BMC Genomics

RESEARCH

Open Access

Divided-and-combined association test for pleiotropic effects with GWAS summary statistics

Boheng Hu[1†], Yaoyao Chen[1†], Xueqin Zheng[1] and Hongping Guo[1*]

Abstract

Many genome-wide association studies (GWAS) conducted over the past two decades have focused on testing the one-to-one association between genetic variants and complex diseases or phenotypes. Despite their considerable success, such one-to-one testing approaches are underpowered because they do not utilize information on pleiotropic effects. Currently available public GWAS summary statistics provide resources for multi-trait association testing. Quadratic statistics are widely used to detect pleiotropic effects, as they follow a chi-square or weighted chi-square distribution under the null hypothesis. Nevertheless, such methods often fail to achieve superior power. To address this limitation, this study proposes a divided-and-combined association test (DCAT) for detecting pleiotropic effects using GWAS summary statistics. DCAT constructs the covariance matrix of *Z*-scores by combining phenotype correlation and genotype partial correlation via the Kronecker product. It consists of a family of quadratic statistics with different covariance matrix weights and leverages the respective advantages of each statistic. The distribution of each statistic is then approximated using a location-shifted generalized gamma distribution for *p*-value calculation, with the resulting *p*-values integrated via the Cauchy combination test. Simulation results demonstrate that DCAT controls the type I error at a reasonable level and achieves the highest statistical power in most scenarios. In the real-data analysis of four cardiovascular diseases, DCAT identifies more significant genes compared with the other four representative methods.

Keywords Summary statistics, Pleiotropic effect, Divided-and-combined test, Association analysis, Cauchy combination test

Introduction

In the past two decades, genome-wide association studies (GWAS) have successfully identified hundreds of genetic variants associated with complex diseases or traits, providing valuable insights into the underlying genetic mechanisms [1]. However, the majority of these identified variants exert only weak effects and can explain only a small proportion of the heritability of phenotypes [2].

Traditional GWAS focus on testing the "one-to-one" association between phenotypes and variants, and such univariate association tests often have lower statistical power than multivarite tests because they do not

†Boheng Hu and Yaoyao Chen are joint first authors.

*Correspondence:
Hongping Guo
guohongping@hbnu.edu.cn
[1]School of Mathematics and Statistics, Hubei Normal University, Huangshi 435002, P.R. China

utilize information on pleiotropic effects. Pleiotropic effect exists when one genetic variant affects multiple phenotypes [3]. There is growing GWAS evidences that many variants have pleiotropic effects [4]. For example, the SNP (single-nucleotide polymorphism) rs2535629 was reported to be associated with five neuropsychiatric disorders, including autism spectrum disorder, attention deficit hyperactivity disorder, bipolar disorder, major depressive disorder, and schizophrenia [5]. Identifying these effects can provide valuable insights into the genetic architecture of complex traits [6, 7]. Therefore, it is of interest to identify pleiotropic effects by jointly analyzing multiple traits in GWAS.

Existing multivariate analysis methods are mainly classified into two categories. The first is the "one-to-multiple" scenario, which tests the association between one phenotype and multiple genetic variants, with representative methods including SKAT [8], ACAT [9], and sum-based tests [10]. The second is the "multiple-to-one" scenario, which tests the association between multiple phenotypes and one genetic variant, with typical methods such as TATES [11], MTAG [12], and principal component (PC) based tests [13].

Notably, due to the protection of individual privacy, only GWAS summary statistics are publicly available. These data include information such as genetic variant IDs, allele frequencies, genetic effects and their standard errors, and *p*-values [14]. This study focuses on "multiple-to-multiple" association tests based on summary statistics, i.e., analyzing the associations between multiple variants and multiple phenotypes. Several "multiple-to-multiple" association tests have been proposed in existing literature, such as metaCCA [15] based on canonical correlation, MGAS based on *p*-value combination [11], MAT [16] based on variance component test and burden test, and TWT which integrates genetic structure, pleiotropy, and potential information for testing [17]. Quadratic statistics are widely used to detect pleiotropic effects in GWAS summary statistic analysis of multiple phenotypes, as they follow a chi-square or weighted chi-square distribution under the null hypothesis [17–19]. For example, MuGenT employs a quadratic statistic to test the association between a set of SNPs and multiple genetic ancestries, and uses a gamma distribution for statistical inference [19]. However, such quadratic-based methods often fail to yield high statistical power.

To address this limitation, we propose a divide-and-combine association test (DCAT) for pleiotropic effects with GWAS summary statistics. The motivation of DCAT is shown in the Supplementary Note 1. Specifically, we perform DCAT via the following steps. Firstly, we construct the covariance matrix of *Z*-scores by combining phenotype correlation and genotype partial correlation via the Kronecker product, thereby capturing complex correlations in jointly analysis of multiple phenotypes. Furthermore, we integrate a family of quadratic statistics with different covariance matrix weights and leverages the respective advantages of each statistic. Then, we use a location-shifted generalized gamma distribution to approximate the distribution of these statistics, and integrate the resulting *p*-values via the Cauchy combination test. Finally, we conduct simulations and real data analysis to evaluate the performance of our proposed tests and other representative methods.

Methods

Notations

Consider a gene or pathway comprising m SNPs and q phenotypes. For the j-th phenotype ($j = 1, 2, \cdots, q$) and the k-th SNP ($k = 1, 2, \cdots, m$), let β_{jk} denote the effect size of the SNP and Z_{jk} denote the Z-score of the corresponding summary statistic. Z_{jk} is the Wald test statistic calculated as $Z_{jk} = \frac{\hat{\beta}_{jk}}{se(\hat{\beta}_{jk})}$, where $\hat{\beta}_{jk}$ and $se(\hat{\beta}_{jk})$ denote the estimated effect and its standard error, respectively. The effect sizes of the SNPs and the summary statistics can be expressed in matrix form as: $\boldsymbol{B} = (\beta_{jk})$ and $\boldsymbol{Z} = (Z_{jk})$ $(j = 1, 2, \cdots, q; \quad k = 1, 2, \cdots, m)$, respectively.

To investigate the association between the candidate gene or pathway and multiple phenotypes, we define the null hypothesis H_0 as: none of the m SNPs is associated with any of the q phenotypes. The alternative hypothesis H_1 is defined as: at least one of the m SNPs is associated with at least one of the q phenotypes. Denote $\boldsymbol{\beta} = \mathrm{vec}(\boldsymbol{B})$, where $\mathrm{vec}(\cdot)$ represents the column-wise vectorization operator for a matrix. Then the hypothesis test of interest can be written as:

$$H_0 : \boldsymbol{\beta} = \mathbf{0}_{qm} \longleftrightarrow H_1 : \boldsymbol{\beta} \neq \mathbf{0}_{qm}.$$

where $\mathbf{0}_{qm}$ denotes a qm-dimensional zero vector.

Let $\boldsymbol{Z}_* = \mathrm{vec}(\boldsymbol{Z})$, under H_0, existing study [17] has shown that $\boldsymbol{Z}_*$ follows a multivariate normal distribution with a mean of zero and an unknown $qm \times qm$ covariance matrix $\boldsymbol{\Delta}$, i.e., $\boldsymbol{Z}_* \sim \mathcal{N}(\mathbf{0}_{qm}, \boldsymbol{\Delta})$.

Estimation of covariance matrix

Consider two Z-scores Z_{jk} and $Z_{j'k'}$ derived from the summary statistics. The correlation coefficient between Z_{jk} and $Z_{j'k'}$ is expressed as [17]:

$$\mathrm{corr}(Z_{jk}, Z_{j'k'}) = \rho_{jj'}\theta_{kk'}, \tag{1}$$

where $\rho_{jj'}$ denotes the correlation coefficient between phenotypes Y_j and $Y_{j'}$, and $\theta_{kk'}$ denotes the partial correlation coefficient between SNP

genotypes G_k and $G_{k'}$ after adjusting for the covariate $\boldsymbol{C}$ ($j, j' = 1, 2, \cdots, q$; $k, k' = 1, 2, \cdots, m$). Here, $\boldsymbol{C}$ represents a matrix of genetic principal components (PCs) used to account for population stratification.

Estimation of $\rho_{jj'}$

The estimation method for $\rho_{jj'}$ has been well established in existing literature [20, 21], and is described as follows: For L independent null SNPs [22], let $(z_{jl}, z_{j'l})$ represent the pairs of Z-scores corresponding to phenotypes Y_j and $Y_{j'}$ ($l = 1, 2, \cdots, L$). The estimator of $\rho_{jj'}$ is given by:

$$\hat{\rho}_{jj'} = \frac{\sum_{l=1}^{L} (z_{jl} - \bar{z}_j)(z_{j'l} - \bar{z}_{j'})}{\sqrt{\sum_{l=1}^{L} (z_{jl} - \bar{z}_j)^2} \sqrt{\sum_{l=1}^{L} (z_{j'l} - \bar{z}_{j'})^2}}, \quad (2)$$

where $\bar{z}_j = \frac{1}{L}\sum_{l=1}^{L} z_{jl}$ and $\bar{z}_{j'} = \frac{1}{L}\sum_{l=1}^{L} z_{j'l}$ denote the sample means of z_{jl} and $z_{j'l}$, respectively. As described in [20], if the two phenotypes are from different cohorts, then a correlation can be induced only by either overlapping samples or related subjects in the two cohorts. In either case, the $\rho_{jj'}$ estimator can capture the correlation. We prove the asymptotic unbiasedness of the $\rho_{jj'}$ estimator in the Supplementary Note 2.

Estimation of $\theta_{kk'}$

To estimate the partial correlation coefficient $\theta_{kk'}$, we use publicly available reference population datasets (e.g., the 1000 Genomes Project [23], or HapMap3 [24]) for individual-level analysis. Let the reference dataset contain N subjects, the genotype vectors for SNPs G_k and $G_{k'}$ are denoted as $\boldsymbol{g}_k = (g_{1k}, g_{2k}, \cdots, g_{Nk})^\top$ and $\boldsymbol{g}_{k'} = (g_{1k'}, g_{2k'}, \cdots, g_{Nk'})^\top$, respectively, where $^\top$ denotes the transpose of a vector or matrix.

We first extract the top 2 PCs of the $N \times M$ genotype matrix from the reference dataset using the PCoC function in the R package AssocTest, where M denotes the number of SNPs included in the reference dataset. These 2 PCs, combined with a vector of ones (representing the intercept term), constitute the covariate matrix $\boldsymbol{C}$. To account for the effect of $\boldsymbol{C}$, we regress the genotype vectors $\boldsymbol{g}_k$ and $\boldsymbol{g}_{k'}$ on $\boldsymbol{C}$, and obtain the following residual vectors:

$$\begin{aligned} \boldsymbol{h}_k &= \left(\boldsymbol{I}_N - \boldsymbol{C}\left(\boldsymbol{C}^\top\boldsymbol{C}\right)^{-1}\boldsymbol{C}^\top\right)\boldsymbol{g}_k, \\ \boldsymbol{h}_{k'} &= \left(\boldsymbol{I}_N - \boldsymbol{C}\left(\boldsymbol{C}^\top\boldsymbol{C}\right)^{-1}\boldsymbol{C}^\top\right)\boldsymbol{g}_{k'}, \end{aligned} \quad (3)$$

where $\boldsymbol{I}_N$ is the $N \times N$ identity matrix, and $\boldsymbol{h}_k = (h_{1k}, h_{2k}, \cdots, h_{Nk})^\top$, $\boldsymbol{h}_{k'} = (h_{1k'}, h_{2k'}, \cdots, h_{Nk'})^\top$ are the covariate-adjusted residual vectors for $\boldsymbol{g}_k$ and $\boldsymbol{g}_{k'}$, respectively.

The estimator of the partial correlation coefficient $\theta_{kk'}$ is then calculated as:

$$\hat{\theta}_{kk'} = \frac{\sum_{i=1}^{N} (h_{ik} - \bar{h}_k)(h_{ik'} - \bar{h}_{k'})}{\sqrt{\sum_{i=1}^{N} (h_{ik} - \bar{h}_k)^2} \sqrt{\sum_{i=1}^{N} (h_{ik'} - \bar{h}_{k'})^2}}, \quad (4)$$

where $\bar{h}_k = \frac{1}{N}\sum_{i=1}^{N} h_{ik}$ and $\bar{h}_{k'} = \frac{1}{N}\sum_{i=1}^{N} h_{ik'}$ denote the sample means of h_{ik} and $h_{ik'}$, respectively.

Expression of covariance matrix $\boldsymbol{\Delta}$

The final covariance matrix $\boldsymbol{\Delta}$ is defined as:

$$\boldsymbol{\Delta} = \boldsymbol{\Delta}_\rho \otimes \boldsymbol{\Delta}_\theta, \quad (5)$$

where $\otimes$ denotes the Kronecker product operator. Specifically, $\boldsymbol{\Delta}_\rho$ is the phenotype correlation matrix, whose (j, j')-th entry corresponds to the correlation coefficient between phenotypes Y_j and $Y_{j'}$, and it will account for GWAS sample overlap. $\boldsymbol{\Delta}_\theta$ is the genotype partial correlation matrix, whose (k, k')-th entry corresponds to the partial correlation coefficient between SNPs G_k and $G_{k'}$ after covariate adjustment for $\boldsymbol{C}$.

The explicit forms of $\boldsymbol{\Delta}_\rho$ and $\boldsymbol{\Delta}_\theta$ are given by:

$$\boldsymbol{\Delta}_\rho = \begin{pmatrix} 1 & \hat{\rho}_{12} & \cdots & \hat{\rho}_{1q} \\ \hat{\rho}_{12} & 1 & \cdots & \hat{\rho}_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\rho}_{1q} & \hat{\rho}_{2q} & \cdots & 1 \end{pmatrix}, \quad \boldsymbol{\Delta}_\theta = \begin{pmatrix} 1 & \hat{\theta}_{12} & \cdots & \hat{\theta}_{1m} \\ \hat{\theta}_{12} & 1 & \cdots & \hat{\theta}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\theta}_{1m} & \hat{\theta}_{2m} & \cdots & 1 \end{pmatrix}, \quad (6)$$

where $\hat{\rho}_{jj'}$ (from Eq. (2)) and $\hat{\theta}_{kk'}$ (from Eq. (4)) are the estimated correlation coefficients defined in the preceding sections.

Divided-and-combined test statistic

Quadratic statistics are widely used for association testing using GWAS summary statistics across multiple phenotypes [17–19, 21, 25]. The classical chi-square test $\mathrm{CHI} = \boldsymbol{Z}_*^\top \boldsymbol{\Delta}^{-1} \boldsymbol{Z}_*$ follows a chi-square distribution with qm degrees of freedom under the null hypothesis. It is an omnibus test suitable for most scenarios, but not for dense signals. When qm is large, CHI suffers from a loss of power due to the heavy-tailed nature of the chi-square distribution with high degrees of freedom [18], which may limit its ability to detect pleiotropic effects. The variance component score test $\mathrm{VC} = \boldsymbol{Z}_*^\top \boldsymbol{\Delta}^{-1} \boldsymbol{\Delta}^{-1} \boldsymbol{Z}_*$ exhibits greater power than CHI to detect heterogeneous effects when the last eigenvector of $\boldsymbol{\Delta}$ contains elements of different signs and magnitudes in pratical settings [21]. The sum of squared score test, defined as $\mathrm{SSU} = \boldsymbol{Z}_*^\top \boldsymbol{Z}_*$, is well suited for detecting heterogeneous effects and dense signals, particularly under weak phenotypic correlation [25]. Next, we provide a simple example to demonstrate that the covariance quadratic test $\mathrm{CQ} = \boldsymbol{Z}_*^\top \boldsymbol{\Delta} \boldsymbol{Z}_*$ can be more

powerful than CHI in some cases. Consider the scenario with two phenotypes and a gene that contains only one SNP, i.e., $q = 2$ and $m = 1$. Let the Z-scores of the two phenotypes be Z_1 and Z_2, so that $\boldsymbol{Z}_* = (Z_1, Z_2)^\top$. The correlation matrix between the two Z-scores is denoted by $\boldsymbol{\Delta} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, $0 < \rho < 1$. Consider the two quadratic statistics as follows: $T^{(-1)} = \boldsymbol{Z}_*^\top \boldsymbol{\Delta}^{-1} \boldsymbol{Z}_*$ (also called CHI above) and $T^{(1)} = \boldsymbol{Z}_*^\top \boldsymbol{\Delta} \boldsymbol{Z}_*$ (also called CQ above). We show that the statistic $T^{(1)}$ has higher power than the traditional one $T^{(-1)}$ in this scenario. The detailed derivation and computational procedure are provided in the Supplementary Note 1. Here we only give a brief explanation. Under H_0, $T^{(-1)} \overset{d}{=} \zeta_1 + \zeta_2$, and $T^{(1)} \overset{d}{=} (1+\rho)^2\zeta_1 + (1-\rho)^2\zeta_2$, where $\overset{d}{=}$ denotes equality in distribution, ζ_1, ζ_2 are independent chi-square random variables with 1 degree of freedom, i.e., $\zeta_1 \sim \chi^2(1)$ and $\zeta_2 \sim \chi^2(1)$. Under H_1, let $Z_1 = d_1$ and $Z_2 = d_2$ (d_1, d_2 are not both zero), then $T^{(-1)} = \frac{d_1^2 + d_2^2}{1+\rho}$, and $T^{(1)} = (d_1^2 + d_2^2)(1+\rho)$. The tail probabilities for the two test statistics are given by

$$
\begin{aligned}
p_{-1} &= P\left(\zeta_1 + \zeta_2 > \frac{d_1^2 + d_2^2}{1+\rho}\right), \\
p_1 &= P\left((1+\rho)^2\zeta_1 + (1-\rho)^2\zeta_2 > (d_1^2 + d_2^2)(1+\rho)\right) \\
&= P\left(\zeta_1 + \zeta_2 > \frac{d_1^2 + d_2^2}{1+\rho} + \frac{4\rho}{(1+\rho)^2}\zeta_2\right).
\end{aligned}
$$

For $\rho > 0$, we have $p_{-1} > p_1$. This indicates that the CHI is not powerful than CQ in this case. From the analysis of the above example, we can see that under H_0, quadratic statistics can be expressed as a weighted sum of mutually independent $\chi^2(1)$ random variables by performing an eigenvalue decomposition on the positive semi-definite symmetric matrix in the middle of the quadratic form. Therefore, by appropriately adjusting the weights, the power of the tests can be improved. For this reason, we propose a divided-and-combined strategy that integrates a family of quadratic statistics, with the four aforementioned quadratic statistics (CHI, VC, SSU, CQ) as special cases. Each statistic in this family has distinct advantages. To leverage their respective strengths, we employ the Cauchy transformation to combine several high-power quadratic statistics, thereby constructing a Cauchy statistic. By approximating its null distribution using the Cauchy distribution, we can explicitly compute the p-values of the test statistics.

Suppose the covariance matrix $\boldsymbol{\Delta}$ is symmetric and positive semi-definite. Let its eigenvalue decomposition be $\boldsymbol{\Delta} = \sum_{v=1}^{M} \lambda_v \boldsymbol{q}_v \boldsymbol{q}_v^\top$, where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M$ are the eigenvalues of $\boldsymbol{\Delta}$, with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{M_0} > \lambda_{M_0+1} = \cdots = \lambda_M = 0$ for some $M_0 \leq M$; $\boldsymbol{q}_1, \boldsymbol{q}_2, \cdots, \boldsymbol{q}_M$ are the corresponding orthonormal eigenvectors.

For any $\omega \in \mathbb{R}$, the matrix power $\boldsymbol{\Delta}^\omega$ can be expressed via the eigenvalue decomposition as: $\boldsymbol{\Delta}^\omega = \sum_{v=1}^{M} \lambda_v^\omega \boldsymbol{q}_v \boldsymbol{q}_v^\top$. We define the following statistic parameterized by ω:

$$
T^{(\omega)} = \boldsymbol{Z}_*^\top \boldsymbol{\Delta}^\omega \boldsymbol{Z}_*, \tag{7}
$$

which is asymptotically distributed as $\sum_{v=1}^{M_0} \lambda_v^{\omega+1} \zeta_v$ [26], i.e.,

$$
T^{(\omega)} \overset{d}{=} \sum_{v=1}^{M_0} \lambda_v^{\omega+1} \zeta_v, \tag{8}
$$

where $\zeta_v \sim \chi^2(1)$ (independent chi-square random variables with 1 degree of freedom). We select ω from the set $\Omega = \left\{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2\right\}$, following the choice from the previous studies [26, 27]. In fact, $T^{(-2)}$ corresponds to the VC statistic, $T^{(-1)}$ to the CHI statistic, $T^{(0)}$ to the SSU statistic, and $T^{(1)}$ to the CQ statistic. For each $\omega \in \Omega$, the corresponding p-value is defined as:

$$
p_\omega = P\left(\sum_{v=1}^{M_0} \lambda_v^{\omega+1} \zeta_v > t_\omega\right), \tag{9}
$$

where t_ω is the observed value of $T^{(\omega)}$.

We then combine these p-values using the Cauchy combination test [28], yielding the following combined statistic:

$$
T = \sum_{\omega \in \Omega} \frac{1}{|\Omega|} \tan\left((0.5 - p_\omega)\,\pi\right), \tag{10}
$$

where $|\Omega|$ denotes the cardinality of Ω (i.e., $|\Omega| = 7$), and $\tan(\cdot)$ is the tangent function. Under H_0, T follows a standard Cauchy distribution. For an observed value t of T, its p-value is calculated as:

$$
P(T > t) = \frac{1}{2} - \frac{\arctan(t)}{\pi}. \tag{11}
$$

Calculation of p_ω

Since the test statistic $T^{(\omega)}$ follows the same distribution as a weighted sum of M_0 independent $\chi^2(1)$ random variables (i.e., $\sum_{v=1}^{M_0} \lambda_v^{\omega+1} \zeta_v$), its cumulative distribution function (CDF) is not analytically tractable. We therefore approximate this distribution using a location-shifted generalized gamma distribution. Let $\mathcal{G}(\xi, \eta, s)$ denote the CDF of a standard generalized gamma distribution with

shape parameters η, s and scale parameter ξ, its probability density function (PDF) is given by:

$$g(x;\xi,\eta,s)=\frac{\eta x^{\eta s-1}}{\xi^{\eta s}\Gamma(s)}\exp\left(-\left(\frac{x}{\xi}\right)^{\eta}\right),\quad x>0, \tag{12}$$

where $\Gamma(s)=\int_0^\infty t^{s-1}e^{-t}dt$(for $s>0$) is the gamma function.

To improve approximation flexibility, we introduce a location parameter r (fixed but unknown) and use the CDF of $X+r$ (where $X\sim\mathcal{G}(\xi,\eta,s)$) to approximate the CDF of $T^{(\omega)}$. The parameters ξ,η,s,r are estimated by matching the first four cumulants of $X+r$ to those of $\sum_{v=1}^{M_0}\lambda_v^{\omega+1}\zeta_v$.

The τ-th cumulants of $X+r$ ($\tau=1,2,3,4$) are:

$$\begin{aligned}
\kappa_1(X+r)&=\xi\cdot\frac{\Gamma\left(\frac{1}{\eta}+s\right)}{\Gamma(s)}+r,\\
\kappa_2(X+r)&=\xi^2\left(\frac{\Gamma\left(\frac{2}{\eta}+s\right)}{\Gamma(s)}-\left(\frac{\Gamma\left(\frac{1}{\eta}+s\right)}{\Gamma(s)}\right)^2\right),\\
\kappa_3(X+r)&=\xi^3\left(\frac{\Gamma\left(\frac{3}{\eta}+s\right)}{\Gamma(s)}-3\frac{\Gamma\left(\frac{2}{\eta}+s\right)\Gamma\left(\frac{1}{\eta}+s\right)}{\Gamma(s)^2}\right.\\
&\quad\left.+2\left(\frac{\Gamma\left(\frac{1}{\eta}+s\right)}{\Gamma(s)}\right)^3\right),\\
\kappa_4(X+r)&=\xi^4\left(\frac{\Gamma\left(\frac{4}{\eta}+s\right)}{\Gamma(s)}-4\frac{\Gamma\left(\frac{3}{\eta}+s\right)\Gamma\left(\frac{1}{\eta}+s\right)}{\Gamma(s)^2}\right.\\
&\quad-3\left(\frac{\Gamma\left(\frac{2}{\eta}+s\right)}{\Gamma(s)}\right)^2+12\frac{\Gamma\left(\frac{2}{\eta}+s\right)\Gamma\left(\frac{1}{\eta}+s\right)^2}{\Gamma(s)^3}\\
&\quad\left.-6\left(\frac{\Gamma\left(\frac{1}{\eta}+s\right)}{\Gamma(s)}\right)^4\right),
\end{aligned} \tag{13}$$

where $\kappa_\tau(\cdot)$ denotes the τ-th cumulant operator.

For the weighted sum $\sum_{v=1}^{M_0}\lambda_v^{\omega+1}\zeta_v$, its τ-th cumulants are derived as:

$$\kappa_\tau\left(\sum_{v=1}^{M_0}\lambda_v^{\omega+1}\zeta_v\right)=2^{\tau-1}(\tau-1)!\sum_{v=1}^{M_0}\lambda_v^{\tau(\omega+1)},\quad\tau=1,2,3,4. \tag{14}$$

The unknown parameters $\{\xi,\eta,s,r\}$ are estimated by minimizing the sum of normalized squared errors between the two sets of cumulants. The optimization objective is:

$$(\hat{\xi},\hat{\eta},\hat{s},\hat{r})=\arg\min_{\xi,\eta,s,r}\sum_{\tau=1}^{4}\left(\frac{\kappa_\tau\left(\sum_{v=1}^{M_0}\lambda_v^{\omega+1}\zeta_v\right)-\kappa_\tau(X+r)}{\max\left(\left|\kappa_\tau\left(\sum_{v=1}^{M_0}\lambda_v^{\omega+1}\zeta_v\right)\right|,\varsigma\right)}\right)^2, \tag{15}$$

where ς is a small positive constant (to avoid division by zero when the cumulant value is near zero), and the numerator quantifies the discrepancy between the target cumulants (from Eq. (14)) and the approximate cumulants (from Eq. (13)).

Using the estimated parameters $(\hat{\xi},\hat{\eta},\hat{s},\hat{r})$, the p-value p_ω is computed as the upper-tail probability of the approximated distribution:

$$\begin{aligned}
p_\omega&=P\left(\sum_{v=1}^{M_0}\lambda_v^{\omega+1}\zeta_v>t_\omega\right)=P(X+\hat{r}>t_\omega)\\
&=\int_{t_\omega-\hat{r}}^{\infty}\frac{\hat{\eta}\,x^{\hat{\eta}\hat{s}-1}}{\hat{\xi}^{\hat{\eta}\hat{s}}\Gamma(\hat{s})}\exp\left(-\left(\frac{x}{\hat{\xi}}\right)^{\hat{\eta}}\right)dx=\frac{\Gamma\left(\hat{s},\left(\frac{t_\omega-\hat{r}}{\hat{\xi}}\right)^{\hat{\eta}}\right)}{\Gamma(\hat{s})},
\end{aligned} \tag{16}$$

where t_ω is the observed value of $T^{(\omega)}$, and $\Gamma(s,\phi)=\int_\phi^\infty t^{s-1}e^{-t}dt$ (for $s,\phi>0$) is the upper incomplete gamma function. This calculation applies to all $\omega\in\Omega=\{-2,-1,-\frac{1}{2},0,\frac{1}{2},1,2\}$.

Simulation study

To evaluate the performance of the proposed DCAT method, we conducted simulation studies where we benchmarked its type I error rate and statistical power against four representative summary statistics-based methods: TWT [17], MGAS [11], metaCCA [15], and MAT [16]. The core principles of these competing methods are as follows: TWT uses a Cauchy combination strategy to integrate three different kinds of signals, thereby realizing robust detection of pleiotropic effects in association tests; MGAS improves the min-p test to combine SNP-set association p-values derived from unvariate analyses with adjustments for inter-phenotype correlations; metaCCA leverages Canonical Correlation Analysis (CCA) to test the joint effect of multiple SNPs across multiple phenotypes; MAT integrates variance component tests and burden tests to detect gene-level pleiotropic effects in multi-trait SNP-set analyses.

Simulation setting

In this study, genotype data are obtained from the 1000 Genomes Project [23], which includes 2504 samples and 196036 SNPs. Phenotype data are generated as the following model:

$$\begin{aligned}
y_{ij}&=\alpha_j+\sum_{k=1}^{m}g_{ik}\beta_{jk}+c_i\gamma_j+\epsilon_{ij},\\
i&=1,2,\cdots,n,\ \ j=1,2,\cdots,q,
\end{aligned} \tag{17}$$

where α_j denotes the intercept term for the j-th phenotype; g_{ik} represents the genotype of the k-th SNP in the i-th individual. β_{jk} is the effect coefficient of g_{ik}, and the coefficient matrix is denoted as $\boldsymbol{B}=(\beta_{jk})\quad(j=1,2,\cdots,q;\ k=1,2,\cdots,m)$; The

Table 1 Simulation scenarios with different phenotype covariance structures and effect patterns under H_0

Scenario	No. of Phenotypes (q)	Covariance ($\boldsymbol{\Delta}_\rho$)	Effect Pattern
S_{1-1}	10	AR ($\rho = 0.2$)	NONE
S_{1-2}		AR ($\rho = 0.4$)	
S_{1-3}		AR ($\rho = 0.6$)	
S_{2-1}	10	ARCS ($\rho = 0.2, \delta = 0.01$)	NONE
S_{2-2}		ARCS ($\rho = 0.4, \delta = 0.015$)	
S_{2-3}		ARCS ($\rho = 0.6, \delta = 0.005$)	
S_{3-1}	20	AR ($\rho = 0.2$)	NONE
S_{3-2}		AR ($\rho = 0.4$)	
S_{3-3}		AR ($\rho = 0.6$)	
S_{4-1}	20	ARCS ($\rho = 0.2, \delta = 0.01$)	NONE
S_{4-2}		ARCS ($\rho = 0.4, \delta = 0.015$)	
S_{4-3}		ARCS ($\rho = 0.6, \delta = 0.005$)	

AR represents autoregressive covariance structure, and ARCS represents autoregressive compound symmetric structure

covariate vector c_i follows a standard normal distribution $N(0, 1)$, with a fixed effect coefficient $\gamma_j = 1$; ϵ_{ij} is the error term. Let $\boldsymbol{\varepsilon} = (\epsilon_{ij})$, $\mathrm{vec}(\boldsymbol{\varepsilon}) \sim \mathcal{N}(\mathbf{0}_{nq}, \boldsymbol{I}_n \otimes \boldsymbol{\Delta}_\rho)$, where $\boldsymbol{\Delta}_\rho$ is defined in Eq. (6).

Phenotype data y_{ij} are generated using Eq. (17). Subsequently, ordinary least squares estimation is applied to obtain the estimated effect $\hat{\beta}_{jk}$ and its standard error $se\left(\hat{\beta}_{jk}\right)$, the Wald test statistic is then calculated as $Z_{jk} = \frac{\hat{\beta}_{jk}}{se\left(\hat{\beta}_{jk}\right)}$. The parameters varied across simulation scenarios included m (number of SNPs), $\boldsymbol{\Delta}_\theta$ (genotype partial correlation matrix), q (number of phenotypes), and $\boldsymbol{\Delta}_\rho$ (phenotype correlation matrix).

Simulations are based on the ACOT7 gene, which contains 6 SNPs (i.e., $m = 6$) along with its corresponding genotype correlation matrix $\boldsymbol{\Delta}_\theta$. The number of phenotypes are fixed at $q = 10$ and $q = 20$. And two types of phenotype correlation structures $\boldsymbol{\Delta}_\rho$ are considered for scenario design:

(1) Autoregressive (AR) matrix: $\boldsymbol{\Delta}_\rho = \left(\rho^{|i-j|}\right)_{q\times q}$, ρ denotes the correlation decay parameter, and it is set to 0.2, 0.4, and 0.6 to describe different phenotype correlation strength.

(2) Autoregressive Compound Symmetric (ARCS) matrix: $\boldsymbol{\Delta}_\rho = \left(\rho^{|i-j|} + \delta\right)_{q\times q}$, δ represents the offset parameter, and the two parameters are paired as $(\rho, \delta) = (0.2, 0.01)$, (0.4, 0.015), and (0.6, 0.005).

On one hand, we evaluate the type I error rate under the null hypothesis $H_0 : \boldsymbol{B} = \mathbf{0}$ (denoted as NONE). On the other hand, we calculate the statistical power under the alternative hypothesis H_1 with "multiple-to-multiple" genetic architectures, where multiple SNPs in the investigated gene or pathway jointly contribute to multiple phenotypes.

Table 2 Simulation scenarios with different phenotype covariance structures and effect patterns under H_1

Scenario	No. of Phenotypes (q)	Covariance ($\boldsymbol{\Delta}_\rho$)	Effect Pattern
S_{1-1-1}	10	AR ($\rho = 0.2$)	Sparse (30%)
S_{1-1-2}			Moderate (50%)
S_{1-1-3}			Dense (70%)
S_{1-2-1}		AR ($\rho = 0.4$)	Sparse (30%)
S_{1-2-2}			Moderate (50%)
S_{1-2-3}			Dense (70%)
S_{1-3-1}		AR ($\rho = 0.6$)	Sparse (30%)
S_{1-3-2}			Moderate (50%)
S_{1-3-3}			Dense (70%)
S_{2-1-1}	10	ARCS ($\rho = 0.2, \delta = 0.01$)	Sparse (30%)
S_{2-1-2}			Moderate (50%)
S_{2-1-3}			Dense (70%)
S_{2-2-1}		ARCS ($\rho = 0.4, \delta = 0.015$)	Sparse (30%)
S_{2-2-2}			Moderate (50%)
S_{2-2-3}			Dense (70%)
S_{2-3-1}		ARCS ($\rho = 0.6, \delta = 0.005$)	Sparse (30%)
S_{2-3-2}			Moderate (50%)
S_{2-3-3}			Dense (70%)
S_{3-1-1}	20	AR ($\rho = 0.2$)	Sparse (30%)
S_{3-1-2}			Moderate (50%)
S_{3-1-3}			Dense (70%)
S_{3-2-1}		AR ($\rho = 0.4$)	Sparse (30%)
S_{3-2-2}			Moderate (50%)
S_{3-2-3}			Dense (70%)
S_{3-3-1}		AR ($\rho = 0.6$)	Sparse (30%)
S_{3-3-2}			Moderate (50%)
S_{3-3-3}			Dense (70%)
S_{4-1-1}	20	ARCS ($\rho = 0.2, \delta = 0.01$)	Sparse (30%)
S_{4-1-2}			Moderate (50%)
S_{4-1-3}			Dense (70%)
S_{4-2-1}		ARCS ($\rho = 0.4, \delta = 0.015$)	Sparse (30%)
S_{4-2-2}			Moderate (50%)
S_{4-2-3}			Dense (70%)
S_{4-3-1}		ARCS ($\rho = 0.6, \delta = 0.005$)	Sparse (30%)
S_{4-3-2}			Moderate (50%)
S_{4-3-3}			Dense (70%)

AR represents autoregressive covariance structure, and ARCS represents the autoregressive compound symmetric structure

In addition, three levels of phenotypic effect sparsity (30%, 50%, 70%) are considered to characterize the heterogeneity of pleiotropic effects. For example, when $q = 10$ (the total number of phenotypes), sparsity levels of 30%, 50%, and 70% indicate that the gene affects 3, 5, and 7 phenotypes, respectively. Detailed configurations of all simulation scenarios are summarized in Tables 1 and 2. To ensure the reliability and stability of the results, each scenario is replicated 10^4 times. For hypothesis testing, three commonly used significance levels ($\alpha = 0.05, 0.01, 0.001$) are applied for analysis.

Type I error rate

The type I error rates of DCAT and four other representative methods (TWT, MGAS, metaCCA, and MAT) across three significance levels are presented in Table 3. When $q = 10$, taking scenario S_{1-1} as an example for scenarios with the AR structure: at the significance level of 0.05, the type I error rates of TWT, MGAS, metaCCA, MAT, and DCAT are 0.0528, 0.0613, 0.0361, 0.0454 and 0.0533, respectively. Except for metaCCA, all other methods in AR-based scenarios keep their type I error rates within the corresponding confidence intervals (95%, 99%, 99.9%) at the three set significance levels ($\alpha = 0.05, 0.01, 0.001$), which indicates effective type I error control ($S_{1-1}, S_{1-2}, S_{1-3}$). The similar results can be obtained for scenarios with the ARCS structure ($S_{2-1}, S_{2-2}, S_{2-3}$). When $q = 20$, consistent findings are obtained for both AR and ARCS structures across all scenarios ($S_{3-1}, S_{3-2}, S_{3-3}, S_{4-1}, S_{4-2}, S_{4-3}$).

In summary, most methods effectively control the type I error rate at the nominal significance levels, the only exception is metaCCA, which exhibits an overly conservative performance.

Statistical power

When $q = 10$, the phenotype covariance matrix has an AR structure, the statistical powers of 9 scenarios are shown in Supplementary Figure S1. As observed, DCAT achieves the highest statistical power in most scenarios. For example, in scenario S_{1-1-1}, characterized by a phenotype correlation strength of $\rho = 0.2$, a sparse effect

Table 3 Type I error rates of DCAT and other four representative methods

Scenario	α	TWT	MGAS	metaCCA	MAT	DCAT
S_{1-1}	0.05	0.0528	0.0613	0.0361	0.0454	0.0533
	0.01	0.0105	0.0133	0.0066	0.0083	0.0100
	0.001	0.0008	0.0013	0.0004	0.0007	0.0010
S_{1-2}	0.05	0.0564	0.0663	0.0359	0.0453	0.0539
	0.01	0.0115	0.0139	0.0067	0.0088	0.0095
	0.001	0.0008	0.0018	0.0004	0.0006	0.0015
S_{1-3}	0.05	0.0548	0.0589	0.0353	0.0456	0.0496
	0.01	0.0118	0.0143	0.0068	0.0088	0.0101
	0.001	0.0006	0.0014	0.0004	0.0007	0.0012
S_{2-1}	0.05	0.0539	0.0617	0.0313	0.0422	0.0485
	0.01	0.0091	0.0126	0.0058	0.0077	0.0084
	0.001	0.0006	0.0019	0.0004	0.0004	0.0009
S_{2-2}	0.05	0.0500	0.0664	0.0295	0.0411	0.0463
	0.01	0.0094	0.0143	0.0055	0.0071	0.0080
	0.001	0.0006	0.0017	0.0004	0.0003	0.0012
S_{2-3}	0.05	0.0524	0.0611	0.0331	0.0433	0.0474
	0.01	0.0111	0.0129	0.0063	0.0084	0.0097
	0.001	0.0011	0.0010	0.0004	0.0005	0.0012
S_{3-1}	0.05	0.0498	0.0589	0.0274	0.0406	0.0639
	0.01	0.0105	0.0137	0.0043	0.0084	0.0114
	0.001	0.001	0.0015	0.0004	0.0011	0.0012
S_{3-2}	0.05	0.0532	0.0602	0.029	0.041	0.0621
	0.01	0.0109	0.0124	0.0041	0.0085	0.0106
	0.001	0.0011	0.0013	0.0004	0.0006	0.0009
S_{3-3}	0.05	0.053	0.0563	0.0286	0.0417	0.0572
	0.01	0.0103	0.0128	0.0044	0.0085	0.0106
	0.001	0.0011	0.0014	0.0004	0.001	0.0014
S_{4-1}	0.05	0.0462	0.0627	0.0234	0.0363	0.0555
	0.01	0.0086	0.0135	0.0035	0.0067	0.0101
	0.001	0.0008	0.0013	0.0004	0.0008	0.001
S_{4-2}	0.05	0.0452	0.0639	0.0214	0.0347	0.0525
	0.01	0.0094	0.014	0.0032	0.0064	0.085
	0.001	0.001	0.0011	0.0004	0.0006	0.0009
S_{4-3}	0.05	0.0506	0.06	0.0257	0.0388	0.0542
	0.01	0.0102	0.013	0.0039	0.0078	0.0099
	0.001	0.0008	0.0018	0.0004	0.0008	0.001

The significance level α is set to 0.05, 0.01 and 0.001

pattern, and a significance level of 0.05, DCAT exhibits a power of 0.962, whereas the powers of TWT, MGAS, MAT and metaCCA are 0.909, 0.912, 0.910 and 0.606, respectively. Even when the significance level α becomes more stringent (0.01 and 0.001) or the effect pattern shifts from sparse to dense, DCAT remains the top-performing method.

As the phenotype correlation strength ρ increases, DCAT's power exhibits the smallest variation compared to the other four methods. For instance, in scenario S_{1-2-2}, defined by $\rho = 0.4$, a moderate effect pattern, and a significance level of 0.01, DCAT achieves a power of 0.798. In contrast, TWT, MGAS, MAT and metaCCA only reach powers of 0.586, 0.584, 0.603 and 0.230, respectively.

However, when ρ further increases to 0.6, DCAT no longer has the highest power in scenarios with a sparse effect pattern. For example, in scenario S_{1-3-1}, characterized by $\rho = 0.6$, a sparse effect pattern, and a 0.05 significance level, DCAT has a power of 0.880, which is lower than that of TWT (0.915), MGAS (0.912), and MAT (0.915). In contrast, DCAT performs consistently well in scenarios with moderate and dense effect patterns across all correlation strengths.

Meanwhile, similar results are observed across the 9 scenarios when the phenotype covariance matrix follows an ARCS structure Supplementary Figure S2. When $q = 20$, DCAT's power performances across the 18 scenarios are consistent with those in $q = 10$ (Supplementary Figures S1-S2). In conclusion, DCAT had the highest power (or close to it) except in the scenario with both strongly correlated phenotypes and sparse non-null SNPs.

Real data analysis

We further validate the performance of DCAT using summary statistics for four cardiovascular diseases, including essential hypertension, heart failure, obesity and type II diabetes mellitus [29]. This dataset contains summary statistics for approximately 6,674,800 SNPs across the four phenotypes, and rigorous quality control is performed. First, SNPs with a GWAS significance level of p-value $< 5.00 \times 10^{-8}$ or a minor allele frequency (MAF) < 0.05 are excluded. Second, SNPs with allele information inconsistent with the HapMap3 reference panel [24] are removed. Third, the remaining SNPs are mapped to genes using a $\pm 5\text{kb}$ window via the MAGMA software [30]. Finally, genes with fewer than 3 SNPs or more than 15 SNPs are filtered out, leaving a total of 119,578 SNPs for analysis. After this preprocessing, a total of 8932 genes are retained for subsequent analysis.

Moreover, the first two PCs are computed to quantify population stratification, these two PCs are then regressed out to obtain residuals, which are used to further estimate the genotype partial correlation matrix $\boldsymbol{\Delta}_\theta$. For the phenotype correlation matrix $\boldsymbol{\Delta}_\rho$, we calculate pairwise phenotype correlations using independent SNPs with association p-values > 0.05 [31, 32]. In addition, Bonferroni correction is used to evaluate the significance. Given the inclusion of 8932 genes, a significance threshold of p-value $< 0.05/8932 \approx 5.59 \times 10^{-6}$ is applied to define associated genes. We test the 8932 genes one at a time, rather than testing all genes simultaneously. For each gene, we jointly test the significance of all SNPs within that gene with respect to the phenotypes, to identify which specific genes are associated with the phenotypes.

We conduct association tests using our proposed DCAT and four other methods with the summary statistics of the four cardiovascular diseases. TWT, MGAS, metaCCA, MAT and DCAT identify 79, 0, 65, 2 and 103 significant genes, respectively. Among these methods, DCAT detectes the largest number of significant genes, followed by TWT and metaCCA; MGAS and MAT exhibits poor performance, with MGAS identifying no significant genes and MAT only 2. The overlap of significant genes detected by the four methods (DCAT, TWT, metaCCA, MAT) is illustrated in Fig. 1. Specifically, DCAT, TWT, metaCCA, and MAT uniquely identify 31, 7, 1, and 0 significant genes, respectively; 8 significant genes are collectively detected by DCAT and TWT; 62 significant genes are identified by the combination of DCAT, TWT, and metaCCA; and only 2 significant genes are common to all four methods. In addition, the genomic inflation factor of DCAT is calculated to be approximately 1.06, indicating well-calibrated test statistic and effective control the population stratification.

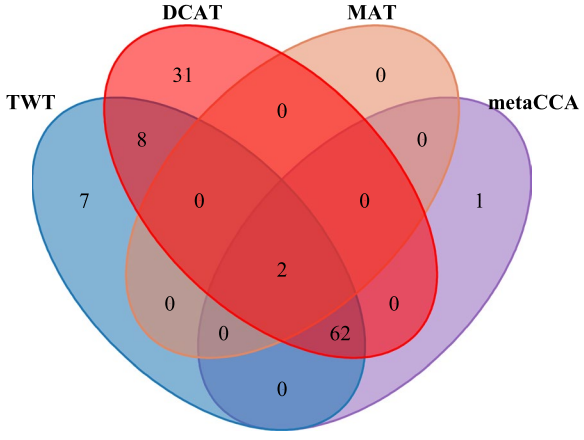

Fig. 1 Venn diagram of significant genes detected by DCAT, TWT, metaCCA and MAT

Table 4 *P*-values of genes that only detected by DCAT at the significance level 5.59×10^{-6} ($\approx 0.05/8932$)

Gene	TWT	MGAS	metaCCA	MAT	DCAT	PMID	Reference
PQLC2	3.22e-3	9.57e-1	7.84e-4	8.69e-1	2.60e-10		
TRAPPC3	1.36e-4	9.94e-1	2.83e-5	9.49e-1	1.28e-7		
SORT1	1.64e-4	2.01e-1	7.08e-5	4.10e-1	3.98e-9	35670343	Su et al. (2023)
RGS21	9.96e-1	9.56e-1	5.45e-1	9.94e-1	1.81e-8		
YWHAQ	1.26e-2	8.40e-1	2.55e-2	6.13e-1	3.97e-6		
ATIC	5.42e-3	3.49e-1	2.10e-1	2.27e-1	1.21e-8	35870639	Surapaneni et al. (2022)
MTMR14	2.53e-3	9.69e-1	3.50e-4	9.24e-1	9.17e-10		
TRAT1	4.79e-5	4.47e-4	1.89e-4	4.36e-2	1.91e-6	36187128	Yu et al. (2022)
XRN1	4.66e-3	9.74e-1	3.66e-4	9.54e-1	7.22e-8		
KCNMB3	2.05e-2	1.24e-1	5.01e-2	7.88e-1	1.14e-8		
TCTEX1D2	1.66e-4	9.49e-1	5.97e-2	6.85e-1	1.53e-6		
FBXL5	1.51e-4	4.75e-1	5.05e-4	7.12e-1	8.82e-7		
C4ORF51	9.88e-1	2.17e-1	2.60e-3	3.71e-1	1.19e-8		
STMN4	9.51e-1	5.68e-1	1.63e-1	3.22e-1	3.65e-6		
MELK	8.22e-1	6.65e-1	5.15e-2	4.18e-1	3.48e-6	36606461	Ke et al. (2022)
TGFBR1	5.07e-5	4.30e-1	1.31e-4	6.31e-1	3.29e-9	33225722	Yamaguchi et al. (2021)
SGPL1	7.20e-4	5.17e-2	8.63e-4	2.60e-2	4.42e-6	38163490	Chen et al. (2024)
C10ORF32	7.61e-4	1.20e-1	2.84e-4	1.64e-1	2.81e-7		
PLEKHA1	1.62e-3	4.02e-2	6.98e-4	5.86e-2	3.41e-6	38152129	Zhu et al. (2023)
KNDC1	2.45e-2	8.66e-1	2.38e-2	8.94e-1	1.21e-6		
RDH16	1.02e-2	2.52e-2	2.20e-3	4.05e-1	4.23e-7	38374256	Suzuki et al. (2024)
RBM25	4.75e-4	9.55e-1	8.25e-4	9.65e-1	1.73e-10	22939879	Gao et al. (2013)
ST20-MTHFS	1.61e-3	4.38e-1	2.94e-1	2.02e-1	2.51e-9		
SMAD4	2.99e-1	6.38e-1	2.46e-1	6.81e-1	3.10e-6	26245826	Du et al. (2025)
ONECUT2	3.46e-5	2.08e-1	2.38e-4	3.17e-1	2.64e-6		
USHBP1	1.99e-4	3.93e-1	1.57e-5	1.09e-2	7.17e-7		
ZNF880	1.26e-1	2.60e-1	5.00e-2	1.26e-1	2.63e-6		
ITPA	9.03e-5	1.31e-1	6.18e-4	6.61e-2	1.31e-6	35870639	Surapaneni et al. (2022)
JAG1	3.78e-2	3.08e-1	6.48e-3	1.92e-1	1.95e-6	34219868	Al-Awaida et al. (2021)
RTCB	9.92e-1	7.46e-1	2.63e-2	9.91e-1	9.07e-12		
CRELD2	2.62e-2	9.04e-1	4.85e-3	4.25e-1	1.43e-10	40695744	Wang et al. (2025)

The bolded genes are the ones that have been reported. PMID indicates the published reference number

We identified 7 genes unique to TWT, of which 2 (NT5DC1 [33], MX1 [34]) have been previously documented to be associated with the above four phenotypes in the literature. In contrast, the *p*-values of the 31 genes uniquely identified by our proposed DCAT are listed in Table 4, among which 13 have been reported in previous studies. For instance, the SORT1 gene is known to inhibit fat production and prevent obesity [35]. The *p*-values of SORT1 gene across different methods are as follows: TWT (1.64×10^{-4}), MGAS (2.01×10^{-1}), metaCCA (7.08×10^{-5}), MAT (4.10×10^{-1}), and DCAT (3.98×10^{-9}). Furthermore, the RBM25 gene has been confirmed to increase the risk of arrhythmia and accelerate the progression of heart failure [36]. The *p*-values of RBM25 gene are: TWT (4.75×10^{-4}), MGAS (9.55×10^{-1}), metaCCA (8.25×10^{-4}), MAT (9.65×10^{-1}), and DCAT (1.73×10^{-10}). Notice that the genes not bolded in Table 4 are newly discovered in this study, and their functions remain to be further explored.

Discussion

Summary statistics have emerged as publicly available data resources since individual-level data in GWAS are typically unavailable due to privacy and legal constraints. Multi-trait SNP-set association analyses can help to uncover potential genetic associations by integrating intrinsic genetic structure and pleiotropic effects, which would be overlooked in "one-to-one" association analyses.

This study proposes a divided-and-combined association test (DCAT) for pleiotropic effects with GWAS summary statistics. Through theoretical derivation, simulation study and real data analysis, we evaluate its performance in genetic association analyses of multiple phenotypes. Methodologically, DCAT considers that the covariance of *Z*-scores arises from both phenotypes and genotypes, and constructs the covariance matrix $\boldsymbol{\Delta}$ of the *Z*-score by combining the phenotype correlation matrix $\boldsymbol{\Delta}_\rho$ and genotype partial correlation matrix $\boldsymbol{\Delta}_\theta$

via Kronecker product, making it more comprehensive and practically meaningful than existing methods in the literature that solely use phenotype covariance as a substitute for the covariance of *Z*-scores [20, 25]. To verify the estimation accuracy of $\boldsymbol{\Delta}_\rho$ following the approach in the previous work [20], we compare the simulated and estimated phenotype correlations ($q = 3$) for DCAT under scenario S_{1-3}. This comparison is based on test statistics derived from 24,142 independent SNPs in the 1000 Genomes Project reference dataset. The results show that the estimation biases are all below 0.05, indicating satisfactory estimation performance. The detailed results are shown in Supplementary Table S1.

To improve the statistical power, we integrate a family of quadratic statistics by introducing a set of parameters ω ($\omega \in \Omega = \{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2\}$) to build stratified statistics $T^{(\omega)}$. As it is illustrated in Supplementary Note 1, quadratic statistics can be decomposed into a weighted sum of mutually independent $\chi^2(1)$ random variables, the core factor influencing the power of quadratic statistics is the weights related to the correlations among *Z*-scores from different phenotypes. DCAT focuses on improving the test power by appropriately adjusting the weights of the covariance matrix $\boldsymbol{\Delta}$ and integrating various types of quadratic statistics. It approximates the distribution of $T^{(\omega)}$ using a location-shifted generalized gamma distribution to calculate *p*-values, and finally integrates these *p*-values via the Cauchy combination test. To examine whether the first four cumulants of the shifted generalized gamma distribution closely match those of $\zeta_v \sim \chi^2(1)$, we calculate the ARMAE (average relative mean absolute error) under simulation scenarios S_{1-2} with $\omega \in \{-2, -1, -\frac{1}{2}, 0, \frac{1}{2}, 1, 2\}$ and each experiment repeated 1,000 times. All ARMAE values are below 0.07, demonstrating that the first four cumulants of the shifted generalized gamma distribution match those of $\chi^2(1)$ with high accuracy. The corresponding results are summarized in Supplementary Table S2.

There are also several limitations of DCAT. Firstly, DCAT is not computationally efficient. We performed a runtime comparison under scenario $S_{4\text{-}1}$ with 10,000 replications on Intel(R) Core(TM) i5-8300H CPU 2.30GHz. The runtimes were 66.93 minutes for DCAT, 9.69 minutes for TWT, 55.72 minutes for MGAS, 0.42 minutes for metaCCA, and 7.68 minutes for MAT, respectively. These results indicate that DCAT has a substantially higher computational burden compared with the other methods. The most time-consuming step is the calculation of the null distribution of the test statistic. As a divided-and-combined statistic, each component $T^{(\omega)}$ does not follow a traditional distribution but instead adheres to a weighted chi-square distribution. We use a location-shifted generalized gamma distribution to approximate the null distribution of $T^{(\omega)}$, enabling the generation of arbitrarily small *p*-values that meet the stringent significance level of GWAS. However, we still need to generate random samples to estimate the parameters of this location-shifted generalized gamma distribution. Although parameter estimation with a smaller sample size saves considerable time, we adopt a large sample size to ensure the accuracy of the final *p*-values. The generation and computation of large samples are the main reasons for the slow computational speed of our proposed method when both the number of phenotypes *q* and the numberof SNPs *m* are large. Therefore, it is highly valuable to construct test statistics specifically designed for in high-dimensional scenarios.

Another important limitation of DCAT is its potential sensitivity to linkage disequilibrium (LD) misspecification and numerical stability issues. The genotype partial correlation matrix $\boldsymbol{\Delta}_\theta$ in DCAT is estimated based on the 1000 Genomes reference panel, but in practical research, the LD structure of the reference panel may differ from that of the study population due to factors such as ancestry mismatch, imputation differences, or sampling noise. This LD misspecification may affect the accuracy of the constructed covariance matrix $\boldsymbol{\Delta}$ and further impact the test performance. Due to the scope of this study, we only evaluated the performance of DCAT under idealized LD specification in simulations, and plan to conduct in-depth and systematic research on these issues in future work to further optimize the method's robustness.

The third limitation is that our proposed DCAT does not achieve the best performance under scenarios characterized by high phenotypic correlation ($\rho = 0.6$) and a sparse effect pattern (30%) (S_{1-3-1} and S_{2-3-1}). Here, we briefly explain the underlying reasons for this phenomenon. As noted earlier, the Cauchy combination method is employed to construct the final test statistic. While this method offers convenience in combining *p*-values without requiring complex bootstrap procedures, it may lead to power loss in certain scenarios due to approximation errors. In general, the Cauchy combination test is highly effective when individual signals are strongly significant. However, it may struggle to exert its advantages when none of the seven statistics $T^{(\omega)}$ exhibit particularly significant signals overall under certain scenarios.

Finally, the objective of this study is to test whether the SNP-set within a gene or a biological pathway has direct pleiotropic effects with multiple phenotypes, rather than the indirect associations arising from the causal relationships among phenotypes. Detected signals can be regarded as evidence of joint genetic influence rather than definitive biological pleiotropy. Furthermore, it remains unclear which specific phenotype is associated with the SNP-set or which SNP is most likely to exhibit

pleiotropy with multiple phenotypes, warranting further in-depth investigations in related fields.

Conclusion

In conclusion, DCAT exhibits robust type I error control and efficient detection of pleiotropic effects in multi-trait SNP-set genetic association analyses. It may serve as a reliable methodological tool for dissecting the shared genetic basis underlying multiple phenotypes.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12883-9.

Supplementary Material 1.

Acknowledgements

The authors are grateful to the three anonymous reviewers and the editor for their valuable comments and constructive suggestions on this paper.

Authors' contributions

H. G. contributed to the conception, design, and revision of the study. B. H. drafted the manuscript, performed the simulation experiments, and conducted the real-data analyses. Y. C. performed the theoretical derivation and revised the manuscript. X. Z. participated in the discussion.

Funding

This work has been supported by National Natural Science Foundation of China (Grant No. 12401344), National Statistical Science Research Project of China (Grant No. 2024LY017), and Hubei Province University's Outstanding Young and Middle-Aged Scientific and Technological Innovation Team Project (Grant No. T2023020).

Data availability

The simulation procedures for generating summary statistics data have been described in the manuscript. The real data analysis utilized publicly available GWAS summary statistics from http://www.nealelab.is/uk-biobank.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 8 November 2025 / Accepted: 21 April 2026
Published online: 30 April 2026

References

1. Visscher PM, Wray NR, Zhang Q, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. Am J Hum Genet. 2017;101(1):5–22.
2. Manolio TA, Collins FS, Cox NJ, et al. Finding the Missing Heritability of Complex Diseases. Nature. 2009;461(7265):747–53.
3. Solovieff N, Cotsapas C, Lee H, et al. Pleiotropy in Complex Traits: Challenges and Strategies. Nat Rev Genet. 2013;14(7):483–95.
4. Sivakumaran S, Agakov F, Theodoratou E, et al. Abundant Pleiotropy in Human Complex Diseases and Traits. Am J Hum Genet. 2011;89:607–18.
5. Cross-Disorder Group of the Psychiatric Genomics Consortium. Identification of Risk Loci with Shared Effects on Five Major Psychiatric Disorders: A Genome-Wide Analysis. Lancet. 2013;381:1371–9.
6. Watanabe K, Stringer S, Frei O, et al. A Global Overview of Pleiotropy and Genetic Architecture in Complex Traits. Nat Genet. 2019;51(9):1339–48.
7. Mackay TFC, Anholt RRH. Pleiotropy, Epistasis and the Genetic Architecture of Quantitative Traits. Nat Rev Genet. 2024;25:639–57.
8. Wu M, Lee S, Cai T, et al. Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. Am J Hum Genet. 2011;89(1):82–93.
9. Liu Y, Chen S, Li Z, et al. ACAT: A Fast and Powerful *p* Value Combination Method for Rare-Variant Analysis in Sequencing Studies. Am J Hum Genet. 2019;104(3):410–21.
10. Han F, Pan W. A Data-Adaptive Sum Test for Disease Association with Multiple Common or Rare Variants. Hum Hered. 2010;70(1):42–54.
11. van der Sluis S, Posthuma D, Dolan CV. TATES: Efficient Multivariate Genotype-Phenotype Analysis for Genome-Wide Association Studies. PLoS Genet. 2013;9(1):e1003235.
12. Turley P, Walters R, Maghzian O, et al. Multi-Trait Analysis of Genome-Wide Association Summary Statistics Using MTAG. Nat Genet. 2018;50(2):229–37.
13. Zhang W, Yang L, Tang LL, et al. GATE: An Efficient Procedure in Study of Pleiotropic Genetic Associations. BMC Genomics. 2017;18(1):552.
14. Leslie R, O'Donnell CJ, Johnson AD. GRASP: Analysis of Genotype-Phenotype Results from 1390 Genome-Wide Association Studies and Corresponding Open Access Database. Bioinformatics. 2014;30(12):i185–94.
15. Cichonska A, Rousu J, Marttinen P, et al. metaCCA: Summary Statistics-Based Multivariate Meta-Analysis of Genome-Wide Association Studies Using Canonical Correlation Analysis. Bioinformatics. 2016;32(13):1981–9.
16. Guo B, Wu B. Powerful and Efficient SNP-Set Association Tests across Multiple Phenotypes Using GWAS Summary Data. Bioinformatics. 2019;35(8):1366–72.
17. Bu D, Wang X, Li Q. Summary Statistics-Based Association Test for Identifying the Pleiotropic Effects with Set of Genetic Variants. Bioinformatics. 2023;39(4):btad182.
18. Bu D, Yang Q, Meng Z, et al. Truncated Tests for Combining Evidence of Summary Statistics. Genet Epidemiol. 2020;44:687–701.
19. Lorincz-Comi N, Song W, Chen X, et al. Combining xQTL and Genome-Wide Association Studies from Ethnically Diverse Populations Improves Druggable Gene Discovery. 2025. https://doi.org/10.21203/rs.3.rs-6700169/v1.
20. Zhu X, Feng T, Tayo BO, et al. Meta-Analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension. Am J Hum Genet. 2015;96(1):21–36.
21. Liu Z, Lin X. Multiple Phenotype Association Tests Using Summary Statistics in Genome-Wide Association Studies. Biometrics. 2018;74(1):165–75.
22. LeBlanc M, Zuber V, Thompson WK, et al. A Correction for Sample Overlap in Genome-Wide Association Studies in a Polygenic Pleiotropy-Informed Framework. BMC Genomics. 2018;19(1):494.
23. The 1000 Genomes Project Consortium. A Global Reference for Human Genetic Variation. Nature. 2015;526(7571):68–74.
24. The International HapMap 3 Consortium. Integrating Common and Rare Genetic Variation in Diverse Human Populations. Nature. 2010;467(7311):52–8.
25. Kim J, Bai Y, Pan W. An Adaptive Association Test for Multiple Phenotypes with GWAS Summary Statistics. Genet Epidemiol. 2015;39:651–63.
26. Wang Z, Wang J, Li Q. Quadratic Form Statistics and Cauchy Statistics. J Appl Stat Manag (in Chinese). 2021;40(3):405–16.
27. Wang J, Long M, Li Q. A Maximum Kernel-Based Association Test to Detect the Pleiotropic Genetic Effects on Multiple Phenotypes. Bioinformatics. 2023;39(5):btad143.
28. 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.
29. Lee CH, Shi H, Pasaniuc B, et al. PLEIO: A Method to Map and Interpret Pleiotropic Loci with GWAS Summary Statistics. Am J Hum Genet. 2021;108(1):36–48.
30. De Leeuw CA, Mooij JM, Heskes T, et al. MAGMA: Generalized Gene-Set Analysis of GWAS Data. PLoS Comput Biol. 2015;11(4):e1004219.
31. Wang K, Abbott D. A Principal Components Regression Approach to Multilocus Genetic Association Studies. Genet Epidemiol. 2008;32(2):170–81.
32. Kwak IY, Pan W. Gene- and Pathway-Based Association Tests for Multiple Traits with GWAS Summary Statistics. Bioinformatics. 2017;33(1):64–71.
33. Geng R, Li Z, Yu S, et al. Weighted gene co-expression network analysis identifies specific modules and hub genes related to subsyndromal symptomatic depression. World J Biol Psychiatry. 2020;21(2):102–10.
34. Yu H, Yu M, Li Z, et al. Identification and analysis of mitochondria-related key genes of heart failure. J Transl Med. 2022;20(1):410.

35. Su X, Wang B, Lai M, et al. Apolipoprotein A1 Inhibits Adipogenesis Progression of Human Adipose-Derived Mesenchymal Stem Cells. Curr Mol Med. 2023;23(8):762–73.
36. Gao G, Dudley SC. RBM25/LUC7L3 Function in Cardiac Sodium Channel Splicing Regulation of Human Heart Failure. Trends Cardiovasc Med. 2013;23(1):5–8.

Publisher's Note