scientific reports

Check for updates

OPEN

Meta-analysis of SNP-environment interaction with heterogeneity for overlapping data

Qinqin Jin[1,2]✉ & Gang Shi[1]

Meta-analysis is a popular method used in genome-wide association studies, by which the results of multiple studies are combined to identify associations. This process generates heterogeneity. Recently, we proposed a random effect model meta-regression method (MR) to study the effect of single nucleotide polymorphism (SNP)-environment interactions. This method takes heterogeneity into account and produces high power. We also proposed a fixed effect model overlapping MR in which the overlapping data is taken into account. In the present study, a random effect model overlapping MR that simultaneously considers heterogeneity and overlapping data is proposed. This method is based on the random effect model MR and the fixed effect model overlapping MR. A new way of solving the logarithm of the determinant of covariance matrices in likelihood functions is also provided. Tests for the likelihood ratio statistic of the SNP-environment interaction effect and the SNP and SNP-environment joint effects are given. In our simulations, null distributions and type I error rates were proposed to verify the suitability of our method, and powers were applied to evaluate the superiority of our method. Our findings indicate that this method is effective in cases of overlapping data with a high heterogeneity.

Genome-wide association studies (GWASs) are effective for the identification of single nucleotide polymorphisms (SNPs) associated with complex traits or disease[1–3]. Meta-analysis[4–8], which combines the results of multiple studies, is a common method used to increase the sample size[5,7,9], which can reduce false positive results, increase power, and increase the probability of finding new associations. The fixed effect model is commonly used for meta-analysis, in which the effects between studies are assumed to be equal. However, in recent studies, meta-analyses have been employed using new designs, by combining different related traits or diseases[10,11], environments[12], populations[13], tissues[14], and cancer types[15,16]. These combinations lead to different effect sizes between studies, which is called heterogeneity[17]. Thus, the fixed effect model is not suitable. The traditional random effect model[4] take heterogeneity into account ,but implicitly assumes a conservative null hypothesis model. Therefore, it provides a power even lower than the fixed model. A modified random effect model method[18] was proposed to overcome this problem and be widely used in various analyses[12,14,17,19,20].

However, in practice, there are many overlapping individuals between studies. This may be caused unintentionally or intentionally by the researchers. If these overlapping individuals exist but they are ignored, spurious associations may occur[21,22]. In recent years, researchers have proposed several methods for overlapping data[16,17,21–25]. These methods are all used to test the main effects of SNPs. Lin[21] proposed a correlation matrix and applied it to the fixed model method for overlapping data among studies. Han[22] transformed the covariance structure of data, which then became a form of diagonal matrix. This transformation makes Lin's method more flexible, which can be widely applied by meta-analysis methods, such as the random effects model. Based on the modified random effect model method[18], Lee[17] proposed a new method for overlapping data. This method combined the fixed effect model and random effect model, which gave a higher power regardless of the heterogeneity.

The meta-regression method (MR)[26] is a powerful and robust method in a fixed-effect model. This method has two steps. First, individuals in each study are divided into several groups based on the distribution of environmental variables, where the number of individuals in each group is equal. In each group, linear regression is used to estimate the main effects, standard errors, and mean environment variables of the SNP. Second, researcher collects all the above results and performs a meta-regression to investigate interactions between SNP and the environment. It can be used to test for the main effects of SNP, SNP-environment interactions, and joint effects. In addition, this method is considered to be a robust method when confounding effects exist, such as interactions

[1]State Key Laboratory of Integrated Services Networks, Xidian University, 2 South Taibai Road, Xi'an 710071, Shaanxi, China. [2]Applied Science College, Taiyuan University of Science and Technology, Taiyuan 030024, Shanxi, China. ✉email: qinqinjin@stu.xidian.edu.cn

between covariates and genetic effects or interactions between covariates and environmental factors[27]. Based on Lin's method[21] and Han's method[22], we extended MR to overlapping MR (OMR)[28]. This method is designed for SNP-environment interactions under the fixed effect model, as well as for overlapping data. We also extended MR to account for heterogeneity; that is, we added random effects of the SNP and SNP-environment interactions to the fixed-effect SNP-environment interaction model. This method is denoted as the random effect MR (RMR)[29], which gives a higher power than MR under the fixed-effect model when heterogeneity exists. The Q-Q plot of the null distribution obtained by MR will shift upward when overlapping data exists. The more overlapping the data, the more obvious the deviation will be. The fixed effect model OMR[28] controls spurious associations caused by overlapping data. When heterogeneity exists and is large, the power of RMR is higher than that of MR. Similarly, when overlapping data and heterogeneity exist, the power obtained by OMR will also be affected by heterogeneity, and the power it provides will be reduced. However, no study has yet considered this condition.

In this paper, inspired by OMR and Lee's method[17] which is proposed for testing SNP main effect with overlapping data, we propose random effect overlapping MR (ROMR) which is a new method to consider overlapping data based on the RMR [29]. Our method is designed to test the SNP-environment interaction effect or the SNP and SNP-environment joint effects with overlapping data. This paper is organized as follows. In the Materials and Methods section, we introduce the correlation matrix into the RMR. We also present a new method to calculate the likelihood function. In the Results section, we carry out simulations to examine the null distribution, type I error rate, and power of our method. We also compare our method with the OMR. In the Discussion and Conclusion section, the results of this paper are analyzed and used to draw conclusions.

Materials and methods

Fixed effect overlapping MR. OMR is a method that extends from fixed effect MR, which is a powerful and robust method under the condition of independent data. This method has two procedures. First, by continuous or dichotomous environmental exposure distribution, each study is divided into several groups. In this process, each group is a subset of the study, and percentiles of the environmental exposure can be used to divide the study into several groups with approximately the same sample sizes. Then, in each group, the coefficient and variance of the main effects of SNPs are estimated. Second, the main effect of SNP and its corresponding standard deviation in each group are collected for regression analysis. Then, either the overall mean SNP-environment effect and its variance is estimated, or the mean SNP and SNP-environment joint effect vector and its variance matrix are estimated.

Assume that the environmental exposure is continuous and $\widehat{\beta}_{ij}$ is set to be the estimation of the main effects of SNPs in the i-th study and j-th group, where subscript $i = 1, 2, \ldots,$ n is the sample size of studies and subscript $j = 1, 2, \ldots, \mathrm{n_i}$ is the sample size of the groups in the i-th study. $\widehat{e}_{ij}$ and E_{ij} are the standard error and mean environment exposure of the i-th study and j-th group, respectively. Under the second OMR procedure, the formula for the environment-dependent SNP effect $\boldsymbol{\beta}$ can be expressed in the following form:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

where

$$\hat{\boldsymbol{\beta}} = \begin{pmatrix} \hat{\boldsymbol{\beta}}_1 \\ \hat{\boldsymbol{\beta}}_2 \\ \vdots \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{n}} \end{pmatrix}, \hat{\boldsymbol{\beta}}_{\boldsymbol{i}} = \begin{pmatrix} \hat{\beta}_{i1} \\ \hat{\beta}_{i2} \\ \vdots \\ \hat{\beta}_{in_i} \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_{\boldsymbol{n}} \end{pmatrix}, \boldsymbol{X}_{\boldsymbol{i}} = \begin{pmatrix} 1 & E_{i1} \\ 1 & E_{i2} \\ \vdots & \vdots \\ 1 & E_{in_i} \end{pmatrix},$$

$$\boldsymbol{\varepsilon} = \begin{pmatrix} \boldsymbol{\varepsilon}_1 \\ \boldsymbol{\varepsilon}_2 \\ \vdots \\ \boldsymbol{\varepsilon}_{\boldsymbol{n}} \end{pmatrix}, \boldsymbol{\varepsilon}_{\boldsymbol{i}} = \begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \\ \vdots \\ \varepsilon_{in_i} \end{pmatrix}$$

$$\boldsymbol{\alpha} = \begin{pmatrix} \alpha_0 \\ \alpha_1 \end{pmatrix}, \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{\Sigma}_{\mathrm{n}} \end{pmatrix}, \boldsymbol{\Sigma}_i = \begin{pmatrix} \hat{e}_{i1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \hat{e}_{in_i} \end{pmatrix}$$

and $\varepsilon_{ij} \sim N(0, \widehat{e}_{ij})\, i = 1, 2, \ldots, \mathrm{n}, \mathrm{j} = 1, 2, \ldots, \mathrm{n_i}$.

Let $\boldsymbol{C}$ be the correlation matrix. The element of this matrix is

$$\gamma_{i_h j_k} \approx n_{i_h j_k} / \sqrt{n_{i_h} n_{j_k}}$$

where the subscript n_{i_h} and n_{j_k} are presented as the size of the h-th group of study i and k-th group of study j, respectively, and $n_{i_h j_k}$ is the size of the overlap individuals between the h-th and k-th group.

The covariance matrix of this method is

$$\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{1/2} \boldsymbol{C} \boldsymbol{\Sigma}^{1/2}$$

It has another form as

$$\boldsymbol{\Omega} = \boldsymbol{diag}(\boldsymbol{e}^{'} (\boldsymbol{\Sigma}^{1/2} \boldsymbol{C} \boldsymbol{\Sigma}^{1/2})^{-1})^{-1}$$

where $\boldsymbol{e} = (1, 1, \ldots, 1)$ and its length is the sum of all group sizes.

The formula of the linear unbiased estimators $\widehat{\boldsymbol{\alpha}}$ and $\mathrm{Cov}\left(\widehat{\boldsymbol{\alpha}}\right)$ are expressed as follows

$$\widehat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^{'}\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}^{'}\boldsymbol{\Omega}^{-1}\widehat{\boldsymbol{\beta}}$$

$$\widehat{\alpha}_2 = (0, 1)\widehat{\boldsymbol{\alpha}}$$

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\alpha}}\right) = \left(\boldsymbol{X}^{'}\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}$$

$$\mathrm{Cov}\left(\widehat{\boldsymbol{\alpha}}\right)_{22} = (0, 1)\mathrm{Cov}\left(\widehat{\boldsymbol{\alpha}}\right)\begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

Under null distribution $\widehat{\alpha}_2 = 0$ and $\widehat{\boldsymbol{\alpha}} = 0$, the Wald statistic of the SNP-environment interaction effect and the SNP and SNP-environment joint effects follow 1 and 2 degrees of freedom (df) χ^2 distribution, respectively.

Random effect overlapping MR. This method is an extension of the OMR[28] and the recently proposed RMR[29]. Under this method, the random effects for the SNP main and SNP-environment interaction are denoted as γ. The environment-dependent SNP effect $\widehat{\boldsymbol{\beta}}$ is presented as follows:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{Z} = \begin{pmatrix} \boldsymbol{Z}_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \boldsymbol{Z}_n \end{pmatrix}, \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\gamma}_1 \\ \boldsymbol{\gamma}_2 \\ \vdots \\ \boldsymbol{\gamma}_n \end{pmatrix}$$

and

$$\boldsymbol{\gamma}_i = \begin{pmatrix} \gamma_{i0} \\ \gamma_{i1} \end{pmatrix}$$

Here, variable γ_{i0} is denoted as the random main effect of SNP in the i-th study and γ_{i1} is denoted as the random effect of SNP-environment interaction in the i-th study. The vector $\boldsymbol{\gamma}_i = \begin{pmatrix} \gamma_{i0} \\ \gamma_{i1} \end{pmatrix}$ follows a bivariate normal distribution with $\begin{pmatrix} \gamma_{0i} \\ \gamma_{1i} \end{pmatrix} \sim \mathrm{N}(0, \boldsymbol{D}_\mathrm{s})$. The variable $\widehat{\boldsymbol{\beta}}$ followed a multivariate normal distribution, as follows:

$$\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\alpha}, \boldsymbol{V})$$

where in the overlapping condition $\boldsymbol{V} = \boldsymbol{Z}\boldsymbol{D}\boldsymbol{Z}^{'} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{C}\boldsymbol{\Sigma}^{1/2}$ is a real symmetric matrix**,** denote $\lambda_1, \lambda_2, \ldots, \lambda_M$ where $M = \sum_{i=1}^{n} n_i$ as the eigenvalues of matrix $\boldsymbol{V}$, denote $\xi_1, \xi_2, \ldots, \xi_M$ as the orthogonal eigenvector of matrix $\boldsymbol{V}$, that is to say, $|(\xi_1, \xi_2, \ldots, \xi_M)| = 1$ and $(\xi_1, \xi_2, \ldots, \xi_M) = (\xi_1, \xi_2, \ldots, \xi_M)^{-1}$. Then $|\boldsymbol{V}| = \lambda_1 * \lambda_2 * \cdots * \lambda_M$.

The likelihood function under this model can be written as

$$l_1 = \sum\nolimits_{i=1}^{M} \ln\left|\lambda_\mathrm{i}^1\right| + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\alpha}\right)^{'} \mathrm{V}^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\alpha}\right) + \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{n_i}\ln(2\pi)$$

The estimation of this likelihood function is given by the minimum variance quadratic unbiased estimator (MIVQUE(0))[30,31] and Newton–Raphson algorithms[32–34]. The detailed process is given in Jin[29].

Test of SNP-environment interaction. Under this test, we suppose that there is no interaction effect and no interaction heterogeneity, that is, $\alpha_1 = 0$ and $\boldsymbol{D}_\mathrm{s} = \begin{pmatrix} u_1^2 & 0 \\ 0 & 0 \end{pmatrix}$. The reduced model can then be given as follows:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{X}\alpha_0 + \boldsymbol{Z}\gamma_0 + \boldsymbol{\varepsilon}$$

where

$$\boldsymbol{X} = \boldsymbol{Z} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \\ \vdots \\ \boldsymbol{X}_n \end{pmatrix}, \boldsymbol{X}_i = (1, 1, \cdots, 1)^{'}$$

and the dimension of $\boldsymbol{X}_i$ is n_i. In this model, $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(\boldsymbol{X}\alpha_0, \boldsymbol{V})$, the covariance matrix is $\boldsymbol{V} = \boldsymbol{ZDZ}^{'} + \boldsymbol{\Sigma}^{1/2}\boldsymbol{C}\boldsymbol{\Sigma}^{1/2}$. Denote $\lambda_1^1, \lambda_2^1, \ldots, \lambda_M^1$ where $M = \sum_{i=1}^{n} n_i$ as the eigenvalues of matrix $\boldsymbol{V}$, then $|\boldsymbol{V}| = \lambda_1^1 * \lambda_2^1 * \cdots * \lambda_M^1$.

The -2 times of the log likelihood for this model is

$$l_2 = \sum_{i=1}^{M} \ln\left|\lambda_{\mathrm{i}}^1\right| + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\alpha_0\right)^{'} \mathrm{V}^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\alpha_0\right) + \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{n_i}\ln(2\pi)$$

As in Jin[29], the likelihood ratio statistic for the test of SNP-environment interaction is given as follows:

$$\mathrm{L_I} = \widehat{l_2} - \widehat{l_1}$$

where $\widehat{l_1}$ is the minimum of l_1 and $\widehat{l_2}$ is the minimum of l_2. The statistic $\mathrm{L_I}$ asymptotically follows an equal mixture of 2 df χ^2 distribution and 3 df χ^2 distribution. Its p-value is calculated by $0.5\ (\mathrm{P}\ (\chi_2^2 > \mathrm{L_I}) + \mathrm{P}(\chi_3^2 > \mathrm{L_I}))$[29].

Joint test of SNP and SNP-environment. Under this test, we suppose that $\boldsymbol{\alpha} = 0$ and $\boldsymbol{D}_{\mathrm{s}} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}$, that is to say, no SNP main, SNP-environment interaction fixed effects, and no corresponding heterogeneity. The null model can be given as follows:

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\varepsilon}$$

Then, $\widehat{\boldsymbol{\beta}} \sim \mathrm{N}(0, \boldsymbol{V})$ and $\boldsymbol{V} = \boldsymbol{\Sigma}^{1/2}\mathbf{C}\boldsymbol{\Sigma}^{1/2}$. The eigenvalues of the covariance matrix $\boldsymbol{V}$ are denoted as $\lambda_1^0, \lambda_2^0, \ldots, \lambda_M^0$, where $M = \sum_{i=1}^{n} n_i$. Then $|\boldsymbol{V}| = \lambda_1^0 * \lambda_2^0 * \cdots * \lambda_M^0$.

The -2 times of the log likelihood for this model is

$$l_0 = \sum_{i=1}^{M} \ln\left|\lambda_{\mathrm{i}}^0\right| + \left(\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\alpha_0\right)^{'} \boldsymbol{V}^{-1}\left(\widehat{\boldsymbol{\beta}} - \boldsymbol{X}\alpha_0\right) + \sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{n_i}\ln(2\pi)$$

The likelihood ratio statistic for the joint test of the SNP main and SNP-environment is given as follows:

$$\mathrm{L_J} = \widehat{l_0} - \widehat{l_1}$$

where $\widehat{l_0}$ is the evaluated value of l_0. The statistic $\mathrm{L_J}$ asymptotically follows a ξ:0.5:(0.5-ξ) mixture of 3 df χ^2 distribution, 4 df χ^2 distribution, and 5 df χ^2 distribution. The value of ξ depends on the given data and is solved by the information matrix. The p-value is calculated by$(0.5\text{-}\xi)\ \mathrm{P}\ (\chi_3^2 > \mathrm{L_J}) + 0.5\mathrm{P}(\chi_4^2 > \mathrm{L_J}) + \xi\ \mathrm{P}\ (\chi_5^2 > \mathrm{L_J})$[29].

Ethics approval. The authors have no ethical conflicts to disclose.

Results

Simulation. The relationship among the quantitative trait Y, the genotype of SNP G, and the environmental variable E is presented as follows:

$$Y = (\beta_G + \gamma_G)G + (\beta_{G\times E} + \gamma_{G\times E})G \times E + \beta_E E + \varepsilon$$

The quantitative trait Y was simulated as a standardized normal distribution, that is, with a mean of 0 and a variance of 1. SNP was assumed as an additive genetic effect, and its minor allele frequency was 0.3. G was coded as the number of minor alleles. In each study, 1000 points following a standard uniform distribution were generated. If these points fell in $[0, 0.3^2]$, then G was set to 2. If these points fell in $[0.3^2, 0.3^2 + 2 \cdot 0.3 \cdot (1 - 0.3)]$, then G was set to 1, else G was set to 0. The environmental variable E was also simulated as a standardized normal distribution. A 10% variation in Y was explained by the environmental term $\beta_E E$. Fixed effects β_G, $\beta_{G\times E}$ and random effects γ_G, $\gamma_{G\times E}$ changed in simulation datasets. The random error ε was normally distributed, with a zero mean and a variance chosen so that the variance of Y was 1. In our simulation, we generated 1000 replications, each replication had 12 studies, each study had 1000 unrelated individuals, and each individual had one quantitative trait Y, one environmental variable E, and one SNP G. Across all the studies, 100 and 400 overlapping individuals were observed. Before analysis, the individuals in each study were divided into five groups according to the distribution of E. The main effects of SNP and standard errors were estimated by linear regression. The mean of the environmental variables E of each stratum was then calculated.

To test the null distribution of statistics for the SNP-environment interaction, we assumed that both the SNP-environment interaction and its corresponding heterogeneity were zero. That is to say, $\beta_{G\times E} = 0$ and $\gamma_{G\times E} = 0$. The main effect of SNP β_G was set to a square root of 0.1, and the random effect of this effect was set to be normally distributed, with a mean of 0 and a variance of 0.02. We calculated the minimum estimates of the likelihood functions l_1 and l_2, such that the statistic $\mathrm{L_I}$ could be obtained. Finally, empirical P-values were calculated using a 0.5:0.5 mixture of 2 and 3 df χ^2 distributions, which was the theoretical distribution of the interaction test. Then, these were compared with expected values following a uniform distribution between 0 and 1. A Q–Q plot was drawn through the two types of P-values. To test the null distribution of statistics for the

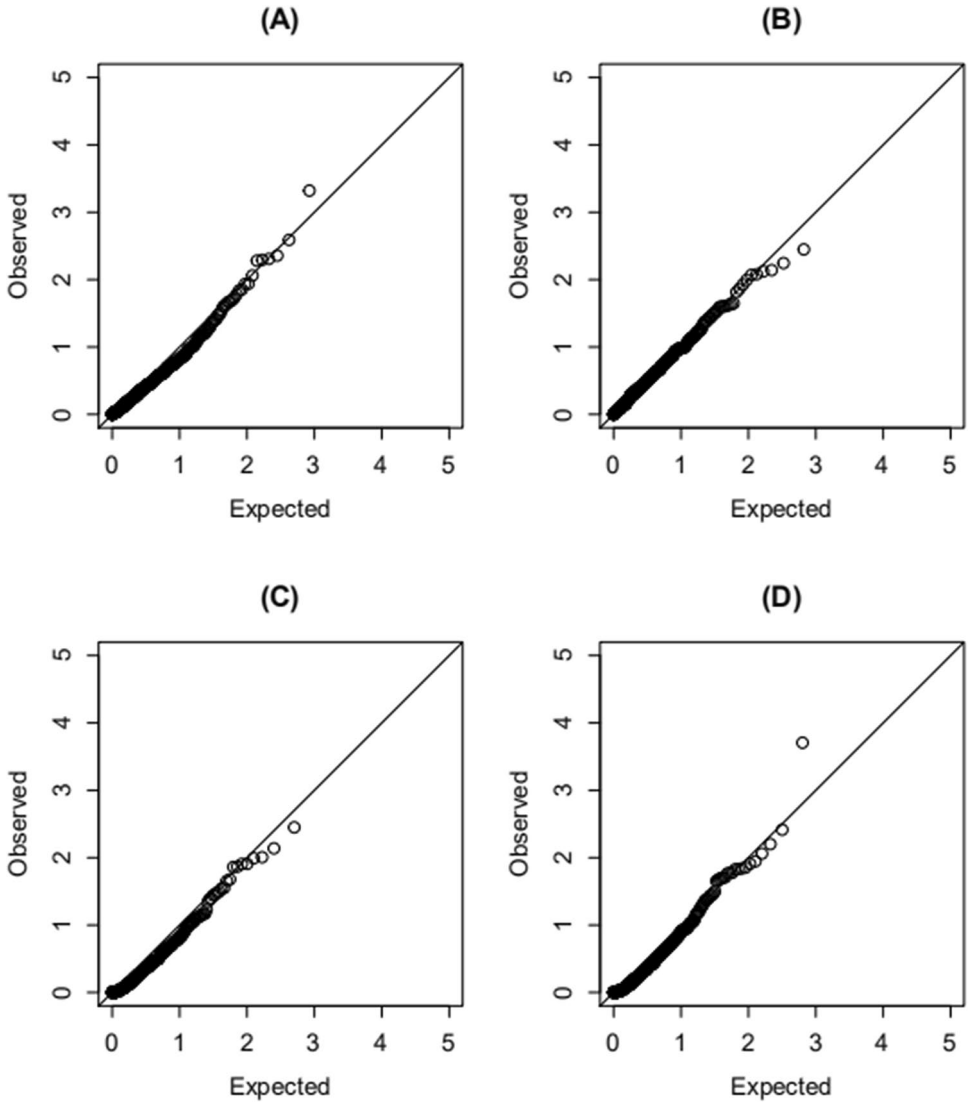

Figure 1. Q–Q plots of the null distributions in the test for SNP-environment interaction effects and the joint test for SNP and SNP-environment interaction effects. (**A**, **B**) The tests for SNP-environment interaction effects with 100 and 400 overlapping individuals between any two studies. (**C**, **D**) The joint tests for SNP and SNP-environment interaction effects with 100 and 400 overlapping individuals between any two studies. The vertical axis is the—log10 (observed P value) from data analyzed under the null hypothesis, and the horizontal axis is the—log10 (expected P value).

SNP and SNP-environment joint effects, we assumed that all the fixed effects and random effects of the SNP and SNP-environment interaction were zero. That is, $\beta_G=\beta_{G\times E}=\gamma_G=\gamma_{G\times E}=0$.We calculated the likelihood ratio statistic L_J by estimating the minimum of the likelihood function $\widehat{l_0}$ and $\widehat{l_1}$. The empirical P-values were calculated as a ξ:0.5:(0.5-ξ) mixture of 3 df χ^2 distribution, 4 df χ^2 distribution, and 5 df χ^2 distribution. The value of ξ was data-dependent and calculated using Fisher information.

To test the powers of the SNP-environment interaction effect, we set the fixed effects of the SNP main β_G and SNP-environment interaction $\beta_{G\times E}$ to $\sqrt{0.002}$. The random effect of SNP main γ_G was normally distributed, followed $E \sim N(0, 0.015)$, variance of random effect of SNP-environment interaction ranging from 0.005 to 0.025, where each increased by 0.005. If the P-value of the test is less than 0.05, it was considered statistically significant. Experiments were repeated 1000 times, and the proportion of statistical significance was called empirical power. The OMR was also tested under this simulation. To test for the SNP and SNP-environment interaction joint effects, the fixed effects of β_G and $\beta_{G\times E}$ were set to a square root of 0.002. The random effects of γ_G and $\gamma_{G\times E}$ were set normally distributed with a mean of 0 and a variance ranging from 0.005 to 0.025.

Null distribution. As shown in Fig. 1A,B, these points are nearly standing on the diagonal line with 100 and 400 overlapping individuals between studies. This verifies that the method presented provides suitable distribu-

Individuals of study	Significance level	SNP main effect	
		$\beta_G = \sqrt{0.1}$	$\beta_G = \sqrt{0.2}$
1000	0.01	0.008	0.0156
	0.05	0.038	0.054
2000	0.01	0.012	0.011
	0.05	0.048	0.046

Table 1. The values of type I error rates at different scenarios for the test of SNP-environment interaction with 10% overlapping data.

Individuals of study	Significance level	SNP main effect	
		$\beta_G = \sqrt{0.1}$	$\beta_G = \sqrt{0.2}$
1000	0.01	0.011	0.009
	0.05	0.048	0.056
2000	0.01	0.011	0.006
	0.05	0.047	0.040

Table 2. The values of type I error rates at different scenarios for the test of SNP-environment interaction with 40% overlapping data.

Individuals of study	Significance level	
	0.01	0.05
1000	0.008	0.038
2000	0.010	0.040

Table 3. The values of type I error rates at different scenarios for the test of SNP and SNP-environment joint effects with 10% overlapping data.

Individuals of study	Significance level	
	0.01	0.05
1000	0.008	0.050
2000	0.011	0.047

Table 4. The values of type I error rates at different scenarios for the test of SNP and SNP-environment joint effects with 40% overlapping data.

tions. In Fig. 1C,D, the empirical P-values are close to the expected ones with 100 and 400 overlapping individuals between any two studies, demonstrating the suitability of our distributions.

Type I error rate. To better illustrate the performance of our method, we considered three different scenarios. In scenario 1, two different sample sizes were considered: (1) a study with 1000 individuals and (2) a study with 2000 individuals. In scenario 2, two different significance levels were considered: (1) 0.01 and (2) 0.05. In scenario 3, two different main effects of SNP were considered: (1) a square root of 0.1 and (2) a square root of 0.2. Table 1 presents the values of the type I error rates in the different scenarios for the test of the SNP-environment interaction with 10% overlapping data. Table 2 presents the values of the type I error rates in the different scenarios for the test of the SNP-environment interaction with 40% overlapping data. Table 3 presents the values of the type I error rates in the different scenarios for the test of the SNP and SNP-environment joint effects with 10% overlapping data. Table 4 presents the values of the type I error rates in the different scenarios for the test of the SNP and SNP-environment joint effects with 40% overlapping data. For the 1000 replications, the 95% confidence intervals for the estimated type I error rates of nominal levels 0.05 and 0.01are (0.036, 0.064) and (0.002, 0.018) respectively. For the 2000 replications, the 95% confidence intervals for the estimated type I error rates of nominal levels 0.05 and 0.01are (0.040, 0.060) and (0.004, 0.016) respectively. From these tables, we can see that all of the estimated type I error rates are in the confidence intervals for interaction tests and joint tests, this indicates that our method is valid.

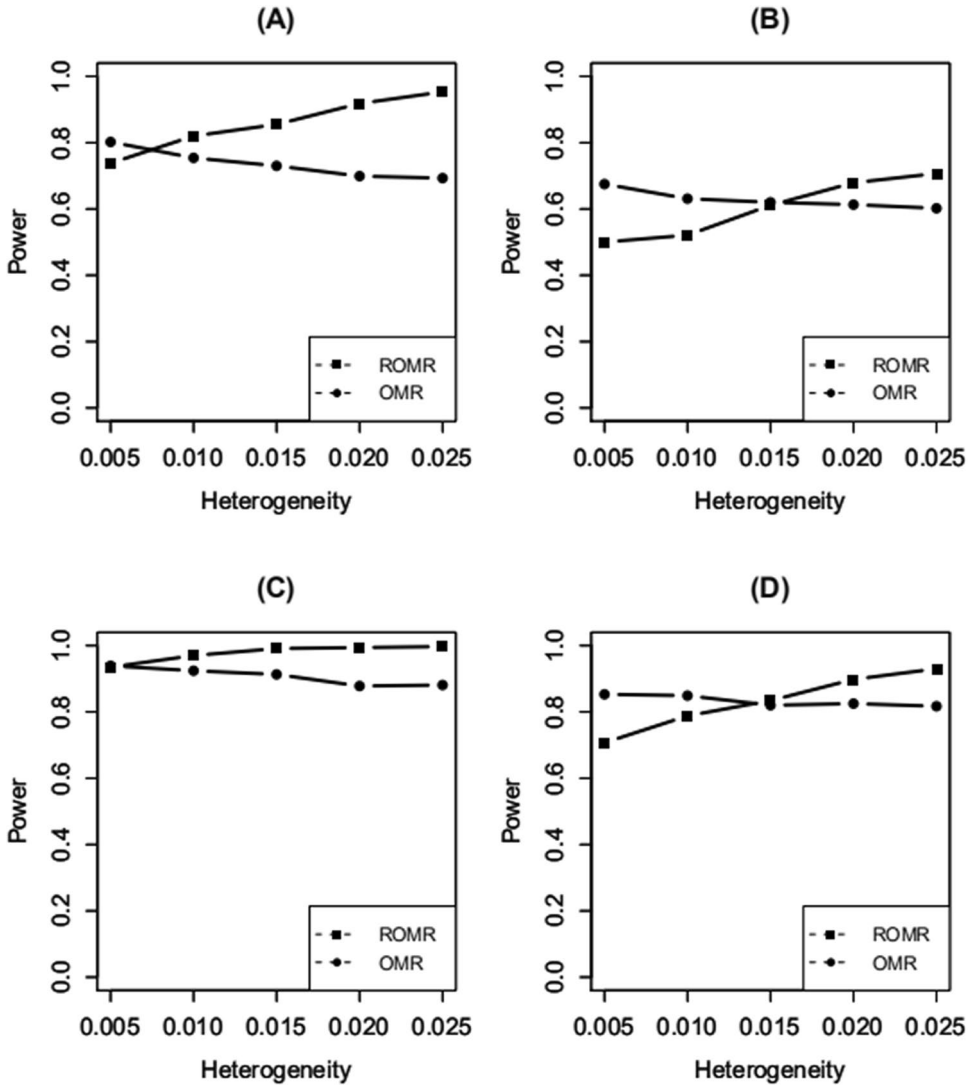

Figure 2. Statistical power of the test for SNP-environment interaction effects and the joint test for SNP and SNP-environment interaction effects. (**A**, **B**) The tests for SNP-environment interaction effects with 100 and 400 overlapping individuals between any two studies. (**C**, **D**) The joint tests for SNP and SNP-environment interaction effects with 100 and 400 overlapping individuals between any two studies. The vertical axis is the statistical power, and the horizontal axis is the heterogeneity effect.

Statistical power. The powers of ROMR and OMR are compared in Fig. 2A,B. The OMR gives higher powers when heterogeneity is low with 100 and 400 overlapping individuals between any two studies. The powers of this method decrease slightly with increase in heterogeneity, however, the powers of the ROMR increase rapidly with increase in heterogeneity. When the heterogeneity is high, the ROMR gives higher powers. This is due to the fact that when the heterogeneity is low, most of the statistical evidence of L_I is obtained from the fixed effect of the interaction. In fact, the test statistics of ROMR are penalized by high degrees of freedom, yielding less power. When the heterogeneity is high, the OMR tested for the fixed effect only, the genetic effect tested by the ROMR is much larger than that of OMR. Thus, ROMR gives higher power[29]. As shown in Fig. 2C,D, although our method provides a similar tendency to interaction simulation, joint tests generally obtain higher results. This is because both the SNP and SNP-environment interaction are tested, thereby including more effects than the test for the interaction only.

Table 5 and 6 present the powers under different levels of heterogeneity with different overlapping data. Table 5 shows the powers of the SNP-environment interaction, showing that the power decreased with an increase in the number of overlapping data. For ROMR, the greatest drop was 0.299; for OMR, the greatest drop was 0.127. However, in any case, when the heterogeneity was large, ROMR gave a higher power than OMR. Table 6 shows the powers of the SNP and SNP-environment interaction joint effects. Similar to the case of SNP-environment interaction, as the number of overlapping data increased, the power of ROMR was reduced faster than OMR. However, when the heterogeneity was large, ROMR gave a higher power than OMR. In order to more

Overlapping individuals	Methods	Heterogeneity				
		0.005	0.01	0.015	0.02	0.025
100	ROMR	0.738	0.820	0.855	0.918	0.953
	OMR	0.802	0.754	0.730	0.699	0.693
200	ROMR	0.622	0.701	0.797	0.832	0.890
	OMR	0.716	0.708	0.692	0.685	0.683
300	ROMR	0.519	0.617	0.670	0.748	0.793
	OMR	0.683	0.673	0.661	0.658	0.644
400	ROMR	0.500	0.521	0.611	0.679	0.707
	OMR	0.675	0.631	0.620	0.613	0.602

Table 5. The powers under different heterogeneity with different overlapping data for the test of SNP-environment interaction.

Overlapping individuals	Methods	Heterogeneity				
		0.005	0.01	0.015	0.02	0.025
100	ROMR	0.935	0.970	0.991	0.994	0.997
	OMR	0.939	0.924	0.913	0.878	0.881
200	ROMR	0.873	0.921	0.951	0.979	0.989
	OMR	0.926	0.910	0.903	0.898	0.876
300	ROMR	0.784	0.865	0.901	0.952	0.957
	OMR	0.898	0.875	0.867	0.861	0.833
400	ROMR	0.707	0.788	0.834	0.898	0.930
	OMR	0.853	0.849	0.820	0.825	0.817

Table 6. The powers under different heterogeneity with different overlapping data for the test of SNP and SNP-environment joint effects.

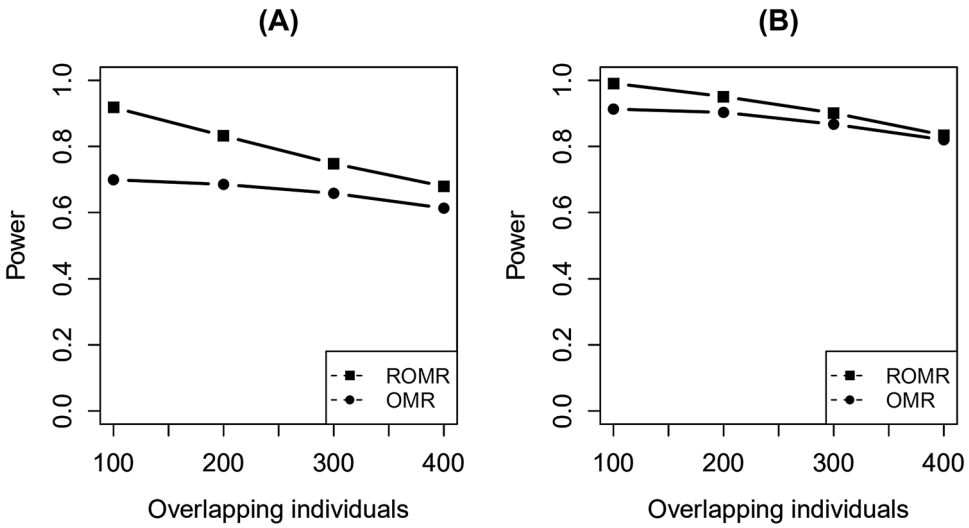

Figure 3. Statistical power of the test for SNP-environment interaction effects and the joint test for SNP and SNP-environment interaction effects with 100, 200, 300, 400 overlapping individuals under a fixed heterogeneity. (**A**) The tests for SNP-environment interaction effects with 100, 200, 300, 400 overlapping individuals under a fixed heterogeneity. (**B**) The joint tests for SNP and SNP-environment interaction effects with 100, 200, 300, 400 overlapping individuals under a fixed heterogeneity.

intuitively understand the impact of different overlapping data, Fig. 3 is given. We selected one set of parameters from Table 5 and 6. The variance of heterogeneity for SNP-environment interaction was fixed as 0.02 and the overlapping individuals were 100, 200, 300 and 400 as in the tables. As can be seen from Fig. 3A,B, with the increase of overlapping individuals, the powers of the two methods are decreasing gradually. The powers of

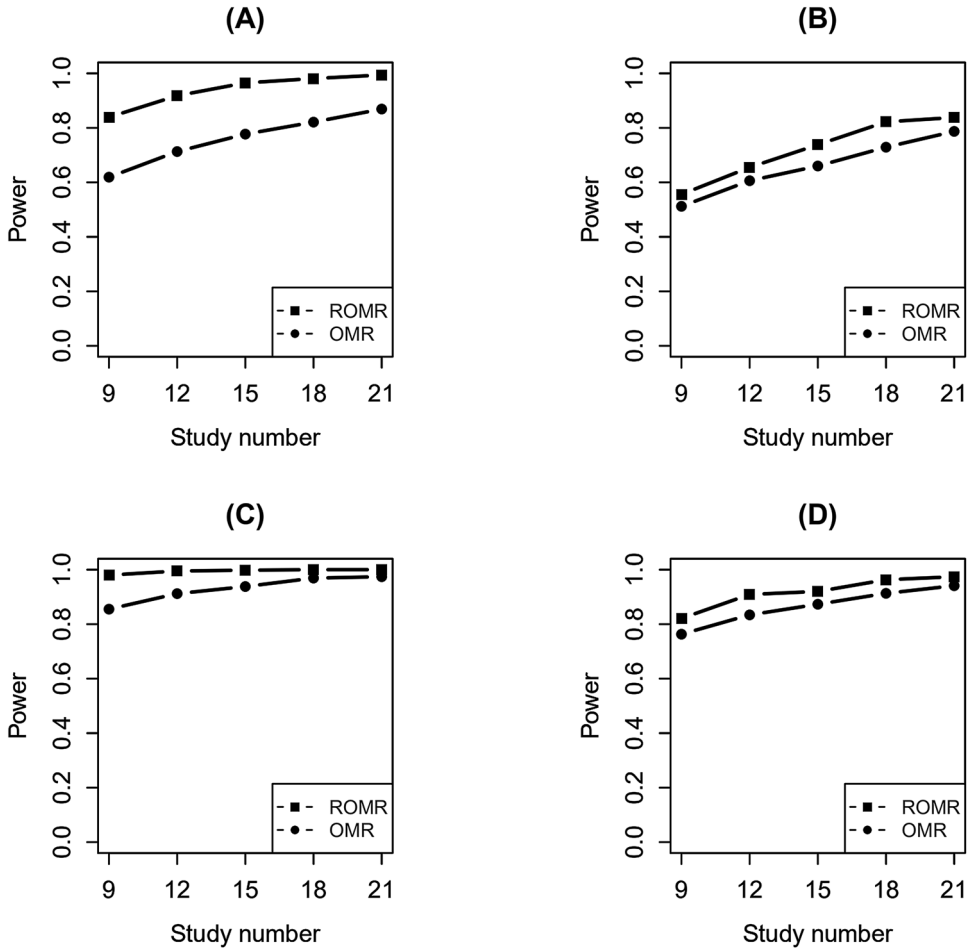

Figure 4. Statistical power of the test for SNP-environment interaction effects and the joint test for SNP and SNP-environment interaction effects with different number of studies (9, 12, 15, 18, 21) and with 100 and 400 overlapping individuals under a fixed heterogeneity. (**A**, **B**) The tests for SNP-environment interaction effects with different number of studies (9, 12, 15, 18, 21) and with 100 and 400 overlapping individuals under a fixed heterogeneity. (**C**, **D**) The joint tests for SNP and SNP-environment interaction effects with different number of studies (9, 12, 15, 18, 21) and with 100 and 400 overlapping individuals under a fixed heterogeneity.

SNP-environment interaction effects and the powers of the SNP and SNP-environment interaction joint effects have the same variation tendency.

Figure 4 present the powers under different number of studies. The variance of heterogeneity for SNP-environment was fixed as 0.02 and the numbers of studies were 9, 12, 15, 18, 21. Figure 4A,B present the powers of SNP-environment interaction effects under 100 and 400 overlapping individuals. Figure 4C,D present the powers of SNP and SNP-environment interaction joint effects under 100 and 400 overlapping individuals. As can be seen from these figures, with the increase of the numbers of studies, the powers of the two methods are increasing gradually.

Discussion

In contrast to the calculation of the likelihood function performed by Lee[17], we only changed the calculation of $\ln|V|$. The covariance matrix V is a real symmetric matrix that can be diagonalized in a similar manner. That is to say, V can be written in the form of $P^{-1}\Lambda P$, where P is the matrix of eigenvectors and Λ is the vector of eigenvalues. Then, $\ln|V| = \sum \ln|\lambda_i|$, where λ_i is the eigenvalue of the covariance matrix V. The other terms of the likelihood function are the same as in the random effect model MR. Thus, the computational complexity of our method is much less than that of Lee's method. We can also compute $\ln|V|$ using OMR or by combining both of the above mentioned methods.

As in Lee[17], our method can also be combined with OMR, providing a higher power, regardless of the level of heterogeneity. This method focuses on heterogeneity, it designs statistic as follows

$$L = \begin{cases} L_R & if \quad p_R \le p_F \\ 0 & p_R > p_F \end{cases}$$

where L_R is the likelihood ratio statistic for the SNP-environment interaction effect under the random effect model with overlapping data, and p_R and p_F are the P-values for test of the SNP-environment interaction effect applying the ROMR and OMR. The P-value for this statistic is similar to that used in Lee[17].

When the data between studies are independent, the correlation matrix C becomes an identity matrix; that is to say, C = I. In this context, the method becomes RMR. However, as a result of the additional judgement process of the correlation matrix, the calculation amount of this method is increased. Therefore, the use of RMR is recommended when the data between studies are independent, while ROMR is recommended when the data is overlapping or it is not certain whether there is overlapping data.

We performed a simulation where the number of overlapping individuals were increased from 100 to 200, 300, and 400. Simulations with 500 or more overlapping individuals were not performed because there are 1000 individuals in our studies. That is, if there are 500 or more overlapping individuals between studies, none of the studies have individuals that are not applied to other studies, thus the correlation matrix cannot be guaranteed to be strictly diagonally dominant. In this case, the non-singularity of the variance matrix cannot be guaranteed; thus, this situation is not considered.

In the present study, we simultaneously evaluated the fixed effect and random effect of the SNP-environment interaction or the fixed effect and random effect of the SNP and SNP-environment interaction. When heterogeneity is high and overlapping data exists, our method provides accurate and valid power. However, our method also has some limitations. First, the calculation cost of our ROMR is much higher than that of the OMR. Second, more than one environmental variable may interact with the genetic effect being tested. Here, only one environmental variable was chosen for the interaction analysis, but other environmental variables can be entered as covariates.

Conclusion

This study generalized the RMR proposed in our previous paper to account for overlapping data. This method was designed to test the SNP-environment interaction effect or the SNP and SNP-environment joint effects with overlapping data. To this end, a correlation matrix was introduced into our random effect model. In addition, a new method to solve the likelihood function was proposed, which allowed the solution to $\ln|V|$ to be obtained more easily. By simulation, we verified that our method was suitable under the conditions of the random effect model of the SNP-environment interaction with overlapping data. As a result, our ROMR obtained a higher power than OMR when the heterogeneity was high. In practice, when data from large-scale meta-analyses originate from different factors, including ethnicities, environments, phenotypes, or some other factors, heterogeneity across studies is likely to exist, our proposed ROMR can be applied.

Received: 24 August 2020; Accepted: 18 January 2021
Published online: 28 January 2021

References

1. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucl. Acids Res.* **45**, D896–D901 (2017).
2. Mannolio, T. A. Genomewide association studies and assessment of the risk of disease. *N. Engl. J.* **363**, 166–176 (2010).
3. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucl. Acids Res.* **42**, D1001–D1006 (2014).
4. DerSimonian, R. & Laird, N. Meta-analysis in clinical trials. *Control Clin. Trials* **7**, 177–188 (1986).
5. Evangelou, E., Ioannidis, J. P. Meta-analysis methods for genome-wide association studies and beyond.*Nat Rev Genet*. **14,** 379–389(2013).
6. Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. *Introduction to Meta-Analysis* 3–14 (Wiley, Hoboken, 2009).
7. Fleiss, J. The statistical basis of meta-analysis. *Stat. Methods Med. Res.* **2**, 121–145 (1993).
8. Field, A. P. The problems in using fixed-effects models of meta-analysis on real-world data. *Underst. Stat.* **2**, 105–124 (2003).
9. Zeggini, E. & Ioannidis, J. P. Meta-analysis in genome-wide association studies. *Pharmacogenomics* **10**, 191–201 (2009).
10. Lee, J. H. *et al.* Genetic susceptibility for chronic bronchitis in chronic obstructive pulmonary disease. *Respir. Res.* **15**, 113 (2014).
11. Kiryluk, K. *et al.* Geographic differences in genetic susceptibility to IgA nephropathy: GWAS replication study and geospatial risk analysis. *PLoS Genet.* **8**, e1002765. https://doi.org/10.1371/journal.pgen.1002765 (2012).
12. Kang, E. Y. *et al.* Meta-analysis identifies gene-by-environment interactions as demonstrated in a study of 4,965 mice. *PLoS Genet.* **10**, e1004022. https://doi.org/10.1371/journal.pgen.1004022 (2014).
13. Liu, J. Z. *et al.* Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).
14. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.* **9**, e1003491. https://doi.org/10.1371/journal.pgen.1003491 (2013).
15. Petersen, G. M. *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nat Genet.* **42**, 224–228 (2010).
16. Bhattacharjee, S. *et al.* A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *Am. J. Hum. Genet.* **90**, 821–835 (2012).
17. Lee, C. H., Eskin, E. & Han, B. Increasing the power of meta-analysis of genome-wide association studies to detect heterogeneous effects. *Bioinformatics* **33**, i379–i388 (2017).
18. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
19. Keller, M. F. *et al.* Trans-ethnic meta-analysis of white blood cell phenotypes. *Hum. Mol. Genet.* **23**, 6944–6960 (2014).
20. Hibar, D. P. *et al.* Genome-wide association identifies genetic variants associated with lentiform nucleus volume in N = 1345 young and elderly subjects. *Brain Imaging Behav.* **7**, 102–115 (2013).
21. Lin, D. Y. & Sullivan, P. F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am. J. Hum. Genet.* **85**, 862–872 (2009).
22. Han, B., Duong, D., Sul, J. H. & de, Bakker, P. I., Eskin, E., Raychaudhuri, S. ,. A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Hum. Mol. Genet.* **25**, 1857–1866 (2016).
23. Zaykin, D. V. & Kozbur, D. O. P-value based analysis for shared controls design in genome-wide association studies. *Genet. Epidemiol.* **34**, 725–738 (2010).

24. Wen, X. Bayesian model selection in complex linear systems, as illustrated in genetic association studies. *Biometrics* **70**, 73–83 (2014).
25. Kim, E. E. *et al.* FOLD: a method to optimize power in meta-analysis of genetic association studies with overlapping subjects. *Bioinformatics* **33**, 3947–3954 (2017).
26. Xu, X., Shi, G. & Nehorai, A. Meta-regression of gene-environment interaction in genome-wide association studies. *IEEE Trans. Nanobiosci.* **12**, 354–362 (2013).
27. Shi, G. & Nehorai, A. Robustness of meta-analyses in finding gene × environment interactions. *PLoS ONE* **12**, e0171446 (2017).
28. Jin, Q. & Shi, G. Meta-Analysis of SNP-Environment Interaction with Overlapping Data. *Front. Genet.* **10**, 1400. https://doi.org/10.3389/fgene.2019.01400 (2019).
29. Jin, Q. & Shi, G. Meta-analysis of SNP-environment interaction with heterogeneity. *Hum. Hered.* **84**, 117–126 (2019).
30. Wolfinger, R., Tobias, R. & Sall, J. Computing Gaussian likelihoods and their derivatives for general linear mixed models. *SIAM J. Sci. Comput.* **15**, 15–17 (1994).
31. Rao, C. R. Estimation of variance and covariance components in linear models. *J. Am. Stat. Assoc.* **67**, 112–115 (1972).
32. Gumedze, F. N. & Dunne, T. T. Parameter estimation and inference in the linear mixed model. *Linear Algebra Appl.* **435**, 1920–1944 (2011).
33. Jennrich, R. I. & Schluchter, M. D. Repeated-measures models with structured covariance matrices. *Biometrics* **4**, 805–820 (1986).
34. Lindstrom, M. J. & Bates, D. M. Newton–Raphson and EM algorithms for linear mixed-effects models for repeated measures data. *J. Am. Stat. Assoc.* **404**, 1014–1022 (1988).

Author contributions

Q.J.: conceived the concept, designed and conducted the simulation studies, and drafted the manuscript. G.S.: conceived the concept, supervised the work, reviewed and revised the manuscript.

Funding

This work was supported by the national Thousand Youth Talents Plan.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Q.J.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.