Ma *et al. BMC Medical Genomics* (2024) 17:111
https://doi.org/10.1186/s12920-024-01884-w

BMC Medical Genomics

RESEARCH Open Access

# Poor statistical power in population-based association study of gene interaction

Jiarui Ma[1†], Jian Li[1†], Yuqi Chen[1], Zhen Yang[2] and Yungang He[3*]

## Abstract

**Background** Statistical epistasis, or "gene–gene interaction" in genetic association studies, means the nonadditive effects between the polymorphic sites on two different genes affecting the same phenotype. In the genetic association analysis of complex traits, nevertheless, the researchers haven't found enough clues of statistical epistasis so far.

**Methods** We developed a statistical model where the statistical epistasis was presented as an extra linkage disequilibrium between the polymorphic sites of different risk genes. The power of statistical test for identifying the gene–gene interaction was calculated and then compared in different hypothesis scenarios.

**Results** Our results show the statistical power increases with the increasing of interaction coefficient, relative risk, and linkage disequilibrium with genetic markers. However, the power of interaction discovery is much lower than that of regular single-site association test. When rigorous criteria were employed in statistical tests, the identification of gene–gene interaction became a very difficult task. Since the criterion of significance was given to be $p$-value $\leq 5.0 \times 10^{-8}$, the same as that of many genome-wide association studies, there is little chance to identify the gene–gene interaction in all kind of circumstances.

**Conclusions** The lack of epistasis tends to be an inevitable result caused by the statistical principles of methods in the genetic association studies and therefore is the inherent characteristic of the research itself.

**Keywords** Genetic association study, Gene interaction, Statistical epistasis, Statistical power

## Background

Gene interaction, or epistasis, is an important concept in biochemical genetics, population genetics and quantitative genetics [1]. The specific definitions of gene interaction are different in different fields, but its basic concept is the interaction among genetic loci in their effects on phenotypes or fitness [2].

Gene interaction plays an important role in molecular genetics and quantitative genetics [1, 3]. The type of gene interaction determines the functional relationship between genes, genes and pathways, and the corresponding gene products [4, 5]. The degree and nature of gene interactions are also important for the theoretical explanation of the selective advantage of sexual reproduction [6, 7]. In addition, gene interaction has a profound effect on quantitative genetics since it plays an important role in many evolutionary processes, and may be involved in the reproductive isolation of species, the additive genetic variance caused by population bottlenecks, the evolution of trait mutation stability, adaptive gene complexes, sexual reproduction, and genetic system structures [3, 8].

[†] Jiarui Ma and Jian Li contributed equally to this work.

*Correspondence:
Yungang He
heyungang@fudan.edu.cn
[1] Shanghai Key Laboratory of Medical Epigenetics, International Co-Laboratory of Medical Epigenetics and Metabolism (Ministry of Science and Technology), Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China
[2] Center for Medical Research and Innovation of Pudong Hospital, Intelligent Medicine Institute, Fudan University, Shanghai 200032, China
[3] Shanghai Fifth People's Hospital, Intelligent Medicine Institute, Fudan University, Shanghai 200032, PR China

Ronald A. Fisher explained the gene interaction from a statistical perspective for the first time and described gene interaction as a non-additive statistical effect between genetic polymorphic sites on two different genes which affect the same biological phenotype [9]. The modern quantitative genetics proposed by Cockerham et al. further developed the definition of gene interaction and regarded it as an interaction term in the regression of allelic effects [10]. The statistical gene interaction of two genes suggests that they also have physiological gene interaction, which may be caused by direct interactions between proteins or indirect interactions of gene product interaction networks. Therefore, statistical gene interaction can provide insights into the genetic structure of complex phenotypes, which may help to improve the statistical power of genetic association analysis.

In the past ten years, genome-wide genetic association studies (GWAS) have gradually emerged, and a large number of research results on genetic markers about polygenic phenotypes have also emerged in abundance. For example, 45 lung cancer susceptibility loci have been reported with different strength of evidence, highlighting suspected causal genes at each locus [11]. However, the influence of a single genetic marker is limited and cannot fully explain the corresponding phenotypic genetic variance. Therefore, investigators tried to combine multiple genetic markers for statistical analysis, hoping to enhance the statistical power of genetic association analysis through "gene interaction", and find the source of unexplained genetic variance and more undiscovered genetic markers and achieve scientific breakthroughs [12]. The first and only genome-wide two-locus interaction analysis about lung cancer performed so far by Minjie Chu et al. revealed a significant interaction between SNPs rs2562796 and rs16832404 with a small sample size (858 cases and 1115 controls). Since small sample size may easily result in false positives, further studies with large sample size are needed to validate their findings [13]. In large-scale genetic association analysis, gene interactions with statistical meaning have not yet been found.

Some investigators attributed this problem to the limitations of linear models. The linear model plays an important role in modern genetic epidemiology because of its solid theoretical foundation and widespread usage in different software packages [14]. Despite the advantages of using linear models, they do have limitations for explaining genetic models of disease due to limited ability to detect nonlinear patterns of genetic interaction [15]. The limitations of statistical models have promoted the development of calculation methods such as machine learning and data mining. The methods reviewed by Cordell [16] include novel approaches such as combinatorial partitioning [17], logic regression [18], and machine learning approaches such as random forests [19]. Based on logistic regression model, Cordell et al. carried out comprehensive analyses for detecting gene–gene interactions and obtained generally decent power under specific scenarios [20]. These novel methods, nevertheless, have limitations in their application. For example, multifactor dimensionality reduction (MDR), a significant new method for detecting and characterizing patterns of statistical epistasis in genetic association studies, was feasible only for examining two-locus interactions in a filtered data set or for examining higher-level interactions in an even further reduced data set [16]. GWAS poses greater challenges to computational methods. As summarized by Ritchie [21] and Moore [22], combinatorial assessment of SNPs in a GWAS is not computationally feasible beyond exploring two-way and three-way combinations. Analysis tools and software that can detect statistical gene interaction quickly and accurately are lacking yet [14].

Research on the detection of gene interaction in genetic association analysis is fruitful, and there are continuous innovations in technical means, but there are also shortcomings, that is, too much attention is paid to modeling and calculation methods, and little attention is paid to the fundamental question of whether gene interaction exists in genetic association analysis. To overcome these limitations, we established an additive genetic effect model with two alleles and developed a measure of interaction between two unlinked disease alleles under the framework of LD analysis to investigate whether there exists statistical epistasis in genetic association studies. We then calculated the power of statistical testing methods to detect gene interactions in different genetic scenarios. As a reference, we conducted a regular single-site genetic association study between the marker gene and the disease under different genetic scenarios, and obtained decent statistical power with reasonable parameters. For exploring the possible solution for the interaction identification, we evaluated the statistic power in investigation of interaction between two unlinked loci under parameters as same as the aforementioned single-site genetic association study and concluded with a reasonable theoretical explanation for the phenomenon of "missing gene interaction" in genetic association analysis.

## Methods

### Case-control model

To investigate the LD pattern generated by gene–gene interaction, we assume that the disease-susceptibility loci are in Hardy–Weinberg equilibrium (HWE) in the population. Let $X_1$ and $X_2$ be the two independent disease genes located on the same chromosome. We give $P(X_1|D)$ and $P(X_2|D)$ as the probabilities of carrying the disease gene $X_1$ and $X_2$ by the affected individual, respectively. When $P(X_1, X_2|D)$ is the probability of carrying both X$_1$ and X$_2$, the LD between X$_1$ and X$_2$ in the affected individuals due to their interaction can be obtained in $D = P(X_1, X_2|D) - P(X_1|D) \cdot P(X_2|D)$.

When there is a gene–gene interaction between X$_1$ and X$_2$, the coefficient of interaction could be defined in the disease case–control study as $K = \frac{f_{11D}}{f_{1*D}f_{*1D}} - 1$, where $f_{1*D}$, $f_{*1D}$, and $f_{11D}$ is frequency of chromosomes in affected individuals that carrying X$_1$, X$_2$, and both of them, respectively. The notation '*' indicates any allele of '0' or '1' at the current site, i.e. $f_{1*D} = f_{11D} + f_{10D}$.We assume that the relative risks of X$_1$ and X$_2$ are $R_{X1}$ and $R_{X2}$, respectively. We follow the definition of relative risks in the literature [23]. $R_{X1}$ is the ratio of the incidence of individuals carrying X$_1$ in the population to the incidence of individuals without X$_1$. In the same way, we can define $R_{X2}$. The aforementioned frequencies in affected individuals are assumed as $f_{1*D} = \frac{R_{X1}f_{1*}}{R_{X1}f_{1*}+(1-f_{1*})}$, $f_{*1D} = \frac{R_{X2}f_{*1}}{R_{X2}f_{*1}+(1-f_{*1})}$, and $f_{11D}$= $(1+K) \cdot f_{1*D}f_{*1D}$, where $f_{1*}$ and $f_{*1}$ is frequency of chromosomes in the population that carrying X$_1$, X$_2$, respectively. It is clear that the interaction coefficient K indicates the extra LD between X$_1$ and X$_2$ due to their non-additive (multiplicative) risk. Similarly, we can obtain the frequencies $f_{1*d}$, $f_{*1d}$, and $f_{11d}$ for the chromosomes in unaffected individuals.

When the disease-susceptibility loci are usually undetectable, genetic association studies are employed with marker genes. In this study, we assume that the marker gene M$_1$ is in LD with X$_1$ and that the marker gene M$_2$ is in LD with X$_2$. Furthermore, we assume that X$_1$ and M$_2$ are unlinked and that X$_2$ and M$_1$ are unlinked.

### Parameters of the model

For the case–control study, we have N$_D$ chromosomes in the affected group and N$_d$ chromosomes in the control group. According to their carrying alleles of the M$_1$ and M$_2$ genes, the affected group and control group can be further categorized into more subgroups (Table 1). For example, the affected group with alleles '1' of M$_1$ and '0' of M$_2$ have a size of N$_{10D}$; the control group with alleles '0' of M$_1$ and '1' of M$_2$ have a size of N$_{01d}$.

**Table 1** Grouping table of disease population and general population on M$_1$ and M$_2$ Alleles

| Markers | Affected group | Control population |
|---|---|---|
| M$_1$ = 0, M$_2$ = 0 | N$_{00D}$ | N$_{00d}$ |
| M$_1$ = 0, M$_2$ = 1 | N$_{01D}$ | N$_{01d}$ |
| M$_1$ = 1, M$_2$ = 0 | N$_{10D}$ | N$_{10d}$ |
| M$_1$ = 1, M$_2$ = 1 | N$_{11D}$ | N$_{11d}$ |
| Total size | N$_D$ | N$_d$ |

The numbers of chromosomes carrying different genetic markers follow a multinomial distribution with the given sample sizes ($N_D$ and $N_d$) and frequency vector, i.e. $[p_{00D}, p_{01D}, p_{10D}, p_{11D}]^T$ and $[p_{00d}, p_{01d}, p_{10d}, p_{11d}]^T$. For simplicity, in this study, we assumed the cases and controls have the same size, i.e. $N_D$=$N_d$. The frequencies of different chromosomes in affected or unaffected individuals can be obtained with the disease prevalence, the relative risks of disease alleles, the interaction coefficient, LDs and allele frequencies of suspect, and marker genes in the general population. For example, the expected frequency of chromosomes carrying allele '1' at both M$_1$ and M$_2$ genes can be calculated for the affected individuals as below.

$$
\begin{aligned}
p_{11D} = {} & \frac{(p_{1*}f_{1*} + D_{X1,M1})(p_{*1}f_{*1} + D_{X2,M2})f_{11D}}{f_{1*}f_{*1}} + \\
& \frac{[p_{1*}(1-f_{1*}) + D_{X1,M1}](p_{*1}f_{*1} + D_{X2,M2})f_{01D}}{(1-f_{1*})f_{*1}} + \\
& \frac{(p_{1*}f_{1*} + D_{X1,M1})[p_{*1}(1-f_{*1}) + D_{X2,M2}]f_{10D}}{f_{1*}(1-f_{*1})} + \\
& \frac{[p_{1*}(1-f_{1*}) + D_{X1,M1}][p_{*1}(1-f_{*1}) + D_{X2,M2}]f_{00D}}{(1-f_{1*})(1-f_{*1})}
\end{aligned}
$$

where the notation $p$ indicates the frequencies of genetic markers (M$_1$ and M$_2$) and the notation $f$ notes the frequencies of disease genes (X$_1$ and X$_2$). LD between the gene makers and disease genes is presented in $D$. Assume that $D_1'$ is the Lewontin's linkage disequilibrium value between X$_1$ and M$_1$ and $D_2'$ is the Lewontin's linkage disequilibrium value between X$_2$ and M$_2$. More details of the calculation for different groups can be found in the Additional file 1.

### Expectation and variance of effect size

We cited the LD-based definition of genetic interaction of Zhao et al. [12] in this research. The LD between the marker genes M$_1$ and M$_2$ for the affected and unaffected individuals are defined as $\beta_D = p_{11D} - p_{1*D}p_{*1D}$ and $\beta_d = p_{11d} - p_{1*d}p_{*1d}$, respectively. Then $\hat{\beta} = \hat{\beta}_D - \hat{\beta}_d$ is the estimated effect size of gene–gene interaction in a case–control study.

As the case and control samples are collected independently, the expectation of estimated effect size can be calculated as $E(\hat{\beta}) = E(\hat{\beta}_D) - E(\hat{\beta}_d)$. It is easy to obtain expectations.

$$
\begin{aligned}
E(\hat{\beta}_D) &= p_{11D}p_{00D} - p_{10D}p_{01D} - Cov(\hat{p}_{1*D}, \hat{p}_{*1D}) \\
E(\hat{\beta}_d) &= p_{11d}p_{00d} - p_{10d}p_{01d} - Cov(\hat{p}_{1*d}, \hat{p}_{*1d})
\end{aligned}
$$

The covariance between $\hat{p}_{1*D}$ and $\hat{p}_{*1D}$ are defined as $Cov(\hat{p}_{1*D}, \hat{p}_{*1D})$, and the covariance between $\hat{p}_{1*d}$ and

$\hat{p}_{*1d}$ are defined as $Cov(\hat{p}_{1*d}, \hat{p}_{*1d})$, which can be calculated as.

$$Cov(\hat{p}_{1*D}, \hat{p}_{*1D}) = \frac{p_{11D} - p_{1*D}p_{*1D}}{N_D}$$

$$Cov(\hat{p}_{1*d}, \hat{p}_{*1d}) = \frac{p_{11d} - p_{1*d}p_{*1d}}{N_d}$$

More details could be found in the Additional file 1.

As both the sample size $N_D$ and $N_d$ are large enough, the above mentioned covariance is ignorable. We therefore calculate the expectation of effect size as of below.

$$E(\hat{\beta}) = (p_{11D}p_{00D} - p_{10D}p_{01D}) - (p_{11d}p_{00d} - p_{10d}p_{01d})$$

We have $Var(\hat{\beta}) = Var(\hat{\beta}_D) + Var(\hat{\beta}_d)$. As the calculation of variance is complicated, we cited the results from the reference [12] directly as below.

$$Var(\hat{\beta}_D) = \frac{1}{N_D}\begin{bmatrix} (p_{11D} + p_{10D})(1 - p_{11D} - p_{10D})(p_{11D} + p_{01D})(1 - p_{11D} - p_{01D}) + \\ (1 - 2p_{11D} - 2p_{10D})(1 - 2p_{11D} - 2p_{01D}) \cdot (p_{11D}p_{00D} - p_{10D}p_{01D}) - \\ (p_{11D}p_{00D} - p_{10D}p_{01D})^2 \end{bmatrix}$$

$$Var(\hat{\beta}_d) = \frac{1}{N_d}\begin{bmatrix} (p_{11d} + p_{10d})(1 - p_{11d} - p_{10d})(p_{11d} + p_{01d})(1 - p_{11d} - p_{01d}) + \\ (1 - 2p_{11d} - 2p_{10d})(1 - 2p_{11d} - 2p_{01d})(p_{11d}p_{00d} - p_{10d}p_{01d}) - \\ (p_{11d}p_{00d} - p_{10d}p_{01d})^2 \end{bmatrix}$$

### Statistical test and power calculation

We propose the statistic as below.

$$S = \frac{\hat{\beta}^2}{Var(\hat{\beta})}$$

when the distribution of effect size can be approximated in a normal distribution. We gave the null hypothesis ($H_0$) $\beta = 0$ that there is no interaction between $X_1$ and $X_2$ and the alternative hypothesis ($H_1$) $\beta \neq 0$, that is, there exists an interaction between $X_1$ and $X_2$. Under the null hypothesis, the statistic S follows a central chi-square distribution with a degree of freedom 1. The statistical power of the hypothesis testing can be obtained for different scenarios with different parameters. Details of the statistical test and power calculation were given in the Additional file 1.

## Results

### Expectation and variance of the effect size

Effect size presents the magnitude of differences found in the genetic analysis of various scenarios. To investigate how the gene–gene interaction coefficient $K$ and relative risk $R_x$ ($R_{X1}$ and $R_{X2}$) affect the effect size, we calculated expectation of effect size $E(\beta)$ with varying $K$ and $R_x$. In this evaluation, we keep other parameters constant. Assume that the $X_1$ and $M_1$, $X_2$ and $M_2$ are fully linked (i.e. $D_1$'=1, $D_2$'=1). The risk allele frequencies of all genes ($X_1$, $X_2$, $M_1$, and $M_2$) are assigned to be 0.05 when the disease prevalence is given as 0.1.

Our results show the expectation of effect size linearly inflate with the increasing of interaction when relative risks of both disease genes are 1.5, 2.0, and 2.5, respectively (Fig. 1A). When the interaction coefficient is only 1.0, the effect sizes are $2.04 \times 10^{-3}$, $4.62 \times 10^{-3}$, and $7.68 \times 10^{-3}$, respectively. However, the effect sizes increase to $4.07 \times 10^{-3}$, $9.25 \times 10^{-3}$, and $1.54 \times 10^{-2}$ when the interaction coefficient was improved to 2.0. The higher the risk, the greater the increasing rate of effect size.

Being different from the aforementioned linear relationship, the effect size shows an exponential increase with higher relative risk (Fig. 1B). The effect sizes are $1.02 \times 10^{-3}$, $2.04 \times 10^{-3}$ and $4.07 \times 10^{-3}$ for the interaction of 0.5, 1.0, and 2.0 when the relative risks of both disease genes are given as 1.5. Alternatively, with the twofold higher relative risks ($R_x$=3.0), the effect sizes become $5.56 \times 10^{-3}$, $1.11 \times 10^{-2}$, and $2.23 \times 10^{-2}$, respectively. It indicates a nonlinear relationship of impact factors.

It is critical to accurately estimate effect size in the genetic analysis. We therefore calculated the variance with varying sample sizes to discover their relationship. In this study, we give the interaction coefficient $K$=1.0 and relative risks of both disease genes $R_x$=2.0 when the other parameters are the same as that of before. Our results show the variance of effect size rapidly decline with the increasing of sample size when the risk allele frequencies of $X_1$ are 0.05, 0.1, and 0.15, respectively (Fig. 2). With the sample size $N$=$N_D$=$N_d$=5000, the variances are $1.14 \times 10^{-5}$, $1.12 \times 10^{-5}$, and $1.09 \times 10^{-5}$, respectively. When the size are the twofold large ($N$= 10000), the variances become $5.71 \times 10^{-6}$, $5.58 \times 10^{-6}$, and $5.45 \times 10^{-6}$, respectively. It indicates a nonlinear relationship between the variance and sample size.

### Statistical power of a single test

Statistical power is the central issue for the discovery of gene–gene interactions in genetic association studies. To investigate how the relative risk $R_x$ contributed the power, we calculated the statistical power with

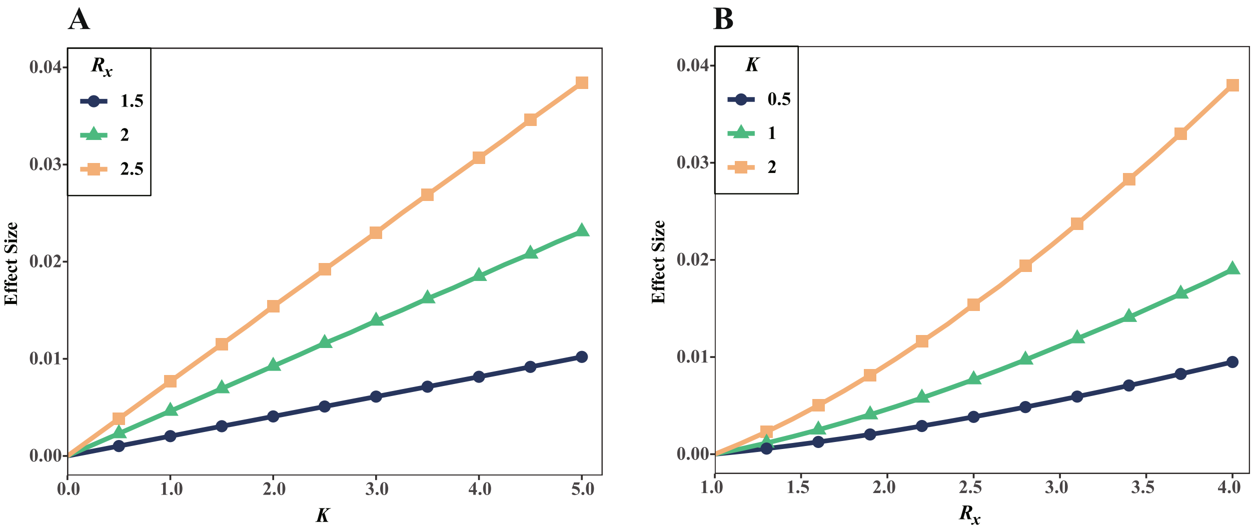

**Fig. 1** Interaction coefficient and relative risks contribute to effect size in case–control studies. **A** The relationships between $E(\beta)$ and $K$ with different $R_x$. **B** The relationship between $E(\beta)$ and $R_x$ under different $K$

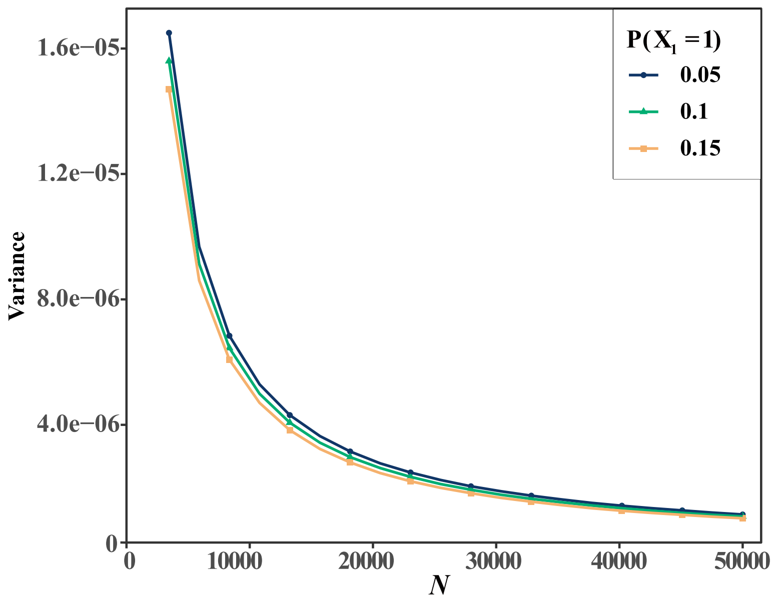

**Fig. 2** Sample sizes contribute to variance in case–control studies

varying $R_x$ and the significant criterion $p$-value $\leq 0.05$. The other parameters are the same as that of before if there lacks further statement. Our results show the substantial increase of statistical power with the inflating of relative risk $R_x$ and the interaction coefficients $K$ under the circumstance where the sample sizes of both case and control groups are 10,000 chromosomes (Fig. 3A). Since the interaction coefficient is as large as 2.0, the statistical power is 0.10 with $R_x = 1.20$ and 0.83 with $R_x = 1.80$. The statistical power is significantly weak when the interaction coefficient was setback to 1.0 or 0.5. Specifically, in the scenario with interaction coefficient 0.5, the power is only 0.05 for $R_x = 1.20$ and 0.12 for $R_x = 1.80$, respectively. Because relative risk of disease gene is generally weak, the aforementioned results indicate the gene–gene interaction is difficult to identify in such circumstances. As sample size could be different in different studies, we further evaluate the power in scenarios with varying sample sizes and the interaction coefficient $K = 1$. When the sample sizes are

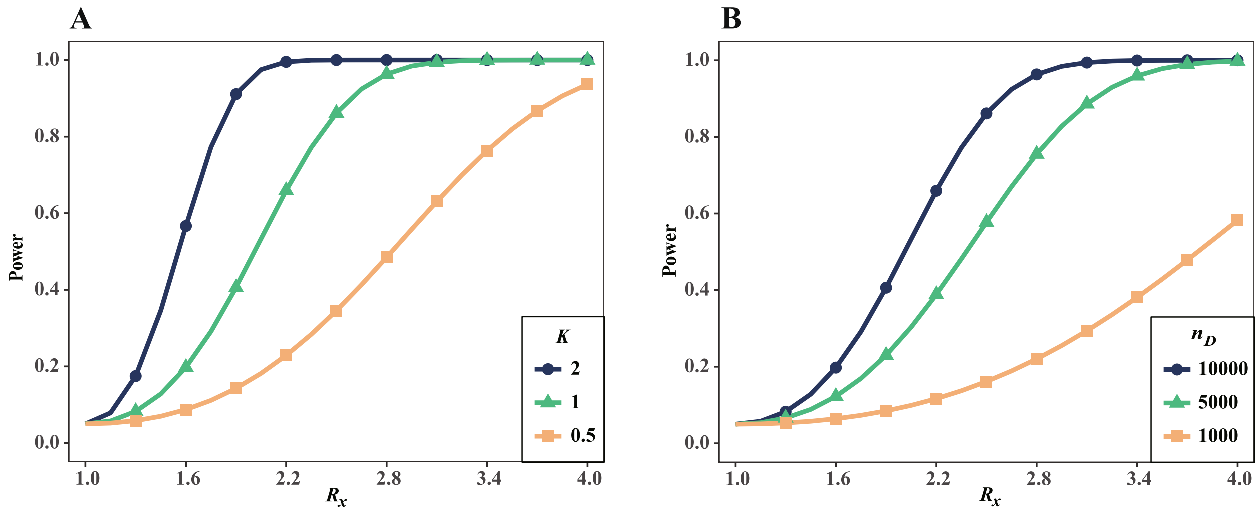

**Fig. 3** Relative risks contribute to power in case–control studies. **A** The relationships between power and $R_x$ with different *K*. **B** The relationship between power and $R_x$ under different sample sizes

1000, 5000, and 10,000, the statistical power are 0.09, 0.28, and 0.49 with relative risk $R_x=2.0$, respectively (Fig. 3B). The results show a great sample size is necessary for identifying gene–gene interaction even if the relative risks of disease genes are enormously large.

To investigate how the interaction coefficient *K* affect the statistical power, we calculated power with varying *K*. In this evaluation, we keep other parameters the same as before. Our results show the power nonlinearly inflate with the increasing of interaction coefficient (Fig. 4A). When the interaction coefficient is weak ($K=1.0$), the power is only 0.06 with $R_x=1.2$, 0.49 with $R_x=2.0$, and 0.86 with $R_x=2.5$, respectively. Alternatively, the power increases to 0.10, 0.96, and 0.99 when the interaction coefficient was improved to $K=2.0$. Our results show the power approximate to 1.0 in the circumstances with the large relative risks and strong interactions. However, the circumstances could be rare in genetic association studies of real world.

We also investigated the statistical power with both varying interaction coefficient and different sample sizes (Fig. 4B). As that was expected, the power is weak when the sample sizes are limited. With the sample size being 1000 and $R_x=2.0$, the power is as small as 0.09 for $K=1.0$, 0.22 for $K=2.0$, respectively.

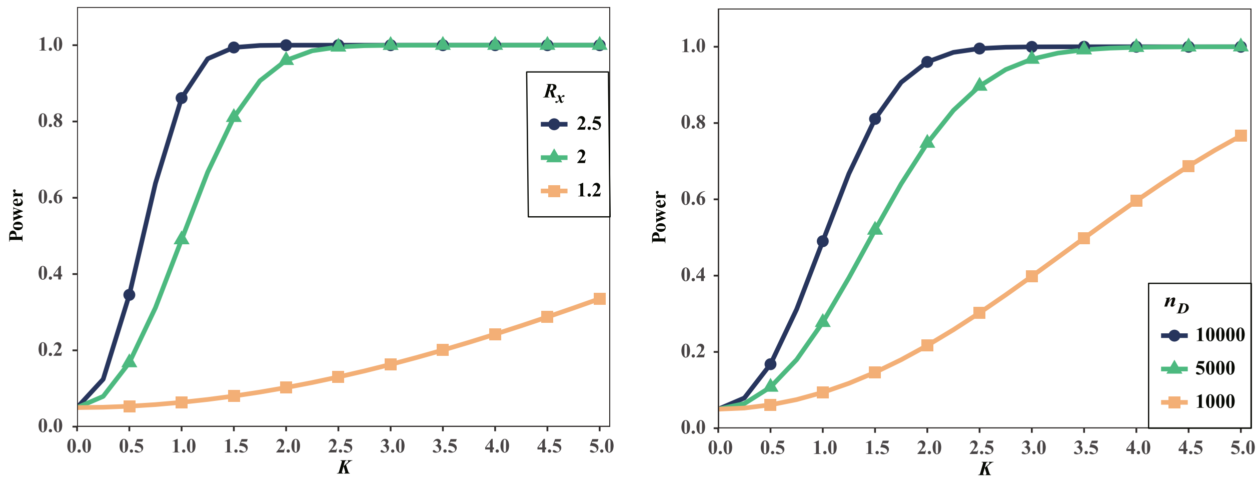

**Fig. 4** Interaction coefficient contributes to power in case–control studies. **A** The relationships between power and *K* with different $R_x$. **B** The relationship between power and *K* under different sample sizes

## Statistical power with multiple testing corrections

Genetic association studies usually involved tens of thousands of genetic markers. It is therefore important to evaluate the statistic power for identification of gene–gene interaction in such stricter criterion in statistical tests. To explore the possible solution for the interaction identification, we evaluated the statistic power with different LD and allele frequencies under varying significant criteria.

By assuming that the $X_1$ and $M_1$ are not fully linked but $X_2$ and $M_2$ are in complete LD (i.e. $D_1' \neq 1$, $D_2' = 1$). We investigate how $D_1'$ affects the statistical power under different levels of significance. In this evaluation, we keep other parameters constant. Among them, the risk allele frequencies of $X_2$, $M_1$, and $M_2$ are assigned to be 0.2 when the sample size are 10,000. We also have $K = 1.0$ and $R_x = 2.0$ with the disease prevalence 0.1.

Our results show the statistical power generally increases with the increasing of $D_1'$ (Fig. 5A, $\text{Power}_1$). When the LD is moderate, say $D_1' = 0.5$, the power is only 0.18, 0.45, and 0.16 for the risk allele frequency of $X_1$ being 0.1, 0.2, and 0.3. Alternatively, with $D_1' = 1.0$, the power was improved to 0.72, 0.95, and 0.87 for the different allele frequencies, respectively. To compare the aforementioned power of interaction identification with that of regular association test, we calculated the power of association test for the marker gene $M_1$ in the same circumstances. Results show the power of regular association test is much higher than that of interaction discovery (Fig. 5A, $\text{Power}_2$). When much stricter criteria ($\textit{sig.level} \leq 5.0 \times 10^{-4}, \leq 5.0 \times 10^{-6}$, and $\leq 5.0 \times 10^{-8}$) were applied in our investigation, the identification of gene–gene interaction became more difficult than ever before (Fig. 5B, C, D). As the significance level was given the same as many genome-wide association studies, $p\text{-value} \leq 5.0 \times 10^{-8}$, there is little opportunity to identify the gene–gene interaction in all kind of circumstances (Fig. 5D). However, as that was expected,

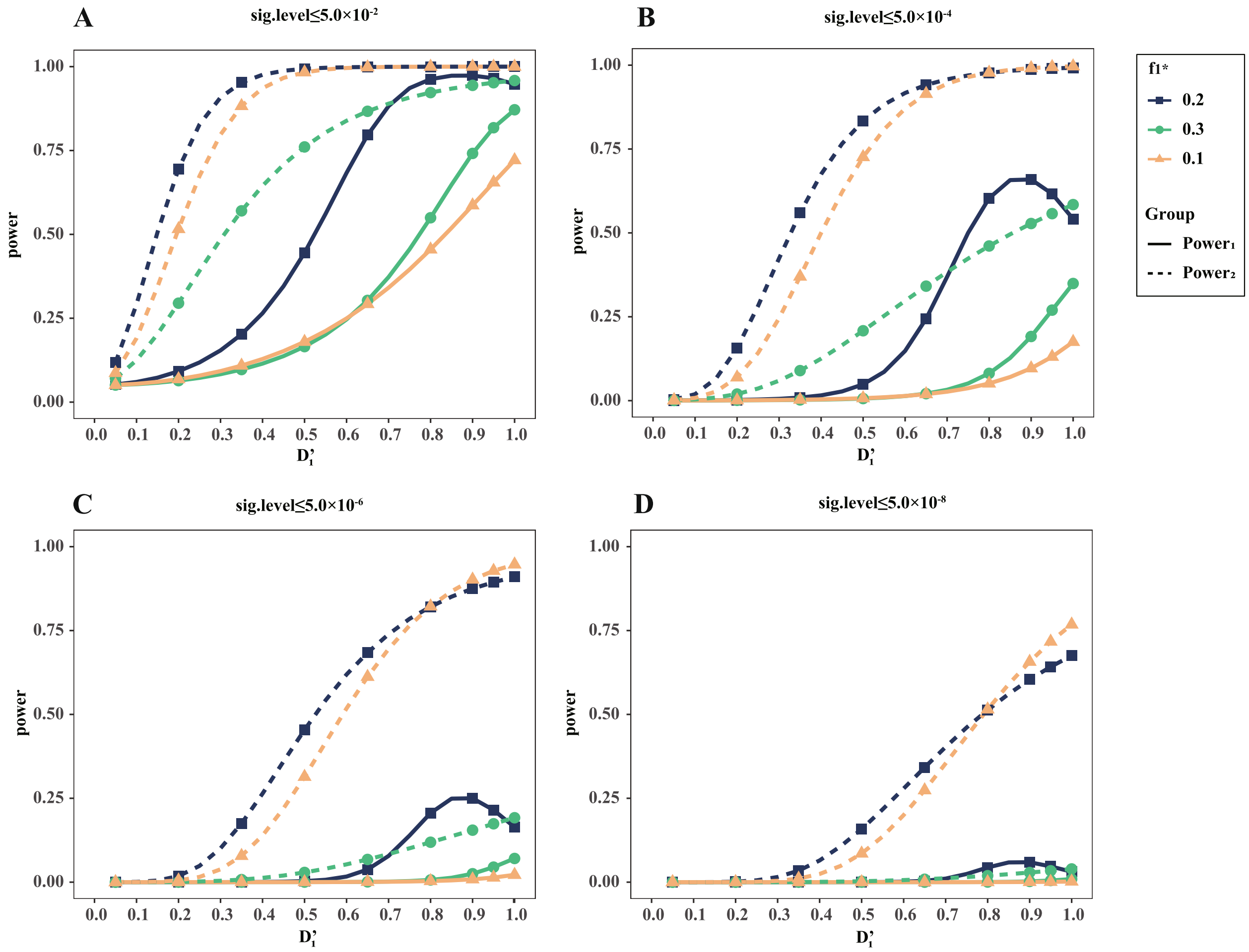

**Fig. 5** $D_1'$ and level of significance contributes to $\text{power}_1$ and $\text{power}_2$. **A-D** The figures show the relationship between $\text{power}_1$, $\text{power}_2$, and $D_1'$ under three groups of $f_{1*}$ when level of significance $\leq 5.0 \times 10^{-2}$, $\leq 5.0 \times 10^{-4}$, $\leq 5.0 \times 10^{-6}$, and $\leq 5.0 \times 10^{-8}$, respectively

it still possible to discover the significant association between the marker gene $M_1$ and the disease (Fig. 5D).

To investigate the effects of interaction coefficient *K*, both the power of interaction identification and the power of regular genetic association test were obtained for different interaction coefficients and different significance levels. Here we gave $D_1'$ and $D_2'$ to be 0.5 and 1.0 respectively, when other parameters were the same as that aforementioned. Our results show the statistical power of both of the tests increase with the increasing of *K* in general scenarios (Fig. 6). However, with the criterion $p$-value $\leq 5.0 \times 10^{-8}$, it is little of power to identify the gene–gene interaction even if the interaction coefficient is as large as 2.0 (Fig. 6D).

## Discussion

Because the influence of a single genetic variant is limited, in order to find the source of unexplained genetic variance and more undiscovered genetic markers, investigators tried to combine multiple genetic markers in their analysis to enhance the statistical power of genetic association study through "gene–gene interaction". Wu and Cui proposed a varying-coefficient model for the detection of nonlinear gene-environment interaction with binary disease traits and got decent power in a variety of genetic settings given reasonable control on false positive rate [24]. However, no similar results have been found in the studies of gene–gene interaction, and the reasons of which have aroused many debates. Some investigators argued that this problem is caused by technical limitations, and a large variety of regularization models for interaction studies have been developed, including regression based, LD based, Bayesian, data filtering and machine learning methods, etc [25, 26]. Nevertheless, the progress of statistical theory and calculation methods have not yet brought substantial help to solve the problem. Our study provided an alternative perspective to the long debates that the gene–gene

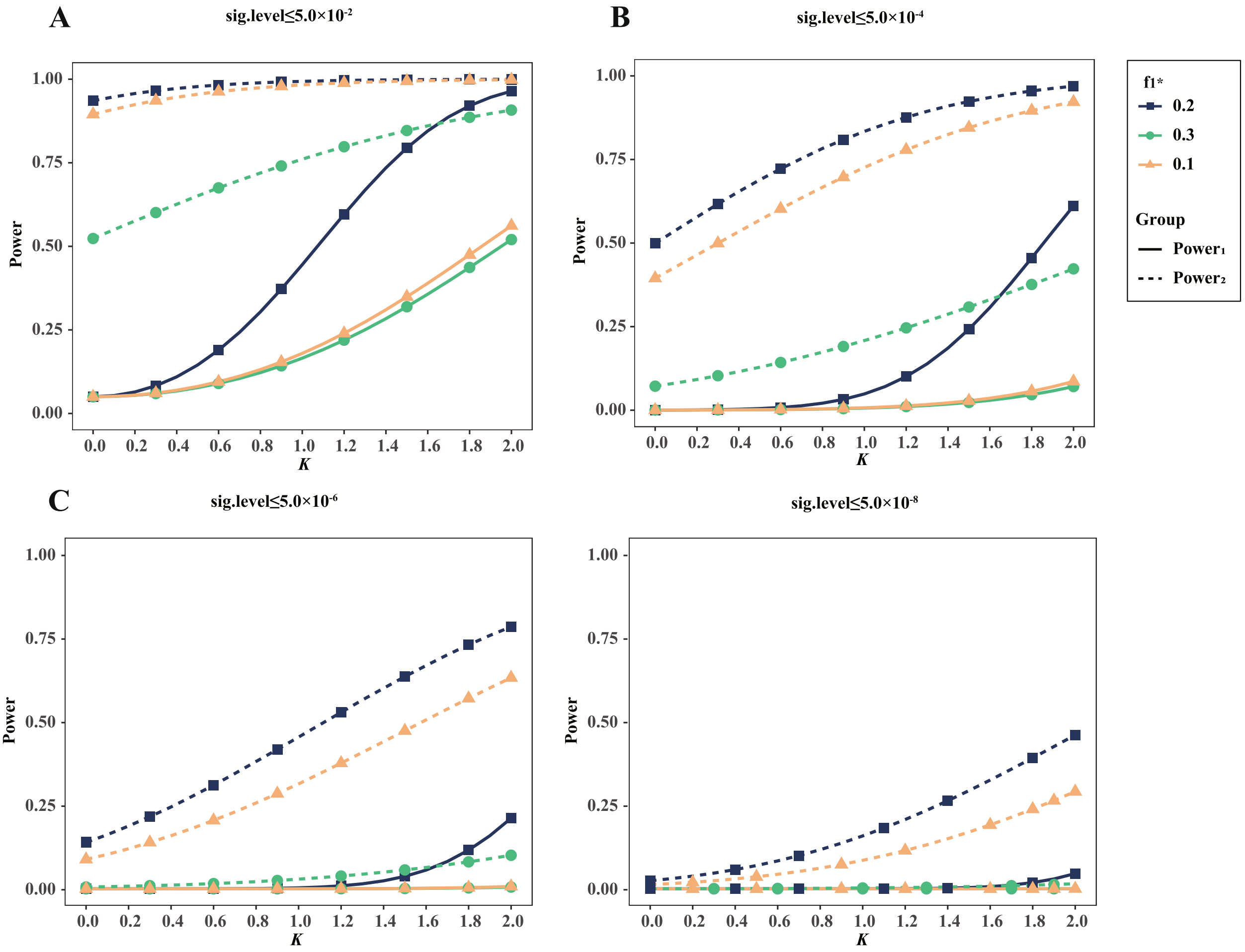

**Fig. 6** The interaction coefficient and level of significance contributes to power$_1$ and power$_2$. **A-D** The figures show the relationship between power$_1$, power$_2$, and *K* under three groups of $f_{1*}$ when level of significance $\leq 5.0 \times 10^{-2}$, $\leq 5.0 \times 10^{-4}$, $\leq 5.0 \times 10^{-6}$, and $\leq 5.0 \times 10^{-8}$, respectively

interaction may not be detectable due to the lack of statistical power.

A large number of genetic markers in actual genetic association studies make it difficult to obtain decent statistical power. For controlling false positive results, multiple test corrections are required to be conducted. Taking the Bonferroni method as an example, for a multiple test with around $1.0\times10^6$ genetic markers, the declared significant threshold after Bonferroni correction is about $5.0\times10^{-8}$. Our results showed that the power of testing gene–gene interaction is close to zero under such rigorous threshold. Kraft et al. developed a joint test of marginal associations and showed that the joint test could have higher power than other joint tests for detecting the gene-environment association [27]. Their method is also helpful to detect the gene–gene interaction due to similar forms of statistical models of the gene-environment and gene–gene interactions. However, as the size of hypotheses for detecting gene–gene interactions is a quadratic function of marker numbers, the chance of finding statistical epistasis will be even lower in the genome-wide analysis with much rigorous significance criterion.

In genetic association studies, it is particularly important to rationally define and measure gene interactions. However, there is no consensus on what values of interaction parameter or epistatic models are appropriate for human disease and investigators have divergent mathematical definition of gene interaction strength. In the study of Gauderman [28, 29], gene interaction strength was related with the interaction regression coefficient of logistic model, while Zhao [12] defined it as a measure of interaction between two unlinked loci associated with the penetrance of different genotypes at the disease-susceptibility locus. In this study, we define coefficient of gene interaction as a concise ratio of allele frequencies about two independent disease genes located on the same chromosome, and the magnitude of which reflects the strength of the interaction, thus providing a foundation for developing statistics for detection of interaction.

An important benefit from investigating the genetics of human disease is to predict the risk that individuals may have of succumbing to a specific disease, the knowledge of which can then be used by the clinician in prevention, diagnosis, prognosis, and treatment [30]. That is why the relative risk is an important parameter in the studies of gene–gene interaction and it is significant to use reasonable value of relative risk when conducting theoretical research. The relative risk of disease due to one allele is typically of the order of 1.1 to 2.0 in the real world [30, 31], and the values in this range has also been used in previous studies of gene–gene interaction [12]. In this study, we evaluated the statistic power with different LD, allele frequencies and interaction coefficient under reasonable relative risks, ensuring the reliability of the results.

We selected the haplotype model as the research model for highlighting the main results. In the reference Hu et al. [32], the authors mentioned three different definitions of genetic interaction as the penetrance-based definition, the logit-based definition, and the LD-based definition. They showed that the 3 definitions are equivalent under the circumstances that linkage equilibrium holds in general population for the two loci dominant (or recessive) disease model. Since the framework of their genotype model researching gene–gene interaction applies as well in haplotype model when haplotype at two loci can be inferred from diplotype without uncertainty [32], our conclusions about gene–gene interaction apply to dominant and recessive inheritance patterns under Hardy–Weinberg equilibrium.

The aforementioned report [32] discussed the power of statistical tests to identify genetic interactions and emphasized the significance of the power in the specific circumstances. In our study, we assumed that the casual variants are not determined but in LD with the detected markers. Alternatively, the previous report gave the effect sizes of disease risk and interaction to the detected markers. The previous report may overestimate the power in general circumstances because there usually lacks a complete association between causal variants and genetic markers. The primary difference explained the different results in our research and the previous studies.

The results in this article are preliminary. Interaction between high-order interactions among multiple loci have not been studied. Further research is required to investigate gene–gene interactions among multiple loci and explore the problem of missing statistical epistasis in genetic association analysis in a more in-depth and comprehensive manner, so as to lay a foundation for the optimization of genetic association analysis methods. In addition, this analysis is not applicable to general disease models with a total of nine genotypes since our research is based on haplotype model. For more general disease models, dominant coding or recessive coding for each of two loci [29, 32] can be employed to study gene–gene interactions as done in the previous reports.

## Conclusions

This study provides a possible explanation for the long-standing important question of "lack of epistasis" in genetic association studies. It points out that the lack of epistasis tends to be an inevitable result caused by the statistical principles of genetic association research methods and is the inherent characteristic of the research method itself. The results of this study suggest that looking for epistasis from genetic association studies is an extremely tough task.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12920-024-01884-w.

**Additional file 1.** Supplementary Methods.

### Acknowledgements
Not applicable.

### Authors' contributions
Y.H. and Z.Y. conceptualized the study; Y.H., J.M. and J.L. developed the methodology; J.M. and J.L. performed the investigation; Y.H. and Z.Y. supervised the project; Y.H., J.M. and J.L. prepared the manuscript.

### Funding
This work was supported by grants from National Natural Science Foundation of China (grant numbers 32170634, and 31871255 to Y.H). Y.H. was also supported by Shanghai Fifth People's Hospital (grant 2021IBSPI02) and Shanghai Municipal Science and Technology (grant 2017SHZDZX01).

### Availability of data and materials
The data used in this study are all simulated data.

The related R program can be obtained from Github at jrma21/Poor-Statistical-Power-in-Population-based-Association-Study-of-Gene-Interaction (github.com).

## Declarations

### Ethics approval and consent to participate
Not applicable.

### Consent for publication
Not applicable.

### Competing interests
The authors declare no competing interests.

Received: 6 July 2023 Accepted: 19 April 2024
Published online: 27 April 2024

### References

1. Cordell HJ. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet. 2002;11:2463–8.
2. Whitlock MC, Phillips PC, Moore FB-G, Tonsor SJ. Multiple Fitness Peaks and Epistasis. Annu Rev Ecol Syst. 1995;26:601–29.
3. Phillips PC. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet. 2008;9:855–67.
4. Boone C, Bussey H, Andrews BJ. Exploring genetic interactions and networks with yeast. Nat Rev Genet. 2007;8:437–49.
5. Mani R, St.Onge RP, Hartman JL, Giaever G, Roth FP. Defining genetic interaction. Proc Natl Acad Sci USA. 2008;105:3461–6.
6. Zeyl C. How missing genes interact. Nat Genet. 2007;39:440–2.
7. Jasnos L, Korona R. Epistatic buffering of fitness loss in yeast double deletion strains. Nat Genet. 2007;39:550–4.
8. Hansen TF, Wagner GP. Modeling genetic architecture: a multilinear theory of gene interaction. Theor Popul Biol. 2001;59:61–86.
9. Fisher RA. XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh. 1919;52(2):399-433. https://doi.org/10.1017/S0080456800012163.
10. Cockerham CC. An extension of the concept of partitioning hereditary variance for analysis of Covariances among relatives when epistasis is present. Genetics. 1954;39:859–82.
11. Bossé Y, Amos CI. A decade of GWAS results in lung cancer. Cancer Epidemiol Biomark Prev. 2018;27:363–79.
12. Zhao J, Jin L, Xiong M. Test for interaction between two unlinked loci. Am J Hum Genet. 2006;79:831–45.
13. Chu M, et al. A genome-wide gene–gene interaction analysis identifies an epistatic gene pair for lung cancer susceptibility in Han Chinese. Carcinogenesis. 2014;35:572–7.
14. Moore JH, Williams SM. Epistasis and its implications for personal genetics. Am J Hum Genet. 2009;85:309–20.
15. Moore JH, Williams SM. New strategies for identifying gene-gene interactions in hypertension. Ann Med. 2002;34:88–95.
16. Cordell HJ. Detecting gene–gene interactions that underlie human diseases. Nat Rev Genet. 2009;10:392–404.
17. Nelson MR, Kardia SL, Ferrell RE, Sing CF. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. Genome Res. 2001;11:458–70.
18. Kooperberg C, Ruczinski I, LeBlanc ML, Hsu L. Sequence analysis using logic regression. Genet Epidemiol. 2001;21(Suppl 1):S626-631.
19. Lunetta KL, Hayward LB, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet. 2004;5:32.
20. Ueki M, Cordell HJ. Improved Statistics for Genome-Wide Interaction Analysis. PLoS Genet. 2012;8:e1002625.

21 Moore JH. The challenges of whole-genome approaches to common diseases. JAMA. 2004;291:1642–3.

22. Moore, J. H. Genome-Wide Analysis of Epistasis Using Multifactor Dimensionality Reduction: Feature Selection and Construction in the Domain of Human Genetics. In J. Tan (Ed.), Medical Informatics: Concepts, Methodologies, Tools, and Applications. Hershey: IGI Global; 2009.
23. Simon SD. Understanding the odds ratio and the relative risk. J Androl. 2001;22:533–6.
24. Wu C, Cui Y. A novel method for identifying nonlinear gene-environment interactions in case-control association studies. Hum Genet. 2013;132:1413–25.
25. Wei W-H, Hemani G, Haley CS. Detecting epistasis in human complex traits. Nat Rev Genet. 2014;15:722–33.
26. Zhou F, Ren J, Lu X, Ma S, Wu C. Gene-environment interaction: a variable selection perspective. Methods Mol Biol. 2021;2212:191–223.
27. Kraft P, Yen Y-C, Stram DO, Morrison J, Gauderman WJ. Exploiting gene-environment interaction to detect genetic associations. Hum Hered. 2007;63:111–9.
28. Gauderman WJ. Sample size requirements for matched case-control studies of gene-environment interaction. Stat Med. 2002;21:35–50.
29. Gauderman WJ. Sample size requirements for association studies of gene-gene interaction. Am J Epidemiol. 2002;155:478–84.
30. Wray NR, Goddard ME, Visscher PM. Prediction of individual genetic risk to disease from genome-wide association studies. Genome Res. 2007;17:1520–8.
31. Bertram L, McQueen MB, Mullin K, Blacker D, Tanzi RE. Systematic meta-analyses of Alzheimer disease genetic association studies: the AlzGene database. Nat Genet. 2007;39:17–23.
32. Hu JK, Wang X, Wang P. Testing gene-gene interactions in genome wide association studies. Genet Epidemiol. 2014;38:123–34.

## Publisher's Note