Omnibus testing approach for gene-based gene-gene interaction

Florian Hébert, David Causeur, Mathieu Emily

Omnibus testing approach for gene-based gene-gene interaction

Florian Hébert | **David Causeur** | **Mathieu Emily**

Department of Statistics and Computer Science, Institut Agro, CNRS, IRMAR, Univ Rennes, F-35000, Rennes, France

Genetic interaction is considered as one of the main heritable component of complex traits. With the emergence of genome-wide association studies (GWAS), a collection of statistical methods dedicated to the identification of interaction at the SNP level have been proposed. More recently, gene-based gene-gene interaction testing has emerged as an attractive alternative as they confer advantage in both statistical power and biological interpretation. Most of the gene-based interaction methods rely on a multidimensional modeling of the interaction, thus facing a lack of robustness against the huge space of inter-action patterns. In this paper, we study a global testing approaches to address the issue of gene-based gene-gene interaction. Based on a logistic regression modeling framework, all SNP-SNP interaction tests are combined to produce a gene-level test for interaction. We propose an omnibus test that takes advantage of (1) the heterogeneity between existing global tests and (2) the complemen-tarity between allele-based and genotype-based coding of SNPs. Through an extensive simulation study, it is demonstrated that the proposed omnibus test has the ability to detect with high power the most common interaction genetic models with one causal pair as well as more complex genetic models where more than one causal pair is involved. On the other hand, the flexibility of the pro-posed approach is shown to be robust and improves power compared to single global tests in replication studies. Furthermore, the application of our procedure to real datasets confirms the adaptability of our approach to replicate various gene-gene interactions.

KEYWORDS
correlated statistics, gene-gene interaction, genome-wide association studies, omnibus, replication studies, welcome trust case control consortium

1 | INTRODUCTION

Genome-wide association studies (GWAS) aim at detecting the genetic variants associated with complex human diseases and traits. For over a decade, GWAS have led to the identification of hundreds of loci involved in the etiology of thousands of diseases, thus providing valuable insights into their genetic architecture.[1] However, the enthusiasm generated by the success of GWAS has rapidly declined since single marker strategy fails at covering a large proportion of the genetic heritability for common complex diseases.[2,3] Nevertheless, it is widely assumed that genetic interactions are likely to play

a major role in phenotype-genotype relationships.[4,5] Since human complex diseases are generally caused by the combined effect of multiple genes, detecting genetic interactions is thus essential to address the issue of the "missing" heritability.[6,7]

Genetic interactions have first been investigated at the SNP level with the development of a large collection of statistical methods to detect SNP-SNP interactions[8-21] that have mainly been compared with respect to their computational and statistical performance.[22,23] However due to the huge amount of tests to be performed, SNP-level methods suffer mainly from computational and statistical burdens. Moreover, findings are hardly interpretable since SNP-SNP interaction can be difficult to translate into functional interaction, such as protein-protein interaction. To circumvent these issues, it has been proposed to investigate genetic interaction at the gene level.[5,24,25] Gene-based gene-gene tests allow for all the SNPs within the region of a gene to be jointly modeled as a set, thus leading to a potential gain in power by accounting for LD structures within genes[26] and by aggregating signals across pairs of variants in a gene.[27]

The issue of gene-gene interaction testing has drawn specific attention in genetic epidemiology where the phenotype corresponds to healthy/diseased status of sampled individuals. In this context, multidimensional methods have first been introduced in several studies to compare the SNP correlation structures between cases and controls populations. First, in Reference 28, the authors proposed to test the interaction between the two SNP-sets by comparing their respective decomposition in principal components. Next, a series of U-like statistics have been proposed to compare gene-gene interaction in cases and controls. Under such framework, several measures of interaction have been used such as the canonical correlation coefficient[29] along with kernelized versions[30,31] or a measure based on coefficients of partial least squares path modeling.[32] Other kernel-based methods have also been proposed in more standard variance component testing frameworks.[33,34] Rather than focusing on a single measure of correlation between genes, it has been proposed to test the association between the phenotype and the interaction between two SNP-sets by comparing the covariance structures in cases and controls.[35] Entropy-based methods have also been developed as an attractive option to detect nonlinear relationship between two genes.[36]

Unlike previously introduced methods, where association is tested based on a multidimensional modeling of the overall SNP-sets, it has been successfully proposed to combine single SNP-SNP interaction tests at the gene level.[37] Testing for gene-gene interaction can indeed be performed by applying SNP-SNP interaction tests to all possible SNP pairs between two SNP-sets. A single *P*-value at the SNP-set level can be obtained by aggregating the*P*-values obtained at the SNP level. It is noteworthy that such approach turns out to be very similar to the issue of signal detection in functional data analysis[38] and falls into the paradigm of global testing introduced in Reference 39. In Reference 37, the method Aggregator has been introduced as a maxT procedure where the aggregate statistic is the maximum of the absolute values of the SNP-SNP single statistics. The significance of the maxT statistic is obtained assuming a multivariate gaussian distribution, thus accounting for the correlation between pointwise SNP-SNP tests. Results obtained with Aggregator have demonstrated the potential of global testing approach in the context of gene-based gene-gene interaction testing,[37] by showing a robustness to the presence of main effects and a high power of detecting signals when few SNP pairs are involved in the genotype/phenotype association.

However, Aggregator suffers from main limitations that prevent approaches combining SNP-SNP interaction tests to be efficient in a wider range of situations. First, Aggregator is based on an allele-based modeling of SNPs where the genotype of a SNP corresponds to the number of the allele. Although such modeling is adapted to the detection of allele-based signals, especially when the genotype/phenotype relationship is linear in the logit scale, it prevents Aggregator from having power in detecting more complex interaction signals. Another limitation of Aggregator is the heuristic approach used to estimate the covariance matrix of pointwise statistics. As quoted in Reference 37, such heuristic prevents Aggregator from (1) correctly controlling for type I error rate in presence of correlation between genes and (2) adjusting the genotype/phenotype relationship for covariates.

To overcome the limitations of Aggregator, we propose a global framework based on a logistic regression modeling where the issue of global testing is addressed by considering a vector of pairwise statistics to summarize the association between a phenotype and a pair of SNP-sets. Our framework comprises two SNP genotype models for bi-allelic SNP with three possible genotypes. In allele-based modeling, the SNP is coded additively with the number of copies of one of the alleles. In genotype-based modeling, the SNP is coded as categorical using two indicator variables. Moreover, covariates can be straightforwardly included in our association models thus allowing for the adjustment to confounding factors.

In global testing, pointwise statistics are ought to be aggregated and various statistics have been proposed in response to this issue.[39] The most popular aggregating methods are the maximum of the absolute values, also called minP,[40] the squared of the L^2-norm,[41] the Higher Criticism (HC),[42] and the Hotelling's t^2 statistic.[43] However, these methods are known to have different statistical power with respect to the interplay between the correlation structure of pointwise

statistics and the pattern of association.[44] By considering different ways of accounting for the correlation between pointwise statistics, these methods are indeed complementary.

Therefore, we propose in our study two omnibus tests that efficiently combine the eight global tests, namely, minP, HC, L^2-norm, and Hotelling for both allele-based and genotype-based coding of SNPs. The aggregation of multiple *P*-values, obtained by the eight global tests, can be performed using a recently proposed Cauchy-combination method.[45] The Cauchy combination test, introduced to overcome the challenge of accounting for features like correlation and sparsity often encountered when aggregating multiple *P*-values, is based on a weighted sum of Cauchy transformation of individual *P*-values. Since the Cauchy combination test does not explicitly account for the correlation between individual methods, we also introduce an other omnibus test based on a resampling procedure and that we called "omnibus by resampling," to aggregate *P*-values in our omnibus strategy.

The implementation of the two omnibus tests, omnibus by Cauchy combination and omnibus by resampling, requires the computation of *P*-values for minP, HC, L^2-norm, and Hotelling. First, to avoid the calculation of the probability distribution of a multivariate normal random variable, known to be unstable,[40] we propose to assess the significance of these individual global tests using permutations. Permutations are feasible in replication studies in which significance criteria for type 1 error control are substantially less stringent than required for high-dimension genome-wide gene-interaction discovery studies. To control the potential main effect of each SNP-set as well as the confounding effect of covariates, we use a parametric bootstrap approach as proposed in Reference 46.

To compute a *P*-value for Hotelling's t^2 statistic, the estimation of the inverse of the correlation matrix of the vector of interaction coefficients is needed. We propose an estimator based on a Kronecker decomposition of the correlation matrix and prove that our estimator converges in probability to the true correlation matrix. Our estimator has the main advantage of only computing the correlation matrices for both marginal SNP-sets, thus reducing the computational cost while improving the stability of the estimation.

In Section 2, the overall statistical framework is detailed by focusing on the various types of coding considered for SNPs and on methods for combining pointwise statistics. We also introduced our estimator for the correlation matrix as well as the sampling procedure used to test for the significance of global test statistics. Section 3 is devoted to our proposed omnibus strategy. After illustrating our motivation for combining global tests, we detail how our omnibus by resampling test is defined. Section 4 presents the main results obtained for demonstrating the correct control of the Type-I error rate by our procedure and for comparing the performance in terms of power of detection of our omnibus by resampling test compared to other tests. Our power study is based on both data-driven simulations, where a large number of disease models have been considered, and on truly observed data, where 25 pairs of SNP-sets in susceptibility with five complex diseases have been tested. The paper ends with a discussion in Section 5.

2 | STATISTICAL FRAMEWORK AND SIMULATION PROCEDURE

The statistical framework introduced in this section relies on a sample of n individuals. The set of observed binary phenotypes is given by the vector $\boldsymbol{y} = [y_1, \dots, y_n]$ where $y_i \in \{0, 1\}$ for all $i = 1, \dots, n$. Let us further consider that each individual has been genotyped for two SNP-sets (for example two genes) where each SNP-set is a collection of, respectively, p_1 and p_2 SNPs. The observed genotypes for the first SNP-set can be represented by a $n \times p_1$ matrix: $\boldsymbol{x^{(1)}} = [x_{ij}^{(1)}]_{i\in 1\dots n; j\in 1\dots p_1}$ where $x_{ij} \in \{0; 1; 2\}$ is the number of copies of the minor allele for SNP j carried by individual i. A similar representation is used for the second SNP-set where $\boldsymbol{x^{(2)}}$ is a $n \times p_2$ matrix. Finally, we assume that a collection of q covariates are likely to be measured for each individual. The observed covariates are stored in a $n \times q$ matrix $\boldsymbol{u} = [u_{ij}]_{i\in 1\dots n; j\in 1\dots q}$ where u_{i1} is constant, thus modeling the intercept in a regression framework.

Throughout this paper, we assume that each y_i is a realization of a two-class binary random variable Y. Furthermore, for each individual i, the genetic profiles $[x_{i1}^{(1)}, \dots, x_{ip_1}^{(1)}]$ and $[x_{i1}^{(2)}, \dots, x_{ip_2}^{(2)}]$ as well as the covariate profile $[u_{i1}, \dots, u_{iq}]$ are assumed to be realizations from random vectors $\boldsymbol{X^{(1)}}$, $\boldsymbol{X^{(2)}}$ and $\boldsymbol{U}$, respectively.

2.1 | Statistical models of phenotype-genotype association

In the remainder of this paper, we assume that the association between phenotype and genotypes is defined through the following general logistic model:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{X^{(1)}} = \boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}} = \boldsymbol{x^{(2)}}]) = \boldsymbol{w}'\boldsymbol{\alpha} + \boldsymbol{s}'\boldsymbol{\beta}, \quad (1)$$

where $\boldsymbol{w}$ is obtained by combining the covariates profile $\boldsymbol{u}$ (a q-vector) and the genotypic profiles $\boldsymbol{x^{(1)}}$ (a p_1-vector) and $\boldsymbol{x^{(2)}}$ (a p_2-vector). The vector $\boldsymbol{s}$ corresponds to the interaction profile, obtained by considering pairwise combinations of elements from marginal genotypic profiles $\boldsymbol{x^{(1)}}$ and $\boldsymbol{x^{(2)}}$. Finally, $\boldsymbol{\alpha}$ characterizes the set of parameters for covariate and main effects while $\boldsymbol{\beta}$ summarizes the interaction parameters.

However, a proper definition of association models depends on the modeling of the SNPs and in the rest of this section we focus on two current characterizations of SNP data: an allele-based modeling and a genotype-based modeling.

2.1.1 | Association model with allele-based modeling of SNPs

SNP data can first be specified as a quantitative variable that is coded additively (0,1,2) corresponding to the number of copies of the minor allelle. In that case, $\boldsymbol{w}$ is a $(q + p_1 + p_2)$-vector given by the raw concatenation of $\boldsymbol{u}$, $\boldsymbol{x^{(1)}}$, and $\boldsymbol{x^{(2)}}$:

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{x^{(1)}} \\ \boldsymbol{x^{(2)}} \end{pmatrix} = \left(u_1, \ldots, u_q, x_1^{(1)}, \ldots, x_{p_1}^{(1)}, x_1^{(2)}, \ldots, x_{p_2}^{(2)} \right)'.$$

Accordingly, the interaction profile $\boldsymbol{s}$ is defined as:

$$\boldsymbol{s} = \boldsymbol{x^{(1)}} \otimes \boldsymbol{x^{(2)}} = \left(x_1^{(1)} x_1^{(2)}, \ldots, x_1^{(1)} x_{p_2}^{(2)}, \ldots, x_i^{(1)} x_1^{(2)}, \ldots, x_i^{(1)} x_{p_2}^{(2)}, \ldots, x_{p_1}^{(1)} x_1^{(2)}, \ldots, x_{p_1}^{(1)} x_{p_2}^{(2)} \right)',$$

where $\otimes$ denotes the Kronecker product between two vectors.

2.1.2 | Association model with genotype-based modeling of SNPs

SNP genotypes can also be coded as categorical with two indicator variables, corresponding to the allele counts of 1 and 2, that compare the heterozygous category and the minor homozygous category, respectively to the reference homozygous category. For each genetic profile $\boldsymbol{x^{(\ell)}}, \ell = 1, 2$, the two following genotype-based profiles are introduced:

$$\boldsymbol{a^{(\ell)}} = (\mathbf{1}_{\{x_1^{(\ell)}=1\}}, \ldots, \mathbf{1}_{\{x_{p_\ell}^{(\ell)}=1\}})',$$
$$\boldsymbol{b^{(\ell)}} = (\mathbf{1}_{\{x_1^{(\ell)}=2\}}, \ldots, \mathbf{1}_{\{x_{p_\ell}^{(\ell)}=2\}})'.$$

Using this coding, the profile $\boldsymbol{w}$ is a $(q + 2p_1 + 2p_2)$-vector given by:

$$\boldsymbol{w} = \begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{a^{(1)}} \\ \boldsymbol{b^{(1)}} \\ \boldsymbol{a^{(2)}} \\ \boldsymbol{b^{(2)}} \end{pmatrix}.$$

Therefore, the interaction profile $\boldsymbol{s}$ is of dimension $4 \times p_1 \times p_2$ and defined as:

$$\boldsymbol{s} = \begin{pmatrix} \boldsymbol{a^{(1)}} \\ \boldsymbol{b^{(1)}} \end{pmatrix} \otimes \begin{pmatrix} \boldsymbol{a^{(2)}} \\ \boldsymbol{b^{(2)}} \end{pmatrix}.$$

2.2 | From SNP-SNP interaction to gene-gene interaction: a global testing approach

In this paper, global testing approaches refers to statistical tests that aim at combining pairwise interaction statistics (namely SNP-SNP interaction statistics) into a single gene-gene interaction statistic. Considering that the first gene has

p_1 SNPs and the second gene has p_2 SNPs, a global testing approach aims at combining $q = p_1 \times p_2$ pairwise test statistics that are jointly estimated in a multiple regression model.

Under the point of view of model (1), testing for the presence of gene-gene interaction effects amounts to testing the following hypotheses:

$$\begin{cases} H_0 \,:\, \boldsymbol{\beta} = \mathbf{0} \\ H_1 \,:\, \boldsymbol{\beta} \neq \mathbf{0}. \end{cases} \tag{2}$$

where $\boldsymbol{\beta}$ is the vector of interaction coefficients assuming the full regression model. For sake of clarity, it is noteworthy that $\boldsymbol{\beta}$ is obtained by considering all SNPs pairs in a single regression model (as in model (1)) rather than considering several regression models where only one SNP pair at a time.

Such set of hypotheses corresponds to a classical context of nested models comparison. However, because of the large dimension of $\boldsymbol{\beta}$ and the expected high correlation between elements of $\boldsymbol{\beta}$, traditional procedures, such as the chi-squared approximation of the likelihood ratio test, are not suitable to our context. We therefore consider that the system (2) falls into the paradigm of global testing which consists in testing for the significance of a subset of regression coefficients.[39] More precisely, global testing aims at combining the elements of a vector $\mathcal{Z}$ into a single statistic $T(\mathcal{Z})$. Then, the significance of $T(\mathcal{Z})$ is tested with respect to its distribution under H_0, thus providing a global test for the vector $\boldsymbol{\beta}$. In the following, we first introduce the vector $\mathcal{Z}$ considered in our context of gene-gene interaction testing. We then introduce several global statistics $T(\mathcal{Z})$ and the resampling procedure for estimating the distribution of $T(\mathcal{Z})$ under H_0 along with our estimator for the correlation matrix of $\mathcal{Z}$.

2.2.1 | Definition of $\mathcal{Z}$: a vector of pointwise statistics

Model (1) can be viewed as an extension of logistic regression models dedicated to SNP-set testing where a vector of score statistics has been used for $\mathcal{Z}$.[40,47,48] Similarly, we consider a vector of pointwise test statistics $\mathcal{Z} = (Z_1, \ldots, Z_p)'$ with $p = p_1 p_2$ (resp. $4p_1p_2$) if the allele-based (resp. genotype-based) coding is considered and where Zki is the score statistic associated with the ith pair in the $\boldsymbol{s}$ profile. More precisely, let us introduce $\mathbb{W}$ and $\mathbb{S}$ the two design matrices corresponding to the main terms and the interaction terms. It can be remarked that the dimensions of $\mathbb{W}$ are $n \times (q + p_1 + p_2)$ under an allele-based coding and $n \times (q + 2p_1 + 2p_2)$ under a genotype-based coding while it equals $n \times (p_1p_2)$ and $n \times (4p_1p_2)$ for $\mathbb{S}$ under the same codings. We further denote $\hat{\boldsymbol{y}}_{\mathbf{0}}$ the vector whose ith coordinate $\hat{y}_{0,i}$ is the estimated probability under H_0 that the ith individual is diseased. Under model (1), $\hat{y}_{0,i}$ is given by:

$$\hat{y}_{0,i} = \frac{\exp(\boldsymbol{w}_{\boldsymbol{i}}'\hat{\boldsymbol{\alpha}}_{\mathbf{0}})}{1 + \exp(\boldsymbol{w}_{\boldsymbol{i}}'\hat{\boldsymbol{\alpha}}_{\mathbf{0}})}, \tag{3}$$

where $\boldsymbol{w_i}$ is the ith row of $\mathbb{W}$ and $\hat{\boldsymbol{\alpha}}_{\mathbf{0}}$ is the estimator of $\boldsymbol{\alpha}$ in model (1) under H_0. Then, by adapting the definition of the test statistics used in Reference 40 to our model (1), we define Z_k as:

$$Z_k = \frac{\boldsymbol{S}_{\boldsymbol{j}}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})}{\sqrt{\hat{\boldsymbol{\Gamma}}_{j,j}}},$$

where $\boldsymbol{S_j}$ is the jth column of $\mathbb{S}$ and $\hat{\boldsymbol{\Gamma}}_{j,j}$ is the jth diagonal term of the estimated correlation matrix of the vector $\mathbb{S}'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})$:

$$\hat{\boldsymbol{\Gamma}} = \hat{\sigma}_Y^2(\mathbb{S}'\mathbb{S} - \mathbb{S}'\mathbb{W}(\mathbb{W}'\mathbb{W})^{-1}\mathbb{W}'\mathbb{S}), \tag{4}$$

with $\hat{\sigma}_Y^2 = \frac{1}{n}(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}_{\mathbf{0}})$.

As quoted in Reference 40, $\mathcal{Z}$ can be assumed to have a multivariate normal asymptotic distribution so that $\mathcal{Z} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Testing for (2) is therefore equivalent to testing for the global nullity of $\boldsymbol{\mu}$:

$$\begin{cases} H_0 \,:\, \boldsymbol{\mu} = \mathbf{0} \\ H_1 \,:\, \boldsymbol{\mu} \neq \mathbf{0}. \end{cases} \tag{5}$$

It can be noted that obtaining a stable estimate for $\boldsymbol{\alpha_0}$ is challenging in situations where q, p_1, and p_2 are large and the SNPs are strongly correlated. To circumvent this issue, usual regularization techniques can be used, such as feature selection or penalized regression methods (ridge regression was used for a very similar problem in Reference 49, in the context of gene-environment interaction effect testing). However, penalized regression methods require the selection of a value of one or several hyperparameters. This is often performed using cross validation, which would be cumbersome in the present context. Moreover, this can also raise issues in terms of type I error rate control. Dimensionality reduction methods can also be useful in this context, such as principal components analysis, thus taking advantage of the strong correlation structure among SNPs. SNP matrices $\boldsymbol{x^{(1)}}$ and $\boldsymbol{x^{(2)}}$ can then be replaced with their corresponding first principal components. In the following, SNP profiles are replaced with the set of corresponding principal components, which individually account for at least 0.1% of the total variance profiles, to estimate $\boldsymbol{\alpha_0}$. Based on simulations, the threshold of 0.1% corresponds to the best choice to correctly control for type I error rate, especially for SNP sets larger than 100 SNPs.

2.2.2 | Methods for combining elements of $\mathcal{Z}$

Under the testing framework (5), a common approach is to compute a global test statistic $T(\mathcal{Z})$ by aggregating the coordinates of $\mathcal{Z}$ as proposed in SNP-set testing.[40,41,47,50] Among the numerous existing aggregation methods, one of the most simple and popular ones is the minP (or maxT) procedure.[37,40,51] This method consists in defining $T(\mathcal{Z})$ as:

$$T_{\max}(\mathcal{Z}) = \max_{1\leq k\leq p} |Z_k|.$$

Despite its simplicity, the minP procedure is generally robust and efficient,[37,40] which makes it one of the most used procedures for such problems. It was also used in Reference 37 for gene-gene interaction testing by considering p_1p_2 logistic models and the corresponding Wald statistics for the coefficient corresponding to the interaction effect.

Another natural way of aggregating statistics consists in considering the squared L^2-norm of $\mathcal{Z}$ (see Reference 41 for example). It is defined as:

$$T_{L^2}(\mathcal{Z}) = \sum_{k=1}^{p} Z_k^2.$$

The HC statistic, introduced in Reference 42, is defined as:

$$T_{\mathrm{HC}}(\mathcal{Z}) = \max_{1\leq k\leq p/2} \sqrt{p}\frac{k/p - p_{(k)}}{\sqrt{p_{(k)}(1-p_{(k)})}},$$

where the $p_{(k)}$ is the p-values sorted in ascending order $p_k = 2(1-\Phi(|Z_k|))$, where Φ is the cdf of the standard normal distribution. It is noteworthy that the computation of aggregating statistics $T_{\max}(\mathcal{Z})$, $T_{L^2}(\mathcal{Z})$ and $T_{\mathrm{HC}}(\mathcal{Z})$ only depends on the computation of the main test statistics Z_k.

Other aggregating statistics have been introduced where the correlation between the Z_k's is explicitly accounted for in the definition of the aggregation. Hotelling's t^2 statistic, used in Reference 43 for example, can then be defined as:

$$T_{\mathrm{Hotelling}}(\mathcal{Z}) = \mathcal{Z}'\hat{\boldsymbol{\Sigma}}^{-1}\mathcal{Z},$$

where $\hat{\boldsymbol{\Sigma}}$ is the estimated correlation matrix of $\mathcal{Z}$. Therefore, to compute the Hotelling's t^2 statistic, the correlation matrix of $\mathcal{Z}$ must be estimated and inverted which should be made with caution.

2.2.3 | Estimation of Σ^{-1}: the inverse of the correlation matrix of $\mathcal{Z}$

The dimension of the correlation matrix Σ is $p \times p$, where $p = (p_1p_2)$ or $p = (4p_1p_2)$ depending on the chosen coding for the SNPs, which increases quadratically with the size of the SNP-sets. Even for SNP-sets of moderate size (ie, between 20 and 50 SNPs), it might be impossible to invert Σ, and the computational cost associated to this operation would be burdensome. However, it should be noticed that $\hat{\boldsymbol{\Sigma}}$ shows a very specific structure resulting from a direct combination

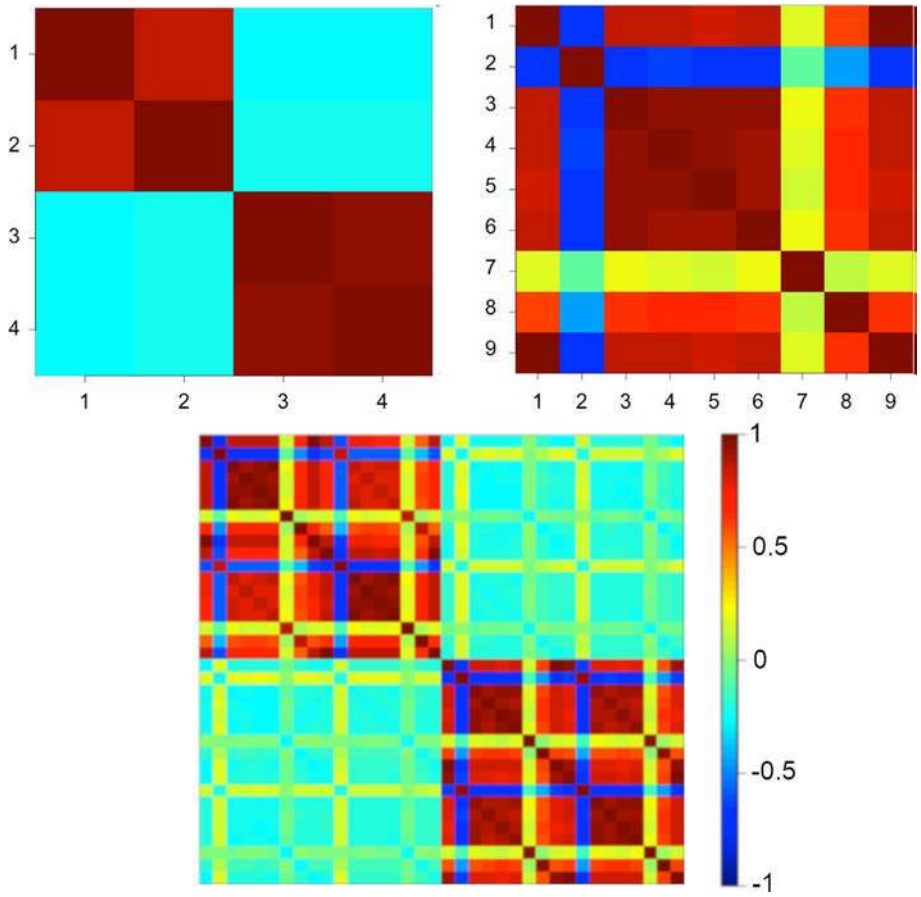

FIGURE 1 Correlation matrices of two observed SNP-sets by considering allele-based coding for SNPs. SNP-sets are composed of four and nine SNPs, respectively. The upper part draws the genotype correlation matrices of two SNP-sets. The lower part displays the correlation matrix for the test statistics vector $\mathcal{Z}$ obtained with simulations

of the correlation structures of the two SNP-sets. Let us illustrate this point by considering 2000 observed genotypes profiles randomly selected in the WTCCC cohorts.[52] Genotype profiles were randomly split into two sets of population, to mimic 1000 cases and 1000 controls, and restricted to two main SNP-sets with four and nine SNPs, respectively. Based on random permutations, the correlation matrix of the vector of test statistics using both allele-based coding (36 pointwise statistics) and genotype-based coding (144 pointwise statistics) was estimated using Equation (4). Indeed, results displayed on Figure 1 for the allele-based coding and on Figure 2 for the genotype-based coding, show that $\hat{\boldsymbol{\Sigma}}$ has a specific structure directly inherited from genotype correlations, namely $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$. Along the diagonal, it can be observed that blocks share the same structure as $\boldsymbol{\Sigma_2}$. Aside from the diagonal blocks, it can be seen that $\hat{\boldsymbol{\Sigma}}$ also has a block structure, where each block has the shape of $\boldsymbol{\Sigma_2}$. Each $\boldsymbol{\Sigma_2}$-like block is weighted by a coefficient that fits with the elements of $\boldsymbol{\Sigma_1}$, thus suggesting that $\hat{\boldsymbol{\Sigma}}$ can be reasonably approximated by $\boldsymbol{\Sigma_1} \otimes \boldsymbol{\Sigma_2}$, the Kronecker product between $\boldsymbol{\Sigma_1}$ and $\boldsymbol{\Sigma_2}$. This result is formalized in the following Theorem 1.

Theorem 1. *Let $\hat{\boldsymbol{\Sigma}}$ be the estimator of the correlation matrix associated to the following covariance matrix:*

$$\hat{\boldsymbol{\Gamma}} = \frac{1}{n}(\mathbb{S}'\mathbb{S} - \mathbb{S}'\mathbb{W}(\mathbb{W}'\mathbb{W})^{-1}\mathbb{W}'\mathbb{S}),$$

introduced in Equation (4). Let us assume that the covariances between the two genetic profiles and covariate profile is zero: $Cov(\boldsymbol{X^{(1)}}, \boldsymbol{X^{(2)}}) = 0$, $Cov(\boldsymbol{X^{(1)}}, \boldsymbol{U}) = 0$ and $Cov(\boldsymbol{X^{(2)}}, \boldsymbol{U}) = 0$. Let $\boldsymbol{\Sigma_1}$ be the correlation matrix of the first gene $\boldsymbol{X}^{(1)}$ and $\boldsymbol{\Sigma_2}$ be the correlation matrix of the second gene $\boldsymbol{X}^{(2)}$. Then

$$\hat{\boldsymbol{\Sigma}} \xrightarrow{\mathbb{P}} \boldsymbol{\Sigma_1} \otimes \boldsymbol{\Sigma_2}.$$

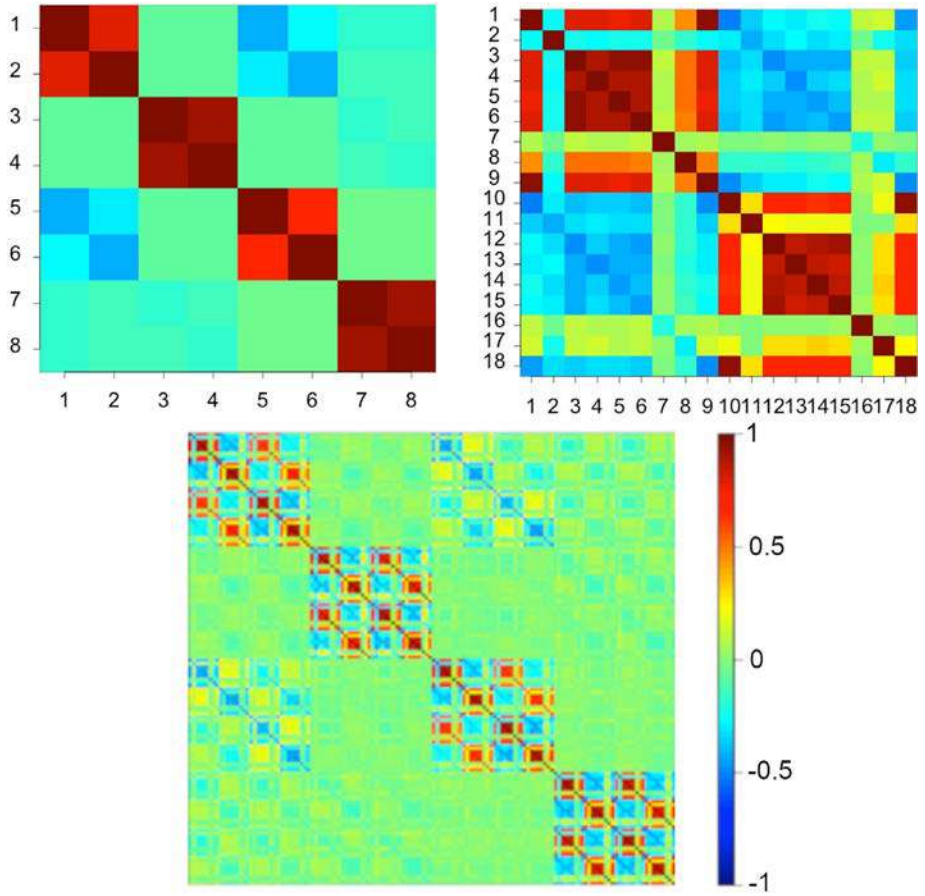

FIGURE 2 Correlations matrices of two observed SNP-sets by considering genotype-based coding for SNPs. The upper part draw the genotype correlation matrices of two SNP-sets. The lower part displays the correlation matrix for the test statistics vector $\mathcal{Z}$ obtained with simulations. SNP-sets are composed of four and nine SNPs, respectively, so that genotype matrices with genotype-based coding have respective dimensions of 8x8 and 18x18. The dimensions for the correlation matrix of interaction parameters are 144×144

Proof of Theorem 1 is given in Appendix S1. It is noteworthy that the hypothesis $\text{Cov}(\boldsymbol{X^{(1)}}, \boldsymbol{X^{(2)}}) = 0$ amounts to considering unlinked SNP-sets which is reasonable since interaction is tested between distant genomic regions along the genome. On the other hand, the two hypotheses $Cov(\boldsymbol{X^{(1)}}, \boldsymbol{U}) = 0$ and $\text{Cov}(\boldsymbol{X^{(2)}}, \boldsymbol{U}) = 0$, that stipulate that the set of covariates are not correlated with any SNP-sets, are also reasonable since considering gene-environment interaction is beyond the scope of the present paper. The decomposition proposed in Theorem 1 offers a clear advantage for approximating the inverse of $\hat{\boldsymbol{\Sigma}}$. Indeed, let us introduce the eigenvalue decompositions of $\Sigma_{\mathbf{1}}$ and $\Sigma_{\mathbf{2}}$ as follows:

$$\Sigma_{\mathbf{1}} = \boldsymbol{U\Lambda U'}, \quad \boldsymbol{U} = [\boldsymbol{u_1} \mid \ \dots \ \mid \boldsymbol{u_{p_1}}], \quad \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_{p_1})$$
$$\Sigma_{\mathbf{2}} = \boldsymbol{V\Omega V'}, \quad \boldsymbol{V} = [\boldsymbol{v_1} \mid \ \dots \ \mid \boldsymbol{v_{p_2}}], \quad \boldsymbol{\Omega} = \text{diag}(\omega_1, \dots, \omega_{p_2}).$$

Then, the eigenvalue decomposition of $\Sigma_{\mathbf{1}} \otimes \Sigma_{\mathbf{2}}$ is $\boldsymbol{A\Theta A'}$ (see Reference 53 for example), where

$$\boldsymbol{A} = [\boldsymbol{a_{11}} \mid \ \dots \ \mid \boldsymbol{a_{1p_2}} \mid \ \dots \ \mid \boldsymbol{a_{i1}} \mid \ \dots \ \mid \boldsymbol{a_{ip_2}} \mid \ \dots \ \mid \boldsymbol{a_{p_1 1}} \mid \ \dots \ \mid \boldsymbol{a_{p_1 p_2}}], \quad \boldsymbol{a_{ij}} = \boldsymbol{u_i} \otimes \boldsymbol{v_j},$$
$$\boldsymbol{\Theta} = \boldsymbol{\Lambda} \otimes \boldsymbol{\Omega}.$$

Consequently, the eigenvalue decomposition of $\Sigma_{\mathbf{1}} \otimes \Sigma_{\mathbf{2}}$, which is a $(p_1p_2) \times (p_1p_2)$ or a $(4p_1p_2) \times (4p_1p_2)$ matrix, can be obtained through the eigenvalue decompositions of a $p_1 \times p_1$ matrix and a $p_2 \times p_2$ matrix. In the remainder of this paper, the previous decomposition is used to estimate the inverse or the eigenvalue decomposition of $\hat{\boldsymbol{\Sigma}}$, especially for computing Hotelling's t^2 statistic.

2.2.4 | Testing for the significance of global test statistics

The use of an aggregating statistic in a global testing approach requires the evaluation of its null distribution. However, for any aggregation method among those introduced in the former section, the distribution of $T(\mathcal{Z})$ under the null hypothesis does not admit a closed-form expression. To overcome such limitation, resampling-based methods, such as random permutations of the phenotype, are widely used to approximate the null distribution of the test statistic. Nevertheless, it can be remarked that in our context, this approach would yield invalid results.[46] Indeed, under H_0, model (1) becomes:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{U} = \boldsymbol{u}, \boldsymbol{X^{(1)}} = \boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}} = \boldsymbol{x^{(2)}}]) = \boldsymbol{w'\alpha}, \tag{6}$$

meaning that even under H_0, covariates and main effects might be present. On the other hand, randomly shuffling the phenotype amounts to assuming that $\boldsymbol{\alpha} = \boldsymbol{\beta} = \boldsymbol{0}$, which is a stronger hypothesis than H_0. Hence, using random permutations of the phenotype could lead to inflated type I error rates. This issue is discussed in details in Reference 46, in which a parametric bootstrap procedure is considered to counteract this problem (see also References 54 and 55 for the similar problem of testing in presence of confounders).

Similarly as in References 46,54, we therefore propose to estimate the null distribution of the global statistic $T(\mathcal{Z})$ by a parametric bootstrap procedure, which can be described as follows. First, main and covariate effects, $\hat{\boldsymbol{\alpha}}_{\boldsymbol{0}}$ of $\boldsymbol{\alpha}$, are estimated under the null hypothesis of no interaction (ie, using model (6)). Given an observed covariate and genotypic profile $\boldsymbol{w_i}$, $\hat{\boldsymbol{\alpha}}_{\boldsymbol{0}}$ is plugged into Equation (3) to estimate $\hat{y}_{0,i}$ (the probability for the ith individual to be diseased). The simulated phenotype for the ith individual is then generated according to the corresponding Binomial distribution, $\%(\hat{y}_{0,i})$. By applying such a procedure to the n observed profiles, a vector of simulated phenotypes under the null hypothesis is obtained where main and covariate effects are preserved. The distribution under H_0 of each aggregating statistic is approximated by simulating a fixed number of phenotypes (eg, 10 000).

2.3 | Data-driven simulation procedure

2.3.1 | Genotype and phenotype simulation

Data-driven simulations have been used to assess for the control of the Type-I error rate as well as for power analysis. Results presented in Section 4 and in Appendix S1 were obtained according to the following simulation procedure. Considering an observed pair of SNP-sets, our data-driven simulation procedure aims at generating matrices of 100 000 genetic profiles for each SNP-set. We used the R package `GenOrd`[56] to simulate these profiles with respect to the truly observed correlation structures and main effects distributions.

For a profile $\boldsymbol{X^{(1)}}$ corresponding to the first gene and a profile $\boldsymbol{X^{(2)}}$ corresponding to the second gene, the phenotype is then generated according to a logistic model as proposed in Equation (7). The disease model is then defined by a vector $\boldsymbol{\alpha}$ and an interacting vector $\boldsymbol{\beta}$; once the phenotype is generated, a sample is obtained by randomly sampling 1000 cases and 1000 controls. For sake of clarity we do not consider covariates.

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X^{(1)}} = \boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}} = \boldsymbol{x^{(2)}}]) = \boldsymbol{w'\alpha} + \boldsymbol{s'\beta}. \tag{7}$$

2.3.2 | Control of the Type-I error

The control of the Type-I error has been assessed under four situations by considering (1) no main effect, (2) additive main effect, (3) recessive main effect, and (4) dominant main effect. In all situations, data are simulated under the null hypothesis of absence of interaction, thus meaning that $\boldsymbol{\beta} = 0$ in (7). To simulate main effects, the SNP i_1 in the first gene and the SNP i_2 are assumed to be marginally associated with the disease. The vector of coefficients $\boldsymbol{\alpha}$ can be decomposed into four vectors $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1^{(1)}, \boldsymbol{\alpha}_2^{(1)}, \boldsymbol{\alpha}_1^{(2)}, \boldsymbol{\alpha}_2^{(2)}]$ corresponding, respectively, to the vectors of main effects for heterozygote genotype in the first gene, the homozygote genotype in the first gene, the heterozygote genotype in the second gene and the homozygote genotype in the second gene. We then consider that $\boldsymbol{\alpha}_1^{(1)} = [0, \ldots, 0, \alpha_{1,i_1}^{(1)}, 0, \ldots, 0]$, $\boldsymbol{\alpha}_2^{(1)} = [0, \ldots, 0, \alpha_{2,i_1}^{(1)}, 0, \ldots, 0]$, $\boldsymbol{\alpha}_1^{(2)} = [0, \ldots, 0, \alpha_{1,i_2}^{(2)}, 0, \ldots, 0]$, $\boldsymbol{\alpha}_2^{(2)} = [0, \ldots, 0, \alpha_{2,i_2}^{(2)}, 0, \ldots, 0]$.

More precisely, the four disease models without interaction used in this study are given by:

- No main effect: $\boldsymbol{\beta} = 0$ and $\boldsymbol{\alpha} = 0$
- Additive main effect: $\boldsymbol{\beta} = 0$, $\alpha^{(1)}_{1,i_1} = 0.5$, $\alpha^{(1)}_{2,i_1} = 1$, $\alpha^{(2)}_{1,i_2} = 0.5$ and $\alpha^{(2)}_{2,i_2} = 1$
- Recessive main effect: $\boldsymbol{\beta} = 0$, $\alpha^{(1)}_{1,i_1} = 0$, $\alpha^{(1)}_{2,i_1} = 1$, $\alpha^{(2)}_{1,i_2} = 0$ and $\alpha^{(2)}_{2,i_2} = 1$
- Dominant main effect: $\boldsymbol{\beta} = 0$, $\alpha^{(1)}_{1,i_1} = 1$, $\alpha^{(1)}_{2,i_1} = 1$, $\alpha^{(2)}_{1,i_2} = 1$ and $\alpha^{(2)}_{2,i_2} = 1$

The estimation of the Type-I error rate is performed by simulating 100 000 datasets. For each dataset, *P*-values are computed and compared to three nominal levels (0.001, 0.01, and 0.05), by using the R packages `GeneGeneInteR`,[57] `SPA3G`,[33] and `GeneGeneInteractions`. For each method, the estimated Type-I error is given by the proportion, over the 100 000 simulations, of *P*-values lower than the corresponding nominal level.

2.3.3 | Power analysis

Power studies presented in this paper are based on disease models where the interaction between at least one SNP pair is associated with the phenotype. For sake of clarity, we do not consider any main effect thus leading to $\boldsymbol{\alpha} = 0$ in Equation (7). The various disease models are then characterized by $\boldsymbol{\beta}$ vectors where some coefficients are different from zero in order to mimic an interaction effect.

Disease models with only one causal SNP pair are first considered. As proposed in Reference 10, such disease models are presented by a 3×3 table of odds where each cell characterizes the odds of the disease with respect to the genotype of the causal pair. Each model has two parameters: γ characterizes the baseline odds and θ quantifies the strength of the disease-genotype relationship thus providing a vector of $\boldsymbol{\beta}$ coefficients plugged into the model in Equation (7) to simulate a genotype/phenotype dataset. For a given $\boldsymbol{\beta}$ vector, power is estimated by the proportion of datasets, over a total of 1000 simulated datasets, with a *P*-value lower that a nominal level. In our studies, the nominal level has been set to 0.05. It can be remarked that power from other nominal levels have been investigated and provided similar conclusions. Power curves are then obtained by computing powers for $\theta \in [0, 0.1, \ldots, 1]$ for eight disease models that cover the wide scope of possible epistatic models (see Section S4 for more details). More precisely, we investigate power on (1) traditional epistatic disease models by considering Recessive-Recessive, Dominant-Dominant and Recessive-Dominant models[9] and (2) other observed models by studying Interaction-Multiplicative, Interface, Threshold, Modifying Effect and XOR models introduced in Reference 58.

For complex disease models where more than one SNP pair is causal, a vector $\boldsymbol{\beta}$ is set into Equation (7). Similar to the one causal SNP pair situation, power is estimated by the proportion, over 1000 simulated datasets, of *P*-values lower than the nominal level of 0.05. Power curves are obtained by considering a gradient of strength of phenotype/genotype association through a weight parameter $\theta \in [0, 0.01, \ldots, 1]$ applied to $\boldsymbol{\beta}$. Therefore, power curves display the estimated power for the model with interacting vector $\theta\boldsymbol{\beta}$ with respect to θ.

3 | AN OMNIBUS TEST

3.1 | Choice of global test statistics

In a preliminary study, we aim at evaluating the performance of the eight global testing approaches introduced in the previous sections, namely minP, L^2-norm, HC and Hotelling for both allele-based and genotype-based modeling, labeled with a letter *c* or *d* as a suffix (minP_*c* therefore corresponds to minP based on a allele-based coding of SNPs). We first investigate the control of the Type-I error using the data-driven simulation procedure presented in Section 2.3. Results, presented in Section S2, show that in the large majority of situations, all methods correctly control the Type-I error rate. We then performed a power study as described in the previous section and also based on the two genomic regions introduced in the previous paragraph and illustrated in Figures 1 and 2. Powers have been compared under various phenotype-genotype associations based on the logistic model defined in Equation (1). The different simulated scenarios correspond to different choices for the vector of interaction parameters $\boldsymbol{\beta}$.

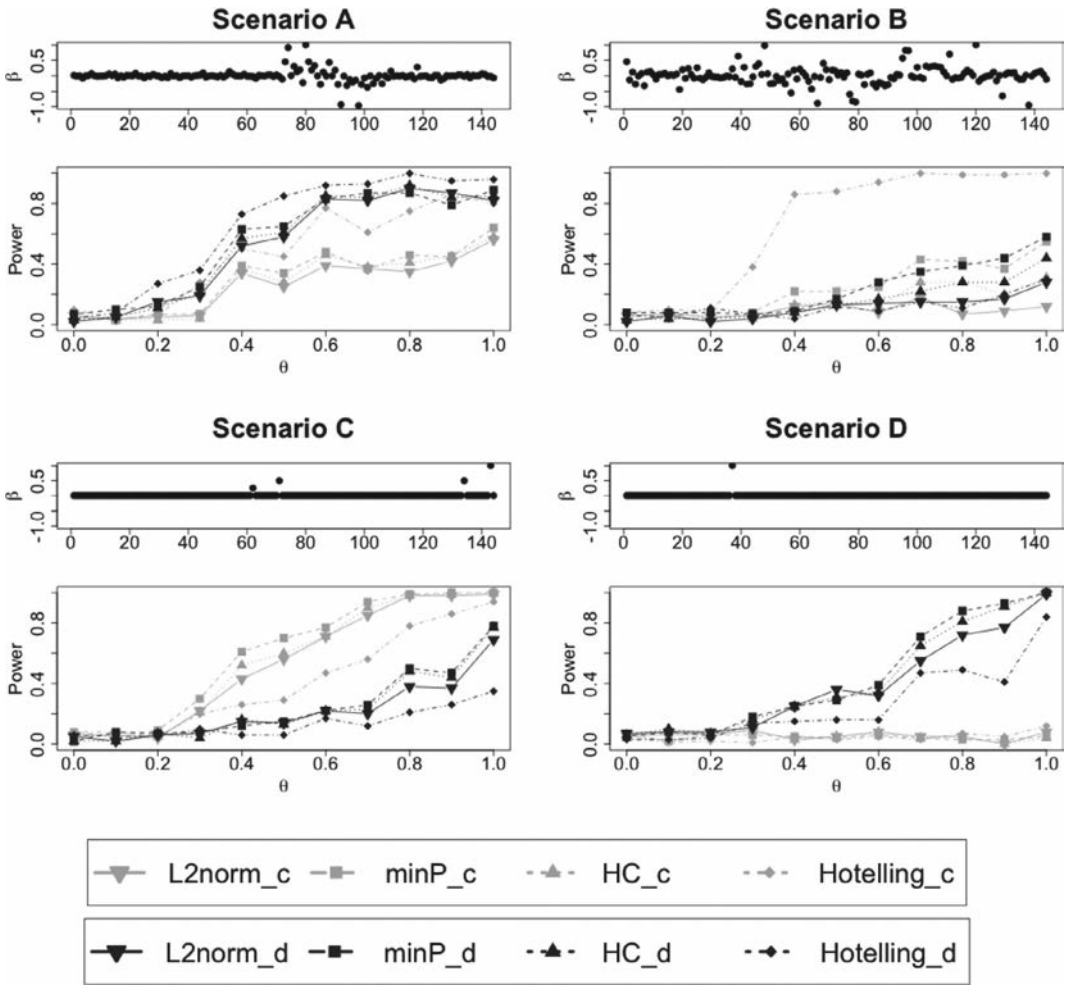

FIGURE 3 Power curves of global regression test statistics for various preliminary phenotype-genotypes association models. Genotypes observed in Figures 2 were used to simulate phenotypes according to four different vectors $\boldsymbol{\beta}$ corresponding to the two scenarios with one causal pair (C and D) and two scenarios with more than one causal pair (A and B) (See Table S1 for details regarding each scenario). Powers have been estimated with 1000 replicates of two genes (with four and nine SNPs, respectively) and sample sizes of 1000 cases and 1000 controls. For each scenario, the upper graphic shows the value of the regression coefficients ordered according to the definition of the interaction profile vector $\boldsymbol{s}$ defined in Section 2.1 for a genotype-based profile

Figure 3 displays the power curves obtained under four scenarios (ie, four different vectors $\boldsymbol{\beta}$ and 11 weights θ as described in the previous section). It can first be remarked that for scenarios A and D, methods based on a genotype-based coding of the interaction outperform methods based an allele-based coding. Among genotype-based methods, Hotelling_d is more powerful than minP_d, L^2_d, and HC_d in some situations (scenario A) and less powerful in other situations (scenario D). Methods based on a allele-based coding can also have more power than methods based on a genotype-based coding as illustrated by scenario C. Moreover, for scenario C, Hotelling_c lacks in power compared to other allele-based methods. However, Scenario B shows another tendency where Hotelling_c largely outperforms all other methods. Therefore, the four illustrative scenarios in Figure 3demonstrate that the variation in power between global testing approaches depends on (1) the coding of the SNPs (allele-based vs genotype-based) and (2) the way the dependence between main test statistics is accounted for (Hotelling vs minP, HC, and L^2-norm).

More generally, our preliminary results indicate that the choice of the most powerful test depends on the interplay between the correlation structure of main tests and the phenotype-genotypes association model. Our preliminary conclusion is supported by an additional power study presented in Section S2. This study further demonstrates that the relative performance of each method also depends on the interplay between the localization of the causal signal and the joint correlation structure. The heterogeneity between global testing methods based on several logistic models is further highlighted by results displayed in Figures S2 and S3.

Therefore, using only one test or one type of coding is likely to lack in power in many situations and, therefore, considering all tests together is necessary to be able to detect a wide range of interaction effects with satisfying power. Other gene-gene interaction methods could also be combined such as PCA or CCA for example. However, introducing a new method in a combining test should be performed with the hope that this new method is complementary to other methods. In our context, we consider that the eight methods minP_*c*, L^2-norm_*c*, HC_*c*, Hotelling_*c*, minP_*d*, L^2-norm_*d*, HC_*d*, and Hotelling_*d* allow for detecting a large variety of disease models. In the following we therefore restrict to the combination of these eight methods.

3.2 | An omnibus strategy

In order to use all global tests in a single procedure, we proposed to combine the eight above introduced tests (minP, L^2-norm, HC, and Hotelling for both allele-based and genotype-based modeling) in an omnibus test. To aggregate the eight global tests, we first used the Cauchy-combination test as proposed in Reference 45. However, to explicitly account for the correlation between the eight individual tests we also introduced the following resampling-based method. First, let $\boldsymbol{q}$ be the vector containing the eight *P*-values associated to the eight former global tests. Then, let φ be a function, which takes as argument the vector $\boldsymbol{q}$ and returns an associated combined *P*-value. We denote $\boldsymbol{q}_{\boldsymbol{k}}^{(\boldsymbol{0})}, 1 \leq k \leq N$ the vectors of *P*-values obtained using simulated phenotypes under the null hypothesis, where *N* is the number of desired simulated phenotypes (eg, $N = 1000$). The final *P*-value of our testing procedure is defined as:

$$P = \frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{\{\varphi(\boldsymbol{q}) \leq \varphi(\boldsymbol{q}_{\boldsymbol{k}}^{(\boldsymbol{0})})\}}. \tag{8}$$

In the following, as recommended in Reference 59, we consider φ to be the Simes' combining procedure,[60] that is:

$$\varphi(\boldsymbol{q}) = \min_{1 \leq i \leq 8} \frac{8q_{(i)}}{i},$$

were $q_{(i)}$ is the ith coordinate of $\boldsymbol{q}$, sorted in ascending order.

In our omnibus strategy, we aim at combining the eight individual global testing procedures. However, since other combinations of individual methods can be considered, we also tested the aggregation of other subsets. It appears that the combination of eight methods is the best combination since it is the most robust in all simulated situations. For example, if Hotelling_*c* is not considered in the aggregation, the omnibus strategy has limited power in Scenario B displayed in Figure 3.

4 | EVALUATION OF THE OMNIBUS TESTS

In this section, the performance of our testing procedure is evaluated using both data-driven simulations and observed genotype-phenotype data. Data-driven simulations are first used to verify that the type I error rate is rightfully controlled by our omnibus by resampling test. Then, the power of our test is evaluated through a large set of association models. Considering an observed pair of SNP-sets, our data-driven simulation procedure aims at generating matrices of 100 000 genetic profiles for each SNP-set. We used the R package `GenOrd`[56] to simulate these profiles with respect to the truly observed correlation structures and main distributions. For a profile $\boldsymbol{X}^{(\boldsymbol{1})}$ corresponding to the first gene and a profile $\boldsymbol{X}^{(\boldsymbol{2})}$ corresponding to the second gene, the phenotype is then generated according to a logistic model as proposed in Equation (1). The disease model is then defined by a vector $\boldsymbol{\alpha}$ and an interacting vector $\boldsymbol{\beta}$; once the phenotype is generated, a sample is obtained by randomly sampling 1000 cases and 1000 controls. This process is repeated 1000 times for each tested vector $\boldsymbol{\beta}$ so that, our omnibus by resampling test and the competitive methods can be compared on the generated samples. It is noteworthy that for sake of clarity we neither considered covariate nor main effects ($\boldsymbol{\alpha} = 0$) unless explicitly mentioned. However, our conclusions remain valid even if $\boldsymbol{\alpha} \neq 0$ (data not shown). Results presented in the next sections were obtained with the pair of SNP-sets previously introduced in Figures 1 and 2.

In this section, the performances of our testing procedure and other methods, namely Cauchy,[45] PCA,[28] CCA,[29] CLD,[35] Aggregator,[37] and SPA[33] methods, are evaluated using both data-driven simulations and observed genotype-phenotype

data. Data-driven simulations are first used to verify that the type I error rate is rightfully controlled by our omnibus by resampling test and other methods. Then, the power of our test is evaluated and compared to other methods through a large set of association models. Results presented in the next sections were obtained with the pair of SNP-sets previously introduced in Figures 1 and 2.

4.1 | Control of the type I error rate

To assess the control of the type I error rate, the phenotypes are generated according to the following model, where no covariate effect is considered:

$$\text{logit}(\mathbb{P}[Y = 1|\boldsymbol{X^{(1)}} = \boldsymbol{x^{(1)}}, \boldsymbol{X^{(2)}} = \boldsymbol{x^{(2)}}]) = \alpha_0 + \boldsymbol{x^{(1)\prime}}\boldsymbol{\alpha^{(1)}} + \boldsymbol{x^{(2)\prime}}\boldsymbol{\alpha^{(2)}}.$$

We investigate the control of the type I error rate by considering either the absence or the presence of main effects. The presence of main effects has been simulated by randomly adding nonzero values if the marginal vectors of coefficients $\boldsymbol{\alpha^{(1)}}$ and $\boldsymbol{\alpha^{(2)}}$. The obtained empirical type I error rates and corresponding confidence intervals are given in Tables 1 and 2. It can be remarked that the empirical type I error rate is always close to the nominal level. Moreover, the nominal level is always in the confidence interval, meaning that the type I error rate is properly controlled, even if there are main genetic effects. It can be remarked that these results are consistent with the study of the type I error rate for the eight components of the omnibus tests presented in Section 3.1 and Tables S2 and S3. This demonstrates that the parametric bootstrap is accurate for taking the main effects into account. Our results are confirmed in Section S5.2 where other pairs of SNP-sets are considered.

Following the simulation procedure described in Section 2.3, control of the type I error rate is investigated by considering either the absence or the presence of main effects. The obtained empirical type I error rates and corresponding confidence intervals are given in Tables 1 and 2 for each method and additive, recessive, and dominant main effects. In can first be remarked that in all situations our omnibus by resampling test correctly controls for type I error since the nominal level is always in the confidence interval. The kernel method also shows a proper control of the type I error in almost all situations. Conversely, Cauchy and CCA methods fail at controlling type I error in all situations. More precisely Cauchy method always slightly overestimates type I error especially for dominant main effect, while CCA largely underestimates type I error. The slight difference between our resampling-based omnibus test and the Cauchy-combination omnibus test can be explained by the fact that the Cauchy-combination test does not explicitly account for the correlation between individual tests. In absence of main effects CLD and PCA have good properties while Aggregator and SPA hardly control type I error at a nominal level of 0.001. In presence of main effects (additive, recessive, or dominant), CLD fails at controlling type I error at the nominal level of 0.001. SPA and PCA have a similar issue with the nominal level of 0.001 and further show an abnormal overestimation of the type I error in presence of dominant main effect. The uncontrolled Type I error for CLD and PCA has previously been observed and discussed in other studies.[61]

4.2 | Power study

In this section, we aim at evaluating and comparing the statistical power of our proposed omnibus procedure to the power of other methods. The global tests introduced in Section 2.2.2 were also included in the comparison to investigate the robustness of our test. Although the results obtained in the previous section show that some methods fail at rightfully controlling the Type I error, we consider all methods in the power studies. Some of the following results have therefore to be interpreted with caution.

4.2.1 | Data-driven simulations with one causal pair

We focus here on disease models where only one pair of SNPs is considered as causal. As proposed in Reference 10, disease models are presented by a 3×3 table of odds where each cell characterizes the odds of the disease with respect to the genotype of the causal pair. Each model has two parameters: γ characterizes

TABLE 1 Empirical type I error rates considering either absence of main effect or additive main effects for three different nominal levels (0.001, 0.01 and 0.05). Methods named Omnibus and Cauchy aim at combining eight global tests derived from logistic multiple regression model while CCA, CLD, PCA, Aggregator, Kernel, and SPA are single global tests. Type I error rates have been estimated with 100 000 replicates of two genes (with four and nine SNPs, respectively). The sample sizes are 1000 cases and 1000 controls. Numbers in italic refer to targeted nominal level. Numbers in bold correspond to situations with a correct control of the type I error. Confidence intervals, in square brackets, are based on normal approximation to the binomial proportion

		Omnibus by resampling	**Omnibus by Cauchy combination**
No main effect	0.001	**0.0010**	0.0016
		[0.0009,0.0011]	[0.013,0.0019]
	0.01	**0.012**	0.015
		[0.009,0.014]	[0.013,0.018]
	0.05	**0.054**	0.066
		[0.049,0.058]	[0.060,0.070]
Additive main effect	0.001	**0.0012**	0.0014
		[0.0010,0.0014]	[0.0012,0.0016]
	0.01	**0.013**	0.015
		[0.010,0.016]	[0.013,0.018]
	0.05	**0.054**	0.063
		[0.049,0.058]	[0.0583,0.0679]
Recessive main effect	0.001	0.0007	0.0017
		[0.0005,0.0009]	[0.0015,0.0019]
	0.01	**0.008**	0.017
		[0.006,0.010]	[0.015,0.018]
	0.05	**0.053**	0.068
		[0.050,0.056]	[0.065,0.073]
Dominant main effect	0.001	**0.0008**	0.0018
		[0.0006,0.0010]	[0.0015,0.0020]
	0.01	**0.009**	0.024
		[0.007,0.011]	[0.02113,0.02724]
	0.05	**0.052**	0.087
		[0.047,0.057]	[0.082,0.093]

the baseline odds, and θ quantifies the strength of the disease-genotype relationship and we restrict our study to 8 disease models that cover the wide scope of possible epistatic models (see Section S4 for more details).[62]

Following the procedure detailed in Section 2.3, we investigate power on (1) traditional epistatic disease models by considering Recessive-Recessive, Dominant-Dominant, and Recessive-Dominant models and (2) other observed models by studying Interaction-Multiplicative, Interface, Threshold, Modifying Effect and XOR models introduced in Reference 58. The power curves, displayed on Figure 4, first show that our omnibus by resampling test has high power in all of the considered disease models. Although our method is rarely the most powerful, it is always very close to the best method, thus demonstrating its ability to cleverly combine individual global tests. For Recessive-Recessive and Threshold models, our omnibus by resampling test can even be considered as the best method together with either L^2_c, minP_c and HC_c or L^2_d, minP_d and HC_d. It can be remarked that the Cauchy-combination omnibus test shows very similar power curves as the resampling omnibus test but with slightly less power in all situations. Furthermore, the promising performances displayed by the Cauchy-combination test has to be taken with caution since it does not properly control type I error.

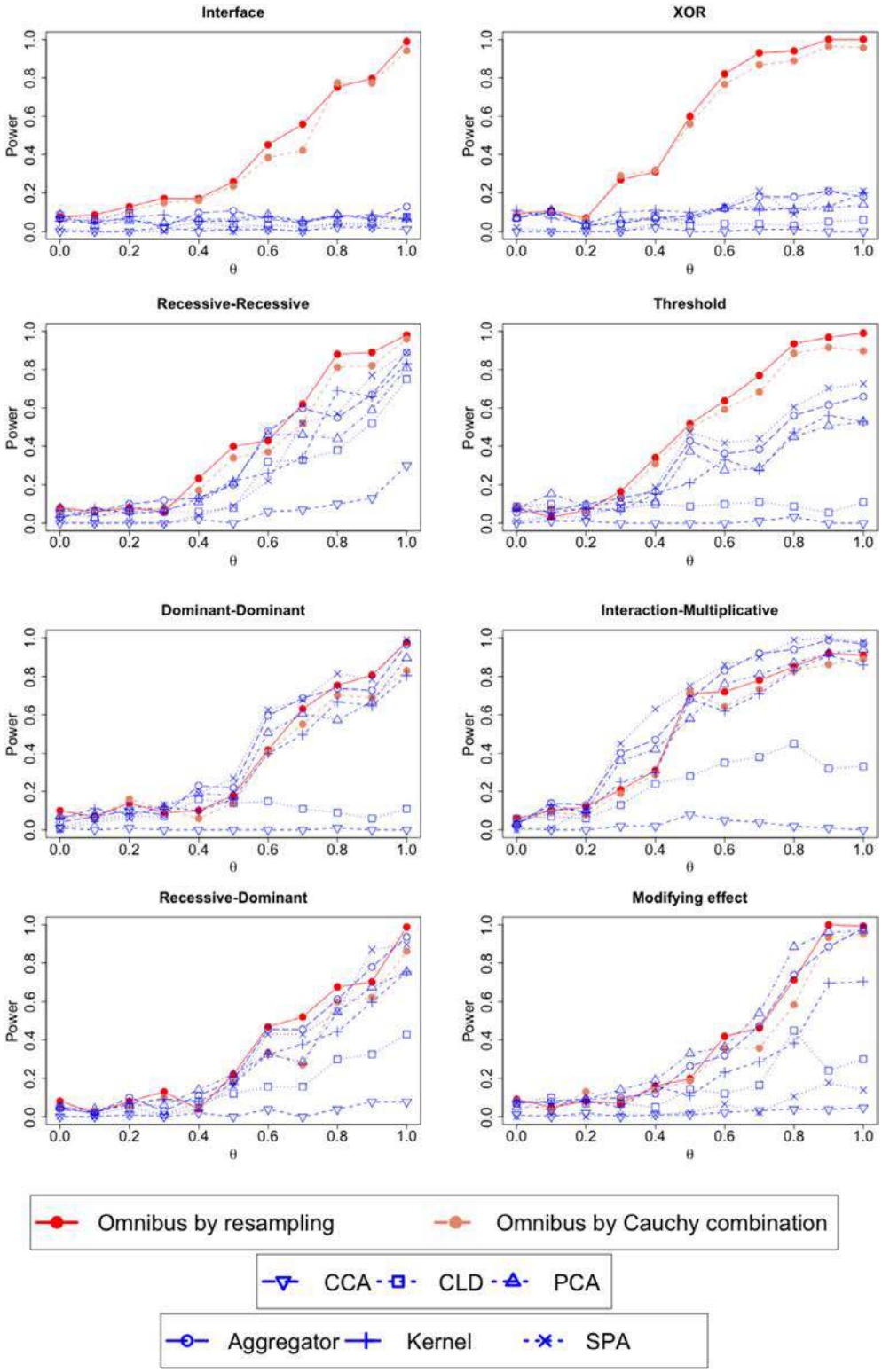

FIGURE 4 Power curves of two omnibus approaches compared with single global tests for 8 disease models involving only one causal pair (see Section S4 for details). Each disease model provides a $\boldsymbol{\beta}$ vector and θ is multiplicative coefficient applied to $\boldsymbol{\beta}$ to quantify the strength of the interaction (see Section 2.3 for details). Powers have been estimated with 1000 replicates of two genes (with four and nine SNPs, respectively) and sample sizes of 1000 cases and 1000 controls

TABLE 2 Empirical type I error rates considering either recessive or dominant main effect for three different nominal levels (0.001, 0.01, and 0.05). Methods named Omnibus and Cauchy aim at combining eight global tests derived from logistic multiple regression model while CCA, CLD, PCA, Aggregator, Kernel, and SPA are single global tests. Type I error rates have been estimated with 100 000 replicates of two genes (with four and nine SNPs, respectively). The sample sizes are 1000 cases and 1000 controls. Numbers in italic refer to targeted nominal level. Numbers in bold correspond to situations with a correct control of the type I error. Confidence intervals, in square brackets, are based on normal approximation to the binomial proportion

		CCA	CLD	PCA	Aggregator	Kernel	SPA
No main effect	0.001	<1e-04	**0.0010**	**0.0011**	0.0019	0.0005	0.0022
		[0,<1e-04]	[0.0008,0.0012]	[0.0009,0.0014]	[0.0017,0.0023]	[0.0004,0.0007]	[0.0019,0.0025]
	0.01	< 1e-03	**0.009**	**0.012**	**0.012**	**0.007**	**0.011**
		[<1e-03,0.001]	[0.008,0.011]	[0.010,0.014]	[0.010,0.014]	[0.006,0.010]	[0.009,0.013]
	0.05	0.006	**0.046**	**0.053**	**0.054**	**0.047**	**0.053**
		[0.004,0.008]	[0.042,0.051]	[0.048,0.057]	[0.049,0.058]	[0.043,0.051]	[0.048,0.057]
Additive main effect	0.001	<1e-04	0.0024	0.0016	0.0023	**0.0009**	0.0018
		[0,<1e-04]	[0.0022,0.0028]	[0.0013,0.0019]	[0.0020,0.0026]	[0.0007,0.0011]	[0.0015,0.0020]
	0.01	0.005	0.014	**0.012**	**0.010**	**0.012**	**0.011**
		[0.004,0.007]	[0.012,0.017]	[0.010,0.014]	[0.008,0.012]	[0.010,0.014]	[0.009,0.014]
	0.05	0.027	0.067	0.058	**0.050**	**0.049**	**0.053**
		[0.024,0.030]	[0.062,0.072]	[0.054,0.063]	[0.047,0.056]	[0.044,0.054]	[0.049,0.057]
Recessive main effect	0.001	< 1e-04	0.0016	0.0018	0.0024	**0.0010**	0.0026
		[0,0.0002]	[0.0013,0.0019]	[0.0016,0.0021]	[0.0021,0.0027]	[0.0007,0.0011]	[0.0023,0.0030]
	0.01	< 0.001	**0.011**	**0.012**	**0.012**	**0.011**	**0.012**
		[0,0.001]	[0.010,0.013]	[0.010,0.014]	[0.010,0.014]	[0.009,0.012]	[0.010,0.014]
	0.05	0.008	**0.052**	**0.051**	0.059	**0.053**	**0.054**
		[0.007,0.010]	[0.049,0.055]	[0.048,0.055]	[0.05648,0.06309]	[0.050,0.056]	[0.050,0.058]
Dominant main effect	0.001	< 1e-04	0.0016	0.0029	0.0045	**0.0011**	0.0052
		[0,< 1e-04]	[0.0014,0.0019]	[0.0026,0.0031]	[0.0041,0.0049]	[0.0009,0.0014]	[0.0047,0.0056]
	0.01	0.001	**0.010**	0.015	**0.009**	**0.010**	0.020
		[<1e-03,0.002]	[0.008,0.012]	[0.013,0.018]	[0.007,0.012]	[0.008,0.012]	[0.018,0.023]
	0.05	0.009	**0.050**	0.073	**0.050**	**0.054**	0.081
		[0.007,0.011]	[0.046,0.055]	[0.068,0.079]	[0.045,0.056]	[0.049,0.058]	[0.076,0.087]

Results of Figures 4 and Figure S2 illustrate that the power of each individual global test highly depends on the interaction effect, thus confirming our preliminary results by demonstrating the lack of robustness of all individual global tests. Taken separately, each global test indeed is powerless in at least one disease-model situation. In particular, when the interaction effect has a strong nonlinear trend in the logit scale (Interface and XOR), the tests based on the genotype-based coding are much more powerful than those based on the allele-based coding. Conversely, tests based on allele-based coding are more powerful when the trend of the effect is mainly linear (Dominant-Dominant, Interaction-Multiplicative, and Recessive-Dominant). For other models (Recessive-Recessive and Threshold), it can be remarked that genotype-based and allele-based Hotelling methods lack in power and that the ranking of the other methods is changed from one disease model to another. We can then consider that four groups of global tests are emerging: (1) L^2_c, minP_*c* and HC_*c*, (2) L^2_d, minP_*d* and HC_*d*, (3) Hotelling_*c*, and (4) Hotelling_*d*.

PCA, CCA, CLD, and Aggregator also show a lack of robustness as having low power for some disease models. Although for models like Dominant-Dominant, Interaction-Multiplicative, Recessive-Dominant and Modifying Effect, PCA and Aggregator are very competitive and among the best methods, our results show that PCA, CCA, CLD, and Aggregator are all powerless for other models (Interface and XOR for example). It can also be remarked that the performance

of both PCA and Aggregator are closely related to the group of allele-based global tests L^2_c, minP_c and HC_c. However, the power for CCA and CLD are globally moderate, thus indicating that these methods are not appropriate for disease models with one causal pair.

PCA, CCA, CLD, Aggregator, Kernel and SPA also show a lack of robustness as having low power for some disease models. Although for models like Dominant-Dominant, Interaction-Multiplicative, Recessive-Dominant and Modifying Effect, PCA, Aggregator, Kernel and SPA are very competitive and among the best methods, our results show that PCA, CCA, CLD, Aggregator, Kernel, and SPA are all powerless for other models (Interface and XOR for example). It can also be remarked that the performance of both PCA, Aggregator and Kernel are closely related to the group of allele-based global tests L^2_c, minP_c, and HC_c. SPA method behaves as L^2_c. However, the power for CCA and CLD are globally moderate, thus indicating that these methods are not appropriate for disease models with one causal pair.

In summary, our results demonstrate that our proposed omnibus method is the only method able to detect an interaction signal in all simulated situations. It provides evidence that our method, by combining several individual global tests is very robust to different types of disease models (with linear and non-linear trends). Similar results obtained for another gene pair (see Section S5.2) confirmed that our omnibus by resampling test is also very robust to different interplay between the SNP-sets correlation structures and the positioning of the signal along the genome.

4.2.2 | Data-driven simulation with complex disease models

We now compare our proposed omnibus test to the other methods using scenarios involving several causal SNP pairs in complex disease models. For that purpose, we consider several scenarios for the vector $\boldsymbol{\beta}$ and estimate power curves by multiplying each $\boldsymbol{\beta}$ by a scalar θ in the range $[0, 1]$, as described in Section 2.3. We thus simulate various strength of interaction signal going from $\theta = 0$, which means no effect of the interaction, to $\theta = 1$, that corresponds to a high interaction effect.

In Figures 5 and S3, power curves are reported for four different vectors $\boldsymbol{\beta}$ characterizing various situations (SCa, SCb, SCc, and SCd) regarding the interplay between signal and correlation structures. For each scenario, $\boldsymbol{\beta}$ is displayed on top of the subgraph along with the corresponding power curve with respect to the scalar θ. It can be remarked that the dimension of $\boldsymbol{\beta}$ is 144 since SNP-sets have four and nine SNPs, respectively, and that each SNP pair has four interaction coefficients in Equation (1).

Our results first confirmed the robustness of our omnibus procedure to adapt to different scenarios of association. Our omnibus by resampling test is always among the most powerful tests and even slightly outperforms the other tests for scenario SCa. It is noteworthy that the resampling omnibus test slightly outperform the Cauchy-combination omnibus test. Moreover the good performances of the Cauchy-combination test have to be contrasted by its lack of control of type I error rate. Furthermore, Figure 5 also comforts the fact that global tests can be divided into the four following groups of similarity (L^2_c, minP_c, HC_c), (L^2_d, minP_d, HC_d), Hotelling_c, and Hotelling_d. In scenario SCa, Hotelling_d has good power and outperforms the two groups (L^2_d, minP_d, HC_d) and Hotelling_c while (L^2_c, minP_c, HC_c) lacks in power. The results for scenario SCc are similar to SCa except that Hotelling_d and (L^2_d, minP_d, HC_d) are closer. On the other hand, for scenario SCb and SCd, only Hotelling_d is powerful and all other methods hardly detect signals. It can be remarked that the group (L^2_c, minP_c, HC_c) has a very low power in the four scenarios which can be explained by the complexity and the nonlinearity of the signals. Finally, results in Figure 5 also confirm that PCA, Aggregator, Kernel and SPA are very similar to the group of allele-based coding global tests (L^2_c, minP_c, HC_c). However, in contrary to results observed for disease models with one causal pair, CLD turns out to be powerful in all scenarios, especially for SCb and SCd. CCA was also revealed to be as powerful as CLD for scenarios SCb and SCd and to a lesser extent in SCa but completely fails at detecting interaction effect for SCc.

4.3 | Application to the WTCCC datasets

In this section, we investigate the performance of our omnibus by resampling test and alternatives by considering observed genotypes and phenotypes. To this end, each test have been applied to the Wellcome Trust Case Consortium Data Set,[52]

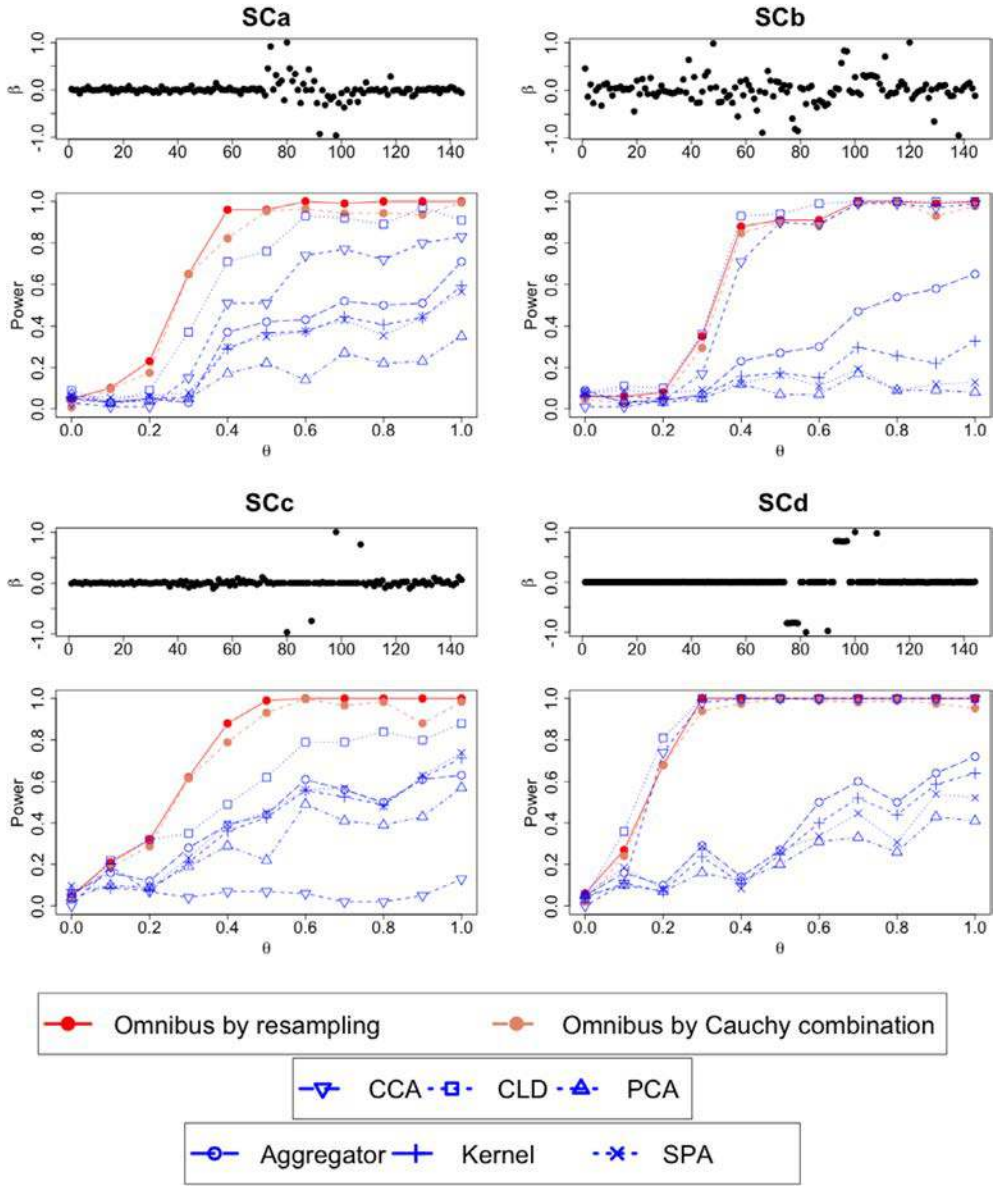

FIGURE 5 Power curves of two omnibus approaches compared with single global tests under four complex scenario of phenotype-genotypes association. The corresponding $\boldsymbol{\beta}$ vector is displayed on top of the four sets of power curves. θ acts as a multiplicative coefficient that weights the the level of phenotype-genotype association. See Section 2.3 for details regarding the simulation procedure given a vector $\boldsymbol{\beta}$. Powers have been estimated with 1000 replicates of two genes (with four and nine SNPs, respectively) and sample sizes of 1000 cases and 1000 controls. For each scenario, the upper graphic shows the value of the regression coefficients ordered according to the definition of the interaction profile vector $\boldsymbol{s}$ defined in Section 2.1 for a genotype-based profile

composed of 3000 controls and 14 000 diseased individuals divided into seven complex diseases: 2000 individuals for each of Bipolar Disorder (BD), Coronary Artery Disease (CAD), Crohn's Disease (CD), Hypertension (HT), Rheumatoid Arthritis (RA), Type-1 Diabetes (T1D), and Type-2 Diabtetes (T2D). We first investigate the control of the Type I error by testing the association of random gene pairs with a permuted phenotype. Results, displayed in Section S6, confirm the correct control of the Type I error rate by our omnibus strategy. In order to evaluate the power of statistical tests to correctly detect validated gene-gene interactions, we restrict our study to pairs of genes previously reported in the literature as interacting in susceptibility with diseases. We thus focus on a total of 25 gene pairs: 12 gene pairs associated with BD,[63,64] three pairs with CAD,[65,66] three pairs with CD,[67-69] six pairs with HT,[70,71] and one pair with RA.[72] For a given gene, the set of cases is the individuals affected by the targeted disease and the set of controls is the shared controls. It is noteworthy that these 25 signals have been detected with datasets independent from the WTCCC

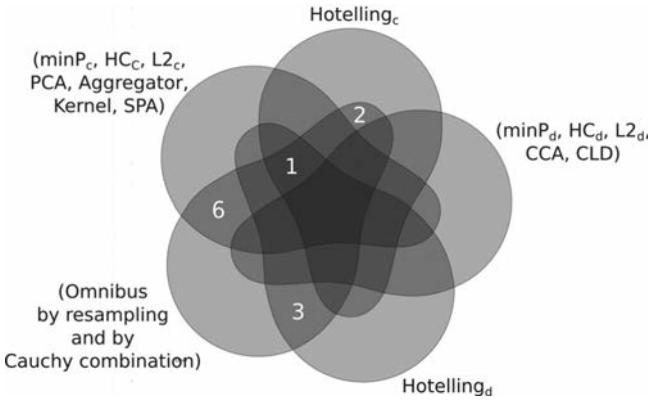

FIGURE 6 Venn diagram of the gene-gene interactions significantly detected by each method on truly observed phenotype and genotype data

dataset. Therefore, our study corresponds to a confirmatory or replication analysis and we do not expect reaching any genome-wide significant level but rather a nominal level of 5% after correction for multiple testing. By assuming that gene pairs are independent we thus applied a Benjamini-Hochberg procedure with an FDR control for multiple testing to the raw *P*-values.

Results are summarized in Table 3, where corrected *P*-values for each method and each gene pair have been reported. It can first be remarked that among the 25 gene pairs, 13 have not been detected as being significantly associated with the corresponding disease at the nominal level of 5% (and even 10%). Among the 12 other gene pairs, our omnibus procedure either with the resampling significant testing or with the Cauchy-combination is the only method robust enough to detect all the interaction signals at a level of 5%. All other methods fail at detecting at least one of these signals.

It can be remarked that the 12 detected interaction signals also show that global tests can be classified into four groups. In agreement with what we observed in our simulation study, Table 3 shows that L^2_c, minP_*c*, and HC_*c* are closely related as well as L^2_d, minP_*d* and HC_*d* while Hotelling_*c* and Hotelling_*d* have singular behaviors. More precisely, six gene pairs (*GABRA4-AFG3L2*, *DCP1A-ATP8A2*, *PTGER4-ATG16L1*, *GABRB1-MANIA1*, *SLIT2-ADRM1*, and *NOD2-NKD1*) are detected only by the group (L^2_c, minP_*c*, HC_*c*) among global tests. Furthermore, three gene pairs (*TNS3-PCDH15*, *OR6B1-ASTN2*, and *LMO3-SNRPN*) are only detected by Hotelling_*d* while two gene pairs (*TAOK2-MIR499B*, *DUOX2-CSF2RB*) are only caught by Hotelling_*c*. The last gene pair (*BANK1-BLK*), is significantly detected by the three groups of global tests (L^2_c, minP_*c*, HC_*c*), Hotelling_*c* and Hotelling_*d*. It is noteworthy that the group (L^2_d, minP_*d*, HC_*d*) does not have power to detect any interacting gene pair.

Finally, the pattern of *P*-values obtained for CCA, CLD, PCA, and Aggregator are similar to those given by our simulation study. First, CCA and CLD lack power in all situations while PCA, Aggregator, Kernel, and SPA depict very similar behaviors as the group of global tests (L^2_c, minP_*c*, HC_*c*) by jointly identifying seven gene pairs.

To sum up, our omnibus by resampling test was the most powerful method to replicate signals in the WTCCC dataset. As confirmed by the Venn diagram presented in Figure 6, our omnibus by resampling test efficiently combines individual global tests to improve the robustness of the global testing approach. Indeed, our results demonstrate that it is able to significantly detect various types of interacting signals.

5 | DISCUSSION/CONCLUSION

In this work, testing gene-based gene-gene interaction in replication studies associated with case-control genome-wide association studies is addressed by a global testing approach. Based on a logistic linear regression model, the statistical framework used in this paper allows to consider allele-based coding and genotype-based coding for SNPs. Furthermore, covariates can directly be introduced in the model so that interaction effects are adjusted for potential confounding

TABLE 3 Significance of the 25 gene pairs known to be associated with one of the five diseases : Bipolar Disorder (BD), Coronary Artery Disease (CAD), Crohn's Disease (CD), Hypertension (HT), Rheumatoid Arthritis (RA). For each gene pair, 2000 cases and 3000 controls (corresponding to the set of the shared controls in the WTCCC dataset) have been tested based on 10 000 resampling replicates (c) stands for the allele-based coding, (d) for the genotype-based coding. Each number corresponds to the *p*-value, obtained by applying the Benjamini-Hochberg procedure with FDR control for multiple testing, and in bold if lower than 0.05. For each gene, the number of SNPs is given in parenthesis

Disease	Gene1 (nb SNPs)	Gene2 (nb SNPs)	Omni by resampling	Omni by Cauchy combination	L^2_c	minP_*c*	HC_*c*	Hotelling_*c*	L^2_d	minP_*d*	HC_*d*	Hotelling_*d*	CCA	CLD	PCA	Aggr.	Kernel	SPA
BD	*DPF3* (28)	*PTPRS* (22)	0.692	0.692	0.713	0.668	0.743	0.726	0.689	0.824	0.667	0.653	0.770	0.734	0.587	0.340	0.713	0.710
BD	*SH3PXD2A* (18)	*ABCC11* (31)	0.354	0.354	0.713	0.805	0.743	0.565	0.562	0.373	0.423	0.423	0.930	0.589	0.399	0.649	0.713	0.705
BD	*SLIT2* (51)	*ADRM1* (6)	**0.023**	**0.023**	**0.017**	**0.015**	**0.022**	0.077	0.579	0.373	0.423	0.466	0.930	0.103	**0.002**	**0.001**	**0.017**	**0.017**
BD	*TNS3* (37)	*PCDH15* (51)	**0.023**	**0.023**	0.857	0.805	0.924	0.110	0.716	0.891	0.667	**0.025**	0.851	0.320	0.886	0.831	0.857	0.805
BD	*RSPO3* (17)	*ASCC2* (8)	0.726	0.726	0.713	0.736	0.743	0.742	0.716	0.891	0.667	0.653	0.930	0.785	0.513	0.448	0.736	0.713
BD	*USP15* (11)	*NDUFA3* (6)	0.612	0.612	0.713	0.742	0.743	0.448	0.707	0.824	0.667	0.496	0.930	0.440	0.199	0.458	0.742	0.713
BD	*GABRA4* (19)	*AFG3L2* (12)	**0.023**	**0.023**	**0.025**	0.068	0.059	0.143	0.579	0.373	0.423	0.284	0.621	0.105	**0.005**	**0.015**	0.068	**0.025**
BD	*MMP27* (11)	*RFX4* (32)	0.996	0.996	0.857	0.805	0.924	0.714	0.998	0.998	0.998	0.899	0.621	0.780	0.906	0.827	0.857	0.821
BD	*TKT* (12)	*NPC1* (17)	0.726	0.726	0.713	0.474	0.569	0.547	0.746	0.908	0.667	0.657	0.621	0.496	0.273	0.177	0.569	0.713
BD	*TAOK2* (8)	*MIR499B* (8)	**0.023**	**0.023**	0.471	0.474	0.464	**0.020**	0.856	0.891	0.907	0.466	0.851	0.411	0.193	0.190	0.470	0.471
BD	*OR6B1* (7)	*ASTN2* (254)	**0.023**	**0.023**	0.627	0.480	0.660	0.555	0.562	0.824	0.423	**0.045**	0.851	0.475	0.462	0.176	0.612	0.621
BD	*DCP1A* (12)	*ATP8A2* (90)	**0.023**	**0.023**	**0.017**	0.077	**0.022**	0.278	0.562	0.373	0.423	0.653	0.851	0.105	**0.008**	**0.015**	**0.022**	**0.017**
CAD	*ANRIL* (32)	*TMEM106B* (9)	0.726	0.726	0.713	0.805	0.789	0.742	0.612	0.824	0.610	0.653	0.851	0.780	0.674	0.676	0.789	0.713
CAD	*ADTRP* (18)	*MIA3* (9)	0.726	0.726	0.713	0.805	0.789	0.742	0.612	0.824	0.610	0.653	0.851	0.780	0.674	0.676	0.789	0.713
CAD	*CYP4A11* (10)	*CYP4F2* (9)	0.996	0.996	0.840	0.805	0.817	0.920	0.998	0.998	0.998	0.899	0.851	0.866	0.751	0.733	0.817	0.845
CD	*NOD2* (12)	*NKD1* (27)	**0.027**	**0.027**	**0.025**	**0.040**	**0.048**	0.078	0.612	0.738	0.610	0.284	0.761	0.105	**0.009**	**0.007**	**0.040**	**0.025**
CD	*PTGER4* (10)	*ATG16L1* (29)	**0.023**	**0.023**	**0.025**	0.083	**0.048**	0.448	0.612	0.824	0.610	0.653	0.930	0.475	**0.010**	**0.020**	**0.048**	**0.025**
CD	*DUOX2* (9)	*CSF2RB* (7)	**0.023**	**0.023**	0.132	0.221	0.187	**0.042**	0.685	0.529	0.631	0.193	0.930	0.103	0.091	0.073	0.182	0.132
HT	*GABRB1* (64)	*MAN1A1* (18)	**0.023**	**0.023**	0.089	**0.040**	0.059	0.547	0.562	0.373	0.423	0.496	0.621	0.589	**0.024**	**0.005**	**0.048**	**0.048**
HT	*LMO3* (10)	*NPAP1* (7)	0.726	0.726	0.713	0.605	0.743	0.565	0.746	0.998	0.755	0.665	0.621	0.589	0.379	0.289	0.713	0.713
HT	*LMO3* (10)	*SNRPN* (25)	**0.023**	**0.023**	0.840	0.805	0.844	0.093	0.562	0.373	0.423	**0.025**	0.930	0.257	0.729	0.666	0.840	0.844
HT	*CHRNA7* (19)	*CDH13* (109)	0.527	0.527	0.349	0.661	0.385	0.411	0.716	0.824	0.667	0.653	0.770	0.440	0.155	0.331	0.385	0.349
HT	*KRT8P5* (8)	*DNAL4* (4)	0.726	0.726	0.840	0.805	0.817	0.852	0.746	0.824	0.667	0.466	0.851	0.793	0.813	0.771	0.817	0.840
HT	*SCGB1A1* (3)	*LPL* (21)	0.726	0.726	0.840	0.805	0.817	0.555	0.716	0.891	0.667	0.695	0.621	0.589	0.708	0.727	0.817	0.840
RA	*BANK1* (121)	*BLK* (15)	**0.023**	**0.023**	**0.023**	**0.030**	**0.022**	**0.020**	0.562	0.603	0.423	**0.028**	0.621	0.103	**0.003**	**0.003**	**0.023**	**0.020**

factors. The issue of the choice of the most appropriate global testing method is tackled by focusing on four usual procedures: minP, HC, L^2-norm, and Hotelling. Compared to the other methods, Hotelling requires the inversion of the correlation matrix of pointwise statistics. A specific model, that relies on a Kronecker decomposition, is proposed to obtain the eigenvalue decomposition or the inverse of this correlation matrix, thus reducing the computational cost while improving stability of the estimation. Based on a comparative study, we first show that none of global testing methods are uniformly powerful. These methods are rather heterogeneous and complementary over the range of interaction models, suggesting that combining individual methods is likely to improve the robustness of global testing approach.

We therefore introduced an omnibus strategy that efficiently aggregates individual global tests. Through a data-driven simulation study, it is demonstrated that the use of a parametric bootstrap resampling method allows for an accurate control of the type I error rate, especially in presence of covariate and/or main effects. Extensive data-driven simulations have also been used to assess the superiority of our omnibus approach in a wide range of phenotype-genotype association models. Moreover, our results show that our global testing approach outperforms previously introduced multidimensional methods. Whether for association models with few causal pairs, where PCA has a good power, or more complex association models, where CCA and CLD can have very high detection rate, our omnibus by resampling test is always very competitive. A comparative study on truly observed datasets confirms the flexibility of our approach that allows the identification of gene-gene interaction in very diverse situations.

Although these results are very promising, combining individual methods in an omnibus strategy raises questions regarding the choice of the individual methods to be aggregated and the way such aggregation is performed. Increasing the number of global tests to be combined is tempting, but it will result in an increase of the computational cost and may tend to consider strongly correlated test statistics in the aggregation step. Furthermore, tentatives to combine different sets of methods than minP, HC, L^2-norm, and Hotelling did not provide better power but increased the computational cost since more single methods, such as Kernel and SPA, have to be tested for example. Properly handling the correlation between global tests is also a difficult task and a safe strategy would consist in a parsimonious choice of methods. The use of the Cauchy-combination method, based on a weighted sum of Cauchy transformation of individual *P*-values, is attractive since it does not require the estimation of the correlation between individual tests. The Cauchy-combination saves computational time compared to our resampling method, especially for large genes, and it shows very similar power than our resampling method. However, the Cauchy-combination interaction testing method can suffer from elevated type 1 error in the presence and absence of genetic main effects.

Nevertheless, the use of resampling-based simulation studies does not allow the validation of our method at a genome-scale. Although the genome-wide significance criteria for genome-wide gene-gene interaction is not established in the literature, it is likely that it requires more than 100 000 replicates. Our results on the control of type I error are therefore to be considered for replication studies where a nominal level of 10^{-3} is acceptable. To generalize our conclusions at the genome-scale, a careful attention should also be paid to the multiple testing issue. As pointed by Reference 7, designing a proper multiple testing correction to reach genome-wide nominal level remains very tricky for epistasis. In addition to widely used adjusted method such as Bonferroni (see BOOST for example[8]) of Benjamini-Hochberg, other methods have been develop to guarantee the control of the type-I error rate in gene-gene interaction. For example, the recent years has seen the application of permutation-based rather than correction based on a theoretical reference distribution.[19] However, in addition to the computational burden, such methods may not be appropriate in situations of high LD and rare variants.[73] Multistage epistasis strategies have also been proposed to reduce the multiple testing problem by considering multiple simple models rather than full interaction models.[74] However, to better account for the complex dependencies between individual tests, the development of more refined methods is still needed.[75]

Although our study is based on simulations that cover a large spectrum gene-gene interaction models, it is impossible to explore all realistic situations and our results are therefore limited to the set of simulated scenarios. At first, truly observed data may lead to situations where the SNPs from a SNP set are highly correlated thus facing the issue of multicollinearity when estimating regression parameters. In our framework, this issue has been addressed by using a principal component analysis. The principal component decomposition has the advantage of having a low computational cost compared to regularization techniques where the regularized coefficient is often estimated by resampling techniques. A threshold is applied on the amount of variability to retain the top principal component and such threshold can also be seen as an hyperparameter of our method. We based the tuning of the threshold on its capacity to control the type I error

rate and, similar to other hyperparameter tuning, it may depend on the training dataset. Secondly, genomic regions of interest might contain a large number of SNPs. The number of individual statistics in gene-gene interaction can therefore be very large thus facing the curse of dimensionality. Our simulations does not consider such high-dimensional situations and should be extended to generalize our results.

Our results on the WTCCC datasets are encouraging in the perspective of a robust replication of candidate gene-gene interaction. However, the computational cost associated with any gene-gene interaction remains an issue for our method to be performed at the genome scale. It should be noticed that computing global tests is decomposed into three main steps: (1) the computation of the vector of pointwise statistics $\mathcal{Z}$, (2) the aggregation into a global statistic $T(\mathcal{Z})$ and (3) the computation of the significance level. For all methods, step (3) is the most expensive since it is based on a resampling parametric bootstrap. Furthermore, since step (1) is shared by all methods and Hotelling, due to the inversion of a correlation matrix, is more consuming in step (2), our omnibus by resampling test is not more time-consuming than the Hotelling method. Therefore, to help performing genome-wide scans, efforts should be put to improve programming efficiency by considering for instance parallel or GPU implementations. More precisely, the computational cost associated with our procedure is linear with the number of tests, the number bootstrap sample and the number of individuals. Although the current implementation of our method, currently available in R package `GeneGeneInteractions` available at https://github.com/fhebert/ does not provide a parallelized version of the functions, parallelization can be performed by computing the statistics from each bootstrap sample in parallel. However, since our method relies on the covariance matrix of the SNP set, parallelization is not feasible on the individuals and the SNP pairs.

Finally, although our omnibus method has been evaluated only on case-control genome-wide association studies, the proposed framework allows for considering any type of generalized linear models. Our study therefore opens perspectives into the detection of gene-gene interaction in association with quantitative traits by extending the results of Reference 76. The scope of the proposed framework is also not limited to gene-gene interaction in genome-wide association studies. As being based on a very general model and on general methods for combining statistics, the proposed framework can be adapted to other applicative situations by considering appropriate definition for the corresponding pointwise statistics. Our work can therefore be useful to address the challenging question of detecting interactions associated with a phenotype, either binary, categorical or continuous, raised for example in precision medicine.[77,78] The use of global testing approaches can indeed help handling with the diversity of interacting signals.

CONFLICT OF INTEREST

The authors declare no potential conflict of interests.

DATA AVAILABILITY STATEMENT

The proposed testing procedure is freely available in a package for the open source R software.[79] Our package, named `GeneGeneInteractions`, is available at https://github.com/fhebert/.

ORCID

Mathieu Emily https://orcid.org/0000-0001-9101-6689

REFERENCES

1. Buniello A, MacArthur JAL, Cerezo M, et al. The NHGRI-EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucl Acids Res*. 2018;47(D1):D1005-D1012. doi:10.1093/nar/gky1120
2. Maher B. Personal genomes: the case of the missing heritability. *Nature*. 2008;456:18-21.
3. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461:747-753.
4. Moore JH. The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Human Heredity*. 2003;56:73-82.
5. Phillips P. Epistasis, the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Rev Genet*. 2008;9:855-867.
6. Zuk O, Hechter E, Sunyaev SR, Lander ES. The mystery of missing heritability: genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*. 2012;109(4):1193-1198.
7. Ritchie MD, Steen KV. The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. *Ann Transl Med*. 2018;6(8):1-14.
8. Wan X, Yang C, Yang Q, et al. BOOST: a fast approach to detecting gene-gene interactions in genome-wide case-control studies. *Am J Human Genet*. 2010;87:325-340.

9. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecul Genet*. 2002;11(20):2463-2468.
10. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet*. 2005;37(4):413-417.
11. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a toolset for whole-genome association and population-based linkage analysis. *Am J Human Genet*. 2007;81:559-575.
12. Emily M. IndOR: a new statistical procedure to test for SNP x SNP epistasis in genome-wide association studies. *Stat Med*. 2012;31(21):2359-2373.
13. Zhao J, Jin L, Xiong M. Test for interaction between two unlinked loci. *Am J Human Genet*. 2006;79(5):831-845.
14. Wu X, Dong H, Luo L, et al. A novel statistic for genome-wide interaction analysis. *PLoS Genet*. 2010;6(9):e1001131.
15. Ueki M, Cordell HJ. Improved statistics for genome-wide interaction analysis. *PLoS Genet*. 2012;8(4):e1002625. doi:10.1371/journal.pgen.1002625
16. Dong C, Chu X, Wang Y, et al. Exploration of gene-gene interaction effects using entropy-based methods. *Eur J Human Genet*. 2008;16(2):229-235.
17. Kang G, Yue W, Zhang J, Cui Y, Zuo Y, Zhang D. An entropy-based approach for testing genetic epistasis underlying complex diseases. *J Theor Biol*. 2008;250(2):362-374.
18. Zhang Y, Liu JS. Bayesian inference of epistatic interactions in case-control studies. *Nature Genet*. 2007;39:1167-1173.
19. Ritchie MD, Hahn LW, Roodi N, et al. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *Am J Human Genet*. 2001;69(1):138-147.
20. Moore J, White B. Tuning reliefF for genome-wide genetic analysis. *Lect Notes Comput Sci*. 2007;4447:166-175.
21. Schwarz D, Konig I, Ziegler A. On safari to random jungle: a fast implementation of random forests for high dimensional data. *Bioinformatics*. 2010;26:1752-1758.
22. Shang J, Zhang J, Sun Y, Liu D, Ye D, Yin Y. Performance analysis of novel methods for detecting epistasis. *BMC Bioinform*. 2011;12(1):475. doi:10.1186/1471-2105-12-475
23. Emily M. A survey of statistical methods for gene-gene interaction in case-control genome-wide association studies. *J de la Société Française de Statistique*. 2018;159(1):27-67.
24. Neale BM, Sham PC. The future of association studies: gene-based analysis and replication. *Am J Human Genet*. 2004;75(3):353-362.
25. Jorgenson E, Witte JS. A gene-centric approach to genome-wide association studies. *Nature Rev Genet*. 2006;7(11):885-891.
26. Huang H, Chanda P, Alonso A, Bader JS, Arking DE. Gene-based tests of association. *PLoS Genet*. 2011;7(7):e1002177. doi:10.1371/journal.pgen.1002177
27. Wu MC, Kraft P, Epstein MP, et al. Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Human Genet*. 2010;86(6):929-942.
28. Li J, Tang R, Biernacka J, de Andrade M. Identification of gene-gene interaction using principal components. *BMC Proc*. 2009;3(Suppl 7):S78.
29. Peng Q, Zhao J, Xue F. A gene-based method for detecting gene-gene co-association in a case-control association study. *Eur J Human Genet*. 2010;18(5):582-587.
30. Yuan Z, Gao Q, He Y, et al. Detection for gene-gene co-association via kernel canonical correlation analysis. *BMC Genet*. 2012;13(1):83. doi:10.1186/1471-2156-13-83
31. Larson NB, Jenkins GD, Larson MC, et al. Kernel canonical correlation analysis for assessing gene-gene interactions and application to ovarian cancer. *Eur J Human Genet*. 2014;22(1):126-131.
32. Zhang X, Yang X, Yuan Z, et al. A PLSPM-based test statistic for detecting gene-gene co-association in genome-wide association study with case-control design. *PLoS One*. 2013;8(4):e62129. doi:10.1371/journal.pone.0062129
33. Li S, Cui Y. Gene-centric gene-gene interaction: a model-based kernel machine method. *Ann Appl Stat*. 2012;6(3):1134-1161.
34. Larson NB, Schaid DJ. A kernel regression approach to gene-gene interaction detection for case-control studies. *Genet Epidemiol*. 2013;37(17):695-703.
35. Rajapakse I, Perlman MD, Martin PJ, Hansen JA, Kooperberg C. Multivariate detection of gene-gene interactions. *Genet Epidemiol*. 2012;36(6):622-630.
36. Li J, Huang D, Guo M, et al. A gene-based information gain method for detecting gene-gene interactions in case-control studies. *Eur J Human Genet*. 2015;23:1566-1572.
37. Emily M. AGGrEGATOr: a gene-based GEne-Gene interAcTiOn test for case-control association studies. *Stat Appl Genet Mol Biol*. 2016;15(2):151-171.
38. Causeur D, Sheu CF, Perthame E, Rufini F. A functional generalized F-test for signal detection with applications to event-related potentials significance analysis. *Biometrics*. 2020;76(1):246-256.
39. Arias-Castro E, Candès EJ, Plan Y. Global testing under sparse alternatives: ANOVA, multiple comparisons and the higher criticism. *Ann Stat*. 2011;39(5):2533-2556.
40. Conneely KN, Boehnke M. So many correlated tests, so little time! rapid adjustment of P values for multiple correlated tests. *Am J Human Genet*. 2007;81(6):1158-1168.
41. Liu JZ, Mcrae AF, Nyholt DR, et al. A versatile gene-based test for genome-wide association studies. *Am J Human Genet*. 2010;87(1):139-145.
42. Donoho D, Jin J. Higher criticism for detecting sparse heterogeneous mixtures. *Ann Stat*. 2004;32(3):962-994.

43. Derkach A, Lawless JF, Sun L. Pooled association tests for rare genetic variants: a review and some new results. *Stat Sci*. 2014;29(2):302-321.
44. Hebert F, Causeur D, Emily M. An adaptive decorrelation procedure for signal detection. *Comput Stat Data Anal*. 2021;153:107082.
45. Liu Y, Xie J. Cauchy combination test: a powerful test with analytic P-value calculation under arbitrary dependency structures. *J Am Stat Assoc*. 2020;115(529):393-402.
46. Buzkova P, Lumley T, Rice K. Permutation and parametric bootstrap tests for gene-gene and gene-environment interactions. *Ann Human Genet*. 2016;75(1):36-45.
47. Wu Z, Sun Y, He S, et al. Detection boundary and higher criticism approach for rare and weak genetic effects. *Ann Appl Stat*. 2014;8(2):824-851.
48. Barnett I, Mukherjee R, Lin X. The generalized higher criticism for testing SNP-set effects in genetic association studies. *J Am Stat Assoc*. 2017;112(517):64-76.
49. Lin X, Lee S, Christiani DC, Lin X. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics*. 2013;14(4):667-681.
50. Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M. Genome-wide gene and pathway analysis. *Eur J Human Genet*. 2010;18(9):1045.
51. Westfall P, Young S. *Resampling-Based Multiple Testing*. New York, NY: Wiley; 1993.
52. WTCCC. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*. 2007;447:661-678.
53. Horn RA, Johnson CR, Elsner L. *Topics in Matrix Analysis*. 1st ed. Cambridge, UK: Cambridge University Press; 1994.
54. Coombes BJ, Biernacka JM. Application of the parametric bootstrap for gene-set analysis of gene–environment interactions. *Eur J Human Genet*. 2018;26(11):1679.
55. Epstein MP, Duncan R, Jiang Y, Conneely KN, Allen AS, Satten GA. A permutation procedure to correct for confounders in case-control studies, including tests of rare variation. *Am J Human Genet*. 2012;91(2):215-223.
56. Barbiero A, Ferrari PA. GenOrd: simulation of discrete random variables with given correlation matrix and marginal Distributions. R package on CRAN; 2015. R package version 1.4.0.
57. Emily M, Sounac N, Kroell F, Houée-Bigot M. Gene-based methods to detect gene-gene interaction in R: the GeneGeneInteR package. *J Stat Softw*. 2020;95(12):1-32. doi:10.18637/jss.v095.i12
58. Hallgrimsdottir IB, Yuster DS. A complete classification of epistatic two-locus models. *BMC Genet*. 2008;9(17):1-15.
59. Li MX, Gui HS, Kwan J, Sham P. GATES: a rapid and powerful gene-based association test using extended simes procedure. *Am J Human Genet*. 2011;88(3):283-293.
60. Simes RJ. An improved Bonferroni procedure for multiple tests of significance. *Biometrika*. 1986;73(3):751-754.
61. Lin X, Lee S, Wu M, et al. Test for rare variants by environment interactions in sequencing association studies. *Biometrics*. 2016;72(1):156-164.
62. Li W, Reich J. A complete enumeration and classification of two-locus disease models. *Human Heredity*. 2000;50(6):334-349.
63. Maj C, Milanesi E, Gennarelli M, Milanesi L, Merelli I. Epistasis analysis reveals associations between gene variants and bipolar disorder. *PeerJ Preprints*. 2017;5:e3242v1.
64. Judy J, Seifuddin F, Pirooznia M, et al. Converging evidence for epistasis between ANK3 and potassium channel gene KCNQ2 in bipolar disorder. *Front Genet*. 2013;4:87.
65. Sirotina S, Ponomarenko I, Kharchenko A, et al. A novel polymorphism in the promoter of the CYP4A11 gene is associated with susceptibility to coronary artery disease. *Disease Markers*. 2018;2018:1-12.
66. Li Y, Cho H, Wang F, et al. Statistical and functional studies identify epistasis of cardiovascular risk genomic variants from genome-wide association studies. *J Am Heart Assoc*. 2020;9(7):e014146.
67. Seiderer J, Glas J, Pasciuto G, et al. First evidence for strong epistasis between two Crohn's disease susceptibility loci: PTGER4-expression-modulating polymorphisms in the 5p13.1 region enhance ATG16L1-associated susceptibility to Crohn's disease. *Z Gastroenterol*. 2008;46(1):022-022.
68. Levine AP, Pontikos N, Schiff ER, et al. Genetic complexity of Crohn's disease in two large Ashkenazi Jewish families. *Z Gastroenterol*. 2016;151:698-709.
69. Abegaz F, Van Lishout F, Mahachie John JM, et al. Epistasis detection in genome-wide screening for complex human diseases in structured populations. *Syst Med*. 2019;2(1):19-27. doi:10.1089/sysm.2019.0003
70. Ndiaye NC, Said ES, Stathopoulou MG, Siest G, Tsai MY, Visvikis-Siest S. Epistatic study reveals two genetic interactions in blood pressure regulation. *BMC Med Genet*. 2013;14:1-7.
71. Meng Y, Groth S, Quinn JR, Bisognano J, Wu TT. An exploration of gene-gene interactions and their effects on hypertension. *Int J Genom*. 2017;2017:1-9.
72. Génin E, Coustet B, Allanore Y, et al. Epistatic interaction between BANK1 and BLK in rheumatoid arthritis: results from a large trans-ethnic meta-analysis. *Plos One*. 2013;8(4):1-8.
73. Mahachie John JM, Van Lishout F, Gusareva ES, Van Steen K. A robustness study of parametric and non-parametric tests in model-based multifactor dimensionality reduction for epistasis detection. *BioData Min*. 2013;6(1):1-17.
74. Franberg M, Gertow K, Hamsten A, Consortium P, Lagergren J, Sennblad B. discovering genetic interactions in large-scale association studies by stage-wise likelihood ratio tests. *PLOS Genet*. 2015;11(9):1-24.
75. Steen KV, Moore JH. How to increase our belief in discovered statistical interactions via large-scale association studies? *Human Genet*. 2019;138:293-305.
76. Ma L, Clark AG, Keinan A. Gene-based testing of interactions in association studies of quantitative traits. *PLoS Genet*. 2013;9(2):e1003321. doi:10.1371/journal.pgen.1003321

77. Li J, Li X, Zhang S, Snyder M. Gene-environment interaction in the era of precision medicine. *Cell*. 2019;177(1):38-44.
78. de Maturana L, Alonso L, Alarcón P, et al. Challenges in the integration of omics and non-omics data. *Genes*. 2019;10(3):1-17.
79. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2013.