain that gene-gene interaction would benefit from it.

Another important aspect in the interpretation of the results is its reproducibility. Most the methods presented in this paper have been implemented in home made softwares that are for most of them available only on request to the authors and at best have a web interface. Thus, searching for gene-gene interaction is not straightforward. Furthermore, a comprehensive comparison of such methods, in terms of power and computational performances, remains hardly feasible. As

 ISSN: 2102-6238

proposed in (Emily et al., 2017), it is important to develop generic computational and statistical tools.

6.4. Concluding words

Although high-throughput technologies evolve rapidly with the emergence of next-generation sequencing, such as RNA and single cell sequencing, GWAS are still widely used. Compared to more recent technologies, GWAS data are indeed less expensive and less noisy, thus having the potential to reliably test for gene-gene interaction. More than computational issues that are likely to be solved in a near future, statistical issues remain challenging. However, improvements in the statistical community provide hope for a better detection of gene-gene interaction. First, the recent advances in the multiple testing community to account for correlation in the correction gives promises in improving the statistical procedures as well as proposing reliable interpretation. Furthermore, the rapid evolution of machine learning techniques open the way to the development of new methods able to embrace the complexity gene-gene interaction (Mackay and Moore, 2014; S. Uppu, 2016).

References

Balding, D. (2006). A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7:781–791.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Ann. Statist.*, 41(3):1111–1141.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Bureau, A., Dupuis, J., Falls, K., Lunetta, K. L., Hayward, B., Keith, T. P., and Van Eerdewegh, P. (2005). Identifying snps predictive of phenotype using random forests. *Genetic Epidemiology*, 28(2):171–182.

Chang, X., Xu, B., Wang, L., Wang, Y., Wang, Y., and Yan, S. (2013). Investigating a pathogenic role for txndc5 in tumors. *International Journal of Oncology*, 43(43):1871–1884.

Chattopadhyay, A. S., Hsiao, C.-L., Chang, C. C., Lian, I.-B., and Fann, C. S. (2014). Summarizing techniques that combine three non-parametric scores to detect disease-associated 2-way snp-snp interactions. *Gene*, 533(1):304 – 312.

Chen, L., Yu, G., Langefeld, C. D., Miller, D. J., Guy, R. T., Raghuram, J., Yuan, X., Herrington, D. M., and Wang, Y. (2011). Comparative analysis of methods for detecting interacting loci. *BMC Genomics*, 12(1):344.

Chen, S.-H., Sun, J., Dimitrov, L., Turner, A. R., Adams, T. S., Meyers, D. A., Chang, B.-L., Zheng, S. L., Grönberg, H., Xu, J., and Hsu, F.-C. (2008). A support vector machine approach for detecting gene-gene interaction. *Genetic Epidemiology*, 32(2):152–167.

Cheverud, J. M. (2001). A simple correction for multiple comparisons in interval mapping genome scans. *Heredity*, 87(1):52–58.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., and Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 6(2):121–133.

Conneely, K. N. and Boehnke, M. (2007). So many correlated tests, so little time! rapid adjustment of p values for multiple correlated tests. *The American Journal of Human Genetics*, 81(6):1158–1168.

Cordell, H. J. (2002). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468.

Cordell, H. J. (2009). Detecting gene-gene interactions that underlie human diseases. *Nature Review Genetics*, 10(2):392–404.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience.

 ISSN: 2102-6238

Dong, C., Chu, X., Wang, Y., Wang, Y., Jin, L., Shi, T., Huang, W., and Li, Y. (2008). Exploration of gene-gene interaction effects using entropy-based methods. *European Journal of Human Genetics*, 16(2):229–235.

Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994.

Emily, M. (2012). Indor: a new statistical procedure to test for snpxsnp epistasis in genome-wide association studies. *Statistics in Medicine*, 31(21):2359–2373.

Emily, M. (2016a). Aggregator: A gene-based gene-gene interacttion test for case-control association studies. *Statistical Application in Genetics and Molecular Biology*, 15(2):151–171.

Emily, M. (2016b). Power comparison of cochran-armitage test of trend against allelic and genotypic tests in case-control genetic association studies. *Statistical Methods in Medical Research*, page In press.

Emily, M. and Friguet, C. (2015). Power evaluation of asymptotic tests for comparing two binomial proportions to detect direct and indirect association in large-scale studies. *Statistical Methods in Medical Research*, page In press.

Emily, M., Mailund, T., Hein, J., Schauser, L., and Schierup, M. H. (2009). Using biological networks to search for interacting loci in genome-wide association studies. *European Journal of Human Genetics*, 17:1231–1240.

Emily, M., Sounac, N., Kroell, F., and Houée-Bigot, M. (2017). *GeneGeneInteR: Tools for Testing Gene-Gene Interaction at the Gene Level*. R package version 1.00.1.

Fan, R., Zhong, M., Wang, S., Zhang, Y., Andrew, A., Karagas, M., Chen, H., Amos, C., Xiong, M., and Moore, J. (2011). Entropy-based information gain approaches to detect and to characterize gene-gene and gene-environment interactions/correlations of complex diseases. *Genetic Epidemiology*, 35(7):706–721.

Ferrario, P. G. and Konig, I. R. (2016). Transferring entropy to the realm of gxg interactions. *Briefings in Bioinformatics*, pages 1–12.

Galwey, N. W. (2009). A new measure of the effective number of tests, a practical tool for comparing families of non-independent significance tests. *Genetic Epidemiology*, 33(7):559–568.

Gao, H., Wu, Y., Li, J., Li, H., Li, J., and Yang, R. (2014). Forward lasso analysis for high-order interactions in genome-wide association study. *Briefings in Bioinformatics*, 15(4):552.

Genz, A. and Bretz, F. (2009). *Computation of Multivariate Normal and T Probabilities*. Springer-Verlag, 1st edition.

Gola, D., Mahachie John, J. M., van Steen, K., and KÃ¶nig, I. R. (2016). A roadmap to multifactor dimensionality reduction methods. *Briefings in Bioinformatics*, 17(2):293.

Hall, P. and Jin, J. (2010). Innovated higher criticism for detecting sparse signals in correlated noise. *Ann. Statist.*, 38(3):1686–1732.

Hallgrimsdottir, I. B. and Yuster, D. S. (2008). A complete classification of epistatic two-locus models. *BMC Genetics*, 9(17).

Hendricks, A. E., Dupuis, J., Logue, M. W., Myers, R. H., and Lunetta, K. L. (2014). Correction for multiple testing in a gene region. *European Journal of Human Genetics*, 22(3):414–418.

Hill, W. G. and Robertson, A. (1968). Linkage diseqilibrium in finite populations. *Theoretical and Applied Genetics*, 38:226–231.

Hindorff, L. A., Sethupathy, P., Junkins, H. A., Ramos, E. M., Mehta, J. P., Collins, F. S., and A., T. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceeding of the National Academy of Sciences*, 106(23):9362–9367.

Huang, H., Chanda, P., Alonso, A., Bader, J. S., and Arking, D. E. (2011). Gene-based tests of association. *PLoS Genetics*, 7(7):e1002177.

Jannot, A., Ehret, G., and Perneger, T. (2015). $p < 5 \times 10^{-8}$has emerged as a standard of statistical significance for genome-wide association studies. *Journal of Clinical Epidemiology*, 68(4):460–465.

Jiang, B., Zhang, X., Zuo, Y., and Kang, G. (2011). A powerful truncated tail strength method for testing multiple null hypotheses in one dataset. *Journal of Theoretical Biology*, 277(1):67 – 73.

Jorgenson, E. and Witte, J. S. (2006). A gene-centric approach to genome-wide association studies. *Nature Review Genetics*, 7(11):885–891.

Koo, C. L., Liew, M. J., Mohamad, M. S., and Salleh, A. H. M. (2013). A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. *BioMed Research International*, page 13.

Kooperberg, C. and Ruczinski, I. (2005). Identifying interacting snps using monte carlo logic regression. *Genetic Epidemiology*, 28:157–170.

Kwon, M.-S., Park, M., and Park, T. (2014). Igent: efficient entropy based algorithm for genome-wide gene-gene interaction analysis. *BMC Medical Genomics*, 7(1):S6.

Larson, N. B., Jenkins, G. D., Larson, M. C., Vierkant, R. A., Sellers, T. A., Phelan, C. M., Schildkraut, J. M., Sutphen,

 ISSN: 2102-6238

R., Pharoah, P. P. D., Gayther, S. A., Wentzensen, N., Goode, E. L., and Fridley, B. L. (2014). Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer. *European Journal of Human Genetics*, 22(1):126–131.

Lewis, C. M. (2002). Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, 3(2):146–153.

Li, J., Huang, D., Guo, M., Liu, X., Wang, C., Teng, Z., Zhang, R., Jiang, Y., Lv, H., and Wang, L. (2015). A gene-based information gain method for detecting gene-gene interactions in case-control studies. *European Journal of Human Genetics*, Online:Online.

Li, J. and Ji, L. (2005). Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95:221–227.

Li, J., Malley, J. D., Andrew, A. S., Karagas, M. R., and Moore, J. H. (2016). Detecting gene-gene interactions using a permutation-based random forest method. *BioData Mining*, 9(1):14.

Li, J., Tang, R., Biernacka, J., and de Andrade, M. (2009). Identification of gene-gene interaction using principal components. *BMC Proceedings*, 3(Suppl 7):S78.

Li, M.-X., Gui, H.-S., Kwan, J., and Sham, P. (2011). Gates: A rapid and powerful gene-based association test using extended simes procedure. *The American Journal of Human Genetics*, 88(3):283–293.

Li, W. and Reich, J. (2000). A complete enumeration and classification of two-locus disease models. *Human Heredity*, 50(6):334–349.

Liu, J. Z., Mcrae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., Hayward, N. K., Montgomery, G. W., Visscher, P. M., Martin, N. G., and Macgregor, S. (2010). A versatile gene-based test for genome-wide association studies. *The American Journal of Human Genetics*, 87(1):139 – 145.

Mackay, T. and Moore, J. (2014). Why epistasis is important for tackling complex human disease genetics. *Genome Medicine*, 6(42).

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

Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F. C., McCarroll, S. A., and Visscher, P. M. (2009). Finding the missing heritability of complex diseases. *Nature*, 461:747–753.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall, London, UK.

McKinney, B. A., Crowe, Jr, J. E., Guo, J., and Tian, D. (2009). Capturing the spectrum of interaction effects in genetic association studies by simulated evaporative cooling network analysis. *PLOS Genetics*, 5(3):1–12.

McKinney, B. A., Reif, D. A., Ritchie, M. D., and Moore, J. H. (2006). Machine learning for detecting gene-gene interactions. *Appl. Bioinformatics*, 5(2):77–88.

Meier, L., Geer, S. V. D., Bühlmann, P., and Zürich, E. T. H. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B*.

Mielniczuk, J. and Rdzanowski, M. (2017). Use of information measures and their approximations to detect predictive gene-gene interaction. *Entropy*, 19(1).

Missiuro, P. V. (2010). *Predicting genetic interactions in Caenorhabditiselegans using machine learning*. PhD thesis, Massachusetts Institue of Technology.

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*, 56:73–82.

Moskvina, V. and Schmidt, K. M. (2008). On multiple-testing correction in genome-wide association studies. *Genetic Epidemiology*, 32(6):567–573.

Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W., and Ritchie, M. D. (2008). Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology*, 32(4):325–340.

Musani, S. K., Shriner, D., Liu, N., Feng, R., Coffey, C. S., Yi, N., Tiwari, H. K., and Allison, D. B. (2007). Detection of gene x gene interactions in genome-wide association studies of human population data. *Human Heredity*, 63:67–84.

Neale, B. M. and Sham, P. C. (2004). The future of association studies: Gene-based analysis and replication. *The American Journal of Human Genetics*, 75(3):353–362.

Niel, C., Sinoquet, C., Dina, C., and Rocheleau, G. (2015). A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, 6:285.

Nyholt, D. R. (2004). A simple correction for multiple testing for single-nucleotide polymorphisms in linkage

 ISSN: 2102-6238

disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765 – 769.
Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics*, 9:30–50.
Peng, Q., Zhao, J., and Xue, F. (2010). A gene-based method for detecting gene-gene co-association in a case-control association study. *European Journal of Human Genetics*, 18(5):582–587.
Phillips, P. (2008). Epistasis, the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Review Genetics*, 9:855–867.
Prabhu, S. and Pe'er, I. (2012). Ultrafast genome-wide scan for snp-snp interactions in common complex disease. *Genome Research*, 22(11):2230–2240.
Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., and Sham, P. C. (2007). Plink: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics*, 81:559–575.
Rajapakse, I., Perlman, M. D., Martin, P. J., Hansen, J. A., and Kooperberg, C. (2012). Multivariate detection of gene-gene interactions. *Genetic Epidemiology*, 36(6):622–630.
Ritchie, M. D. (2015). *Finding the Epistasis Needles in the Genome-Wide Haystack*, pages 19–33. Springer New York, New York, NY.
Ritchie, M. D., Hahn, L. W., Roodi, N., Bailey, L. R., Dupont, W. D., Parl, F. F., and Moore, J. H. (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *American Journal of Juman Genetics*, 69(1):138–147.
Ritchie, M. D., White, B. C., Parker, J. S., Hahn, L. W., and Moore, J. H. (2003). Optimization of neural network architecture using genetic programming improves detection and modeling of gene-gene interactions in studies of human diseases. *BMC Bioinformatics*, 4(1):28.
S. Uppu, A. Krishna, R. P. G. (2016). A deep learning approach to detect snp interactions. *Journal of software*, 11(10):965–975.
Schwender, H. and Ickstadt, K. (2008). Identification of snp interactions using logic regression. *Biostatistics*, 9(1):187.
Schwender, H., Ruczinski, I., and Ickstadt, K. (2011). Testing snps and sets of snps for importance in association studies. *Biostatistics*, 12(1):18.
Shang, J., Zhang, J., Sun, Y., Liu, D., Ye, D., and Yin, Y. (2011). Performance analysis of novel methods for detecting epistasis. *BMC Bioinformatics*, 12(1):475.
Shannon, C. E. (2001). A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55.
Shen, Y., Liu, Z., and Ott, J. (2010). Detecting gene-gene interactions using support vector machines with l1 penalty. In *2010 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW)*, pages 309–311.
Skapura, D. M. (1995). *Building Neural Networks*. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA.
Stanislas, V., Dalmasso, C., and Ambroise, C. (2017). Eigen-epistasis for detecting gene-gene interactions. *BMC Bioinformatics*, 18(1):54.
Steen, K. V. (2011). Travelling the world of gene-gene interactions. *Briefings in Bioinformatics*.
Su, L., Liu, G., Wang, H., Tian, Y., Zhou, Z., Han, L., and Yan, L. (2015). Research on single nucleotide polymorphisms interaction detection from network perspective. *PLOS ONE*, 10(3):1–19.
Thomas, D. C. (2004). *Statistical Methods in Genetic Epidemiology*. Oxford University Press, New York, first edition.
Ueki, M. and Cordell, H. J. (2012). Improved statistics for genome-wide interaction analysis. *PLoS Genet*, 8(4):e1002625.
Upstill-Goddard, R., Eccles, D., Fliege, J., and Collins, A. (2013). Machine learning approaches for the discovery of gene x gene interactions in disease data. *Briefings in Bioinformatics*, 14(2):251.
Wan, X., Yang, C., Yang, Q., Xue, H., Fan, X., Tang, N. L. S., and Yu, W. (2010). Boost: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *The American Journal of Human Genetics*, 87:325–340.
Winham, S. J., Colby, C. L., Freimuth, R. R., Wang, X., de Andrade, M., Huebner, M., and Biernacka, J. M. (2012). Snp interaction detection with random forests in high-dimensional genetic data. *BMC Bioinformatics*, 13(1):164.
Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., and Lin, X. (2010a). Powerful snp-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics*, 86(6):929–942.
Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics*, 25(6):714.
Wu, X., Dong, H., Luo, L., Zhu, Y., Peng, G., Reveille, J. D., and Xiong, M. (2010b). A novel statistic for genome-wide interaction analysis. *PLoS Genetics*, 6(9):e1001131.
Wu, Y., Zhu, X., Chen, J., and Zhang, X. (2013). Einvis: A visualization tool for analyzing and exploring genetic

 ISSN: 2102-6238

interactions in large-scale association studies. *Genetic Epidemiology*, 37(7):675–685.

Yang, C., Wan, X., Yang, Q., Xue, H., and Yu, W. (2010). Identifying main effects and epistatic interactions from large-scale snp data via adaptive group lasso. *BMC Bioinformatics*, 11(1):S18.

Yee, J., Kwon, M.-S., Park, T., and Park, M. (2013). A modified entropy-based approach for identifying gene-gene interactions in case-control study. *PLOS ONE*, 8(7):1–8.

Yuan, Z., Gao, Q., He, Y., Zhang, X., Li, F., Zhao, J., and Xue, F. (2012). Detection for gene-gene co-association via kernel canonical correlation analysis. *BMC Genetics*, 13(1):83.

Yung, L. S., Yang, C., Wan, X., and Yu, W. (2011). Gboost: A gpu-based tool for detecting gene-gene interactions in genome-wide case control studies. *Bioinformatics*.

Zaykin, D., Zhivotovsky, L. A., Westfall, P., and Weir, B. (2002). Truncated product method for combining p-values. *Genetic Epidemiology*, 22(2):170–185.

Zhang, X., Yang, X., Yuan, Z., Liu, Y., Li, F., Peng, B., Zhu, D., Zhao, J., and Xue, F. (2013). A plspm-based test statistic for detecting gene-gene co-association in genome-wide association study with case-control design. *PLoS ONE*, 8(4):e62129.

Zhao, J., Jin, L., and Xiong, M. (2006). Test for interaction between two unlinked loci. *The American Journal of Human Genetics*, 79(5):831–845.

Zhu, R., Zhao, H., and Ma, S. (2014). Identifying gene-environment and gene-gene interactions using a progressive penalization approach. *Genetic Epidemiology*, 38(4):353–368.

 ISSN: 2102-6238