



Effect of non-normality and low count variants on cross-phenotype association tests in GWAS

Debashree Ray^{1,2} · Nilanjan Chatterjee^{2,3}

Received: 3 December 2018 / Revised: 1 September 2019 / Accepted: 5 September 2019 / Published online: 3 October 2019
© The Author(s), under exclusive licence to European Society of Human Genetics 2019

Abstract

Many complex human diseases, such as type 2 diabetes, are characterized by multiple underlying traits/phenotypes that have substantially shared genetic architecture. Multivariate analysis of correlated traits has the potential to increase the power of detecting underlying common genetic loci. Several cross-phenotype association methods have been proposed—some require individual-level data on traits and genotypes, while the others require only summary-level data. In this article, we explore whether non-normality of multivariate trait distribution affects the inference from some of the existing multi-trait methods and how that effect is dependent on the allele count of the genetic variant being tested. We find that most of these tests are susceptible to biases that lead to spurious association signals. Even after controlling for confounders that may contribute to non-normality and then applying inverse normal transformation on the residuals of each trait, these tests may have inflated type I errors for variants with low minor allele counts (MACs). A likelihood ratio test of association based on the ordinal regression of individual-level genotype conditional on the traits seems to be the least biased and can maintain type I error when the MAC is reasonably large (e.g., $MAC > 30$). Application of these methods to publicly available summary statistics of eight amino acid traits on European samples seem to exhibit systematic inflation (especially for variants with low MAC), which is consistent with our findings from simulation experiments.

Introduction

With the availability of rich data on multiple complex traits (or phenotypes) from genome-wide association studies (GWAS) and biobanks, several recent large-scale studies [1–7] have examined genetic associations of multiple traits simultaneously. The advent of advanced technologies that can measure several quantitative traits—such as the automated high-throughput serum nuclear magnetic resonance

(NMR) metabolomics platform that provides quantitative molecular data on hundreds of metabolites—has further led to growing interest in multi-trait genetic analyses [8–10]. Jointly analyzing multiple correlated disease-related traits can increase power (over multiple single-trait analyses) to identify genetic loci influencing at least one of the traits [1, 11]. To address the burgeoning demand for multi-trait analysis in GWAS, several methods—some based on individual-level data and some based on single-trait summary statistics—have been proposed. These cross-phenotype (or multivariate) methods test the null hypothesis of no association of a genetic variant with any of the correlated traits being jointly analyzed against the alternative hypothesis that it is associated with at least one of the traits.

Most of the existing cross-phenotype methods based on individual-level data assume multivariate normality of the traits. However, ensuring multivariate normality is not straightforward, and univariate normality (typically achieved by inverse normalizing traits or trait residuals after covariate adjustment) of each trait does not guarantee that the traits or trait residuals are jointly multivariate normal [12, 13] (see Supplementary S1 for more details).

Supplementary information The online version of this article (<https://doi.org/10.1038/s41431-019-0514-2>) contains supplementary material, which is available to authorized users.

✉ Debashree Ray
dray@jhu.edu

¹ Department of Epidemiology, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA

² Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205, USA

³ Department of Oncology, School of Medicine, Johns Hopkins University, Baltimore, MD 21205, USA

Consequently, as noted earlier [13], “problems with outliers can be more extreme in multivariate settings”, which can be “particularly acute when dealing with strongly correlated phenotypes”. The existing cross-phenotype methods based on single-trait GWAS summary statistics rely on asymptotic normality of each estimated trait effect, and assume asymptotic multivariate normality of the ratios of these effect estimates to their standard errors.

Currently, many GWAS focus their multivariate analyses on variants with minor allele frequency (MAF) $\geq 1\%$ or $\geq 5\%$ irrespective of the sample size or the availability of variants with lower MAFs. Recent focus on sequencing or dense chip-based studies to understand the effect of low-frequency and rare variants on complex traits requires investigation of the calibration and the power of existing multivariate methods across the entire allele-frequency spectrum.

In this article, we explore how some of the existing cross-phenotype methods perform (in terms of type I error control) under deviations from multivariate normality, especially when testing association with a low-count genetic variant. Previous studies have explored the effect of non-normality on rare-variant set-based tests of a single trait [14], and the effect of normalizing traits in single-variant single-trait tests [15]. To the best of our knowledge, no study explored effect of non-normality on various cross-phenotype association tests, especially in the context of rare variant studies where the assumption of multivariate normality can lead to serious inflation in type I error. Similar to what has been reported in previous studies for single-trait test [16], we observe that the minor allele count (MAC) is the key parameter that determines the type I error calibration of multivariate tests as well. The MAC threshold after which a test is well calibrated, however, can be much higher for multivariate than univariate test. In addition, we compare power of these methods under the ideal scenario of multivariate normality of the traits. Finally, we apply some of the existing single-variant cross-phenotype methods on summary data from eight amino acid NMR traits collected on up to 24,295 European samples.

Material and methods

Model and notation

Consider a GWAS on n individuals, genotyped/sequenced on p genetic variants and measured for K traits (possibly correlated). For a given genetic variant, let X_i take values 0, 1, or 2 for individual i , and \mathbf{X} be the $n \times 1$ vector of genotypes for all individuals. Let \mathbf{Y}_k be the $n \times 1$ vector of k th trait and \mathbf{Y} be the $n \times K$ matrix of all traits for all individuals. For simplicity of notation, assume there is no other covariate (note that this assumption can be easily relaxed by

considering trait residuals after regressing out covariate effects). We are interested in testing the association of a single genetic variant with the K traits.

For testing cross-phenotype associations, several methods have been proposed. Some methods require individual-level phenotype-genotype data, while others require only summary-level data (the estimated genetic effect size and its estimated standard error, or the p -value of association). For these tests, the null hypothesis (H_0) of interest is that none of the K traits is associated with a given genetic variant against the alternative hypothesis (H_a) that at least one trait is associated. Here is a brief overview of some of the existing methods with a summary in Table 1 (most other methods are well documented in a recent review article [17]). In this paper, for all our analyses, we apply these methods on both raw traits and inverse normalized traits for comparison. The rank-based inverse normal transformation (INT) of a trait involves ranking the trait values and then mapping the ranks to percentiles of the standard normal distribution. Mathematically, the k th inverse normalized trait for individual i is $Y_{i,k}^{INT} = \Phi^{-1}((r_{i,k} - 0.5)/n_k)$, where $r_{i,k}$ is the rank of i th observation for the k th trait in a sample of size n_k ($n_k < n$ if there are missing values), and $\Phi^{-1}(\cdot)$ is the standard normal quantile function.

Existing methods based on individual-level data

MANOVA

Multivariate analysis of variance (MANOVA) [18, 19] considers the multivariate linear regression model

$$\mathbf{Y}_{n \times K} = \mathbf{1}_{n \times 1} \boldsymbol{\alpha}'_{1 \times K} + \mathbf{X}_{n \times 1} \boldsymbol{\beta}'_{1 \times K} + \boldsymbol{\mathcal{E}}_{n \times K} \quad (1)$$

where $\boldsymbol{\alpha}$ is the vector of intercepts, $\mathbf{1}$ is the corresponding column of 1s, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)'$ is the vector of fixed unknown genetic effects of the K correlated traits, and $\boldsymbol{\mathcal{E}}$ is the matrix of random errors. Each row $\boldsymbol{\mathcal{E}}_i$ of the error matrix $\boldsymbol{\mathcal{E}}$ is assumed independently distributed as a K -variate normal with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ (a $K \times K$ positive definite matrix representing residual correlation among the traits). This assumption imposes the constraint that the individuals are unrelated. Multivariate linear mixed model (mvLMM) uses an additional matrix of random effects to account for sample relatedness and population stratification (e.g., GEMMA [20]). The null hypothesis of no association using Eq. 1 is $H_0 : \boldsymbol{\beta} = \mathbf{0}$, and the likelihood ratio test (LRT) of H_0 gives the MANOVA (or, Wilk's Lambda) test statistic, which has an asymptotic chi-squared distribution with K degrees of freedom (d.f.) (see [11]). It is equivalent to single-variant cross-phenotype association test based on Canonical Correlation Analysis (CCA) [21].

Table 1 Summary of the existing single-variant cross-phenotype association tests that are used in this study

Method [Ref.]	Data type ^a	Trait type	Model/test type	P-value calculation approach	Scenario/s when most powerful
metaMANOVA [33]	Summary	Any	Classical multivariate Wald test (quadratic)	P -value using asymptotic χ^2 distribution	Subset of traits associated; heterogeneous effects
metaUSAT [33]	Summary	Any	Data-adaptive combination of two quadratic tests	Approximate p -value using fast numerical integral	Robust power across many association scenarios
S_{Hom} [31]	Summary	Any	Burden type linear test	P -value using asymptotic χ^2 1-d.f. distribution	All traits associated with homogeneous effects
S_{Het} [31]	Summary	Any	Subset-based approach	Approximate p -value using an estimated gamma distribution	Subset of traits associated; heterogeneous effects
MTAR [37]	Summary	Any	Data-adaptive combination of a quadratic and a linear tests	Analytical p -value using linear combination of χ^2	Robust power across many association scenarios
PCO [43]	Summary	Any	Data-adaptive combination across linear, quadratic, and non-linear tests	P -value estimated using Monte-Carlo simulations	Robust power across many association scenarios
mixAda [42]	Summary	Any	Data-adaptive combination of two quadratic tests from mixed model	Approximate p -value using fast numerical integral	Robust power across many association scenarios
MANOVA/CCA [21]	Individual-level	Continuous	Multivariate linear regression model	P -value using asymptotic χ^2 distribution	Subset of traits associated; heterogeneous effects
USAT [11]	Individual-level	Continuous	Data-adaptive combination of two quadratic tests	Approximate p -value using fast numerical integral	Robust power across many association scenarios
MultiPhen [22]/POM-LRT	Individual-level	Any	Proportional odds regression model (genotype regressed on traits)	P -value using asymptotic χ^2 distribution	Subset of traits associated; heterogeneous effects

^aData type 'summary' means GWAS summary statistics required for each trait; 'individual-level' means individual-level trait and genotype data required

POM-LRT

This approach models the genotype as an ordinal outcome using proportional odds model (POM) assuming unrelated individuals. The LRT statistic for testing no association has an asymptotic chi-squared distribution with K d.f. under the null. In the context of GWAS, this test is known as MultiPhen [22]. One may also use Wald test statistic instead of LRT in this POM framework (implemented in our R program `mvtests`). Other variations of reverse regression of genotype on phenotypes have also been used for cross-phenotype association tests [23–27].

Unified score-based association test (USAT)

USAT [11] is a data-adaptive combination of the MANOVA and the sum of squared score (SSU) [19, 28] tests for unrelated individuals. To account for relatedness among individuals, the USAT framework may be used to combine LRT statistic from mvLMM and SSU test statistic based on linear mixed model [29]. USAT p -value is approximately computed by a fast one-dimensional numerical integral using the fact that both MANOVA and SSU have chi-squared distributions under the null.

Existing methods based on summary-level data

In a typical GWAS, each trait is separately tested for association with a given genetic variant. The association statistic and the p -value for each trait and each variant is reported based on the univariate/marginal model $Y_k = \alpha_k + \beta_k X + \epsilon_k$ with normally distributed errors ϵ_k if the k th trait is continuous, or the logistic model $\text{logit}(P(Y_k = 1|X)) = \alpha_k + \beta_k X$ if it is binary. For the k th trait ($k = 1, 2, \dots, K$), β_k is the genetic effect and the null hypothesis of no genetic association is $H_{0,k} : \beta_k = 0$. Random effects may be included in these models to account for sample relatedness and population structure (as implemented in, say, EMMAX [30]). The Wald test statistic for $H_{0,k}$ is $Z_k = \hat{\beta}_k / \text{se}(\hat{\beta}_k)$ where $\hat{\beta}_k$ is the maximum likelihood estimate (MLE) of β_k and $\text{se}(\hat{\beta}_k)$ is its standard error. Under $H_{0,k}$, Z_k has an asymptotic standard normal distribution. However, for k th and l th traits, summary statistics Z_k and Z_l are not uncorrelated if the traits are correlated [28, 31]. To test the global null hypothesis of no association with any trait ($H_0 : \beta_1 = \dots = \beta_K = 0$), one can form appropriate test statistics based on GWAS summary statistics $\mathbf{Z} = (Z_1, \dots, Z_K)'$ (as summarized below). Under H_0 , we assume \mathbf{Z} has an asymptotic K -variate normal distribution with mean $\mathbf{0}$ and covariance matrix \mathbf{R} . The $K \times K$ matrix \mathbf{R} can be estimated (denoted by $\hat{\mathbf{R}}$) using the Pearson correlation of Z -statistics on a large number of variants across the

genome that are not marginally associated with any of the K traits [28, 32, 33] (call this estimate $\hat{\mathbf{R}}_{\text{pearson}}$). For highly polygenic traits, cross-trait LD-score regression [34] may be used to estimate \mathbf{R} [35, 36]. Guo and Wu [37] argued that the common practice of filtering out large summary statistics (as is done in LD-score regression) is less efficient and may lead to biased estimates, and hence proposed a robust linear regression on LD-scores (call this estimate $\hat{\mathbf{R}}_{\text{LDscore}}$).

metaMANOVA

This method is equivalent to MANOVA or the classical multivariate Wald test but based on summary statistics only. Its test statistic is $\mathbf{Z}'\hat{\mathbf{R}}^{-1}\mathbf{Z}$, which has an asymptotic chi-squared distribution with K d.f. under the null [28, 33, 38], and sometimes referred to as the ‘omnibus chi-square test’. metaCCA [39] is an extension in the sense that it allows multivariate representation of both genotype and phenotype.

S_{Hom}

Much like O'Brien's test [32, 40], S_{Hom} [31] assumes the genetic effects to be homogeneous across traits and its test statistic is proportional to the sum statistic $\mathbf{1}'(\hat{\mathbf{R}}\mathbf{W})^{-1}\mathbf{Z}$, where $\mathbf{1}' = (1, \dots, 1)$ is a row of 1s, and \mathbf{W} is a diagonal matrix of weights (such as square root of sample sizes) for the Z -statistics. S_{Hom} is asymptotically distributed as a chi-squared variable with 1 d.f. under the null.

S_{Het}

Similar in spirit to Xu et al.'s test [38] using truncated Z -statistics, the data-adaptive approach S_{Het} [31] uses the statistic $\max_{\tau>0} S(\tau)$, where $S(\tau)$ is proportional to $\mathbf{1}'_r(\hat{\mathbf{R}}_r\mathbf{W}_r)^{-1}\mathbf{Z}_r$. Here, \mathbf{Z}_r (and similarly $\mathbf{1}_r, \hat{\mathbf{R}}_r, \mathbf{W}_r$) attempts to capture only the traits with a true contribution to the association of a genetic variant under the alternative hypothesis by considering traits whose underlying association statistics exceed some threshold τ (unknown *a priori*). The null distribution of S_{Het} is empirically approximated by a Gamma distribution. ASSET [41] is another such subset-based approach that can additionally provide information on the subset of traits that is associated with the variant.

metaUSAT

Recognizing the non-existence of the uniformly most powerful test for cross-phenotype associations, metaUSAT [33] adaptively combines metaMANOVA and the summary statistic based SSU [28]. It reports a minimum p -value type test statistic and an approximate asymptotic p -value of association.

MTAR

This method [37] is a data-adaptive combination of meta-MANOVA and a principal component (PC) based 1-d.f. chi-squared test to achieve robust performance under various association scenarios.

Other summary-level methods

Recent methods have focused on being data-adaptive to ensure robust power performance across different alternatives. mixAda [42] adaptively combines two independent score statistics (proportional to the forms $\mathbf{1}'\hat{\mathbf{R}}^{-1}\mathbf{Z}$ and $\mathbf{Z}'\hat{\mathbf{R}}^{-1}\hat{\mathbf{R}}^{-1}\mathbf{Z}$) based on linear mixed effects model, while PCO [43] adaptively combines several linear and nonlinear PC based tests together. HIPO [36] finds optimal linear combinations of association coefficients across traits taking into account estimates of heritability, genetic covariance, sample size variations, and overlaps across traits. MTAG [35] obtains an effect estimate for each trait as a weighted sum of GWAS estimates that uses phenotypic and genotypic covariances estimated from cross-trait LD-score regression [34]. Other methods have been well summarized elsewhere [17, 44].

Simulation experiments

We conduct simulation experiments in \mathbb{R} [45] using $K = 10$ continuous traits based on the multivariate linear regression model (Eq. 1) to assess type I errors of MANOVA, POM-LRT (or MultiPhen), USAT, metaMANOVA, S_{Hom} , S_{Het} , metaUSAT, MTAR, mixAda, and PCO under the following four broad scenarios of multivariate trait non-normality, trait outliers, and trait heteroscedasticity. We did not include HIPO and MTAG in our comparison because they belong to a different class of cross-phenotype tests that borrows information across a genome-wide reference panel of variants (e.g., genome-wide estimates of heritability and genetic covariance) [36]. For comparing power, we assumed the ideal scenario—multivariate normality of traits without heteroscedasticity and outliers. Under each simulation scenario, we consider two different trait correlation structures: compound symmetry \mathbf{R}_{cs} (with same pairwise trait correlation) and block diagonal \mathbf{R}_{bd} (with strong pairwise correlation within a block and weak pairwise correlation between blocks). For \mathbf{R}_{cs} , the pairwise trait correlation ρ is either 0.2 (weak correlation) or 0.5 (moderate correlation) or 0.9 (strong correlation). For \mathbf{R}_{bd} , the first five traits have pairwise correlation 0.5, the next 5 traits have pairwise correlation 0.9, and the between-group correlation is 0.2.

We simulate each dataset on $n = 3000$ or 10,000 unrelated individuals based on a single bi-allelic variant in Hardy–Weinberg equilibrium with a fixed population-level

MAF (and corresponding expected MAC) and genetic effects $\beta_1, \dots, \beta_{10}$. We looked at expected MAC 6, 30, and 300, which respectively corresponds to MAF 0.1% (0.03%), 0.5% (0.15%), and 5% (1.5%) for sample size 3000 (10,000). We assume the commonly used additive genetic model in our simulations, and assume that the total variance of a trait is composed of the variance attributable to the genetic variant (σ_g^2) and the residual variance (σ_e^2). Under the null model (for type I error analysis), we set $\sigma_g^2 = 0$ while the alternative models (for power comparison) have $\sigma_g^2 = 0.05$ for the associated traits. We obtain the genetic effect of the k th trait from the relation $\sigma_g^2 = \beta_k^2 \times \text{Var}(X)$, where $\text{Var}(X)$ is the population-level variance of the genetic variant (a function of the MAF) [46]. In other words, we set $\beta_1 = \dots = \beta_{10} = 0$ for type I error analysis while for power comparison, we set genetic effect >0 for the associated traits only. As for the joint covariance structure of the traits, we assume either compound symmetry (or exchangeable) structure $\Sigma_{cs} = \sigma_e^2 \mathbf{R}_{cs}$ or the block-diagonal structure $\Sigma_{bd} = \sigma_e^2 \mathbf{R}_{bd}$, where σ_e^2 is set at 9.95. We simulate 10 million replicates to estimate type I error rates, and 10,000 replicates to estimate asymptotic power at the GWAS threshold $\alpha = 5 \times 10^{-8}$. For simplicity, we do not include any additional covariates.

Scenario 1: multivariate Laplace distributed traits

To mimic a heavy-tailed trait distribution, the random errors in Eq. 1 are drawn from a multivariate Laplace with mean $\mathbf{0}$ and covariance matrix Σ_{cs} or Σ_{bd} . This is a symmetric, elliptical distribution like multivariate normal. This scenario can be considered as a case of slight departure from multivariate normality with respect to tail behavior.

Scenario 2: multivariate t distributed traits

The error distribution is assumed to be multivariate t with 3 d.f., mean $\mathbf{0}$ and covariance matrix Σ_{cs} or Σ_{bd} . Again, it is a symmetric, elliptical distribution but with very heavy tails. When the d.f. of multivariate t is infinitely large, the distribution is multivariate normal.

Scenario 3: multivariate mixture normal traits

To mimic outliers in the trait distribution, we generate error distribution from multivariate normal with mean $\mathbf{0}$ and covariance matrix Σ_{cs} or Σ_{bd} in 95% of the individuals. For the rest 5%, a multivariate normal with mean $\mathbf{0}$ and covariance matrix $\sigma \Sigma_{cs}$ or $\sigma \Sigma_{bd}$ is used. We fix $\sigma = 10$.

Scenario 4: multivariate normal traits with heteroscedasticity

To mimic unequal error variances between genotypes, we simulate error distribution from multivariate normal with mean $\mathbf{0}$ and variance matrix $\tau \Sigma_{cs}$ or $\tau \Sigma_{bd}$, where we fix $\tau = 1, 1.5$ or 2 depending on whether MAC at the variant is 0, 1 or 2 for a given individual.

Scenario 5: multivariate normal traits

This is the ideal scenario where the assumption of multivariate normality of traits is not violated. We simulate error distribution from multivariate normal with mean $\mathbf{0}$ and variance matrix Σ_{cs} or Σ_{bd} . This is the only scenario where we compare power of different methods.

Application to Amino Acids Summary Data

Kettunen et al. [47] analyzed up to 24,295 individuals from 14 European cohorts to perform GWAS for human blood metabolites. They included individuals that had NMR metabolite data and genome-wide single nucleotide polymorphism (SNP) array data. SNPs were imputed up to 39 million variants (build 37). The authors analyzed each cohort separately, and tested univariate associations assuming the additive genetic model. SNPs with accurate imputation (proper info > 0.4) and MAC > 3 were combined in fixed-effects meta-analysis using genomic control correction for both individual cohort results and the meta-analysis results. Their metabolite set covers multiple metabolic pathways.

In this paper, we use the single-trait meta-analysis summary statistics of eight amino acids (alanine, glutamine, histidine, isoleucine, leucine, phenylalanine, tyrosine, and valine) to test multivariate associations with each SNP. We did not include Glycine, another amino acid included in the original study [47], as it has a much larger number of genetic associations compared with other amino acids (Fig. S21) and thus could dominate the cross-phenotype tests. The summary statistics were obtained after adjusting for age, sex, time of last meal (if applicable, as majority of the samples were fasting), and first ten principal components from genomic data, and then inverse normalizing the resulting residuals. For some variants, the effective sample size (and hence the effective MAC) varied widely across the eight traits since some traits are missing from many cohorts included in the original study. To reduce the effect of high or low MAC in a few traits on the joint test of all traits, we discarded variants for which the minimum or maximum MAC across traits was outside $\pm 1.3 \times$ median MAC (Supplementary S5).

Results

Simulation experiments

To describe calibration of methods, we use expected MAC (rather than MAF) because conditional on MAC, the characteristics of methods do not depend on MAF and sample size [16]. For the methods based on summary-level data, we use $\hat{R} = \hat{R}_{\text{pearson}}$ obtained using function `cor.pearson()` from `metaUSAT` software [33]. Parameter choices and implementation details of the methods are given in Supplementary S2. We examine type I error performance of the different methods for varying expected MAC using Quantile-Quantile plot (QQ plot) of observed and expected p -values.

Scenario 1: multivariate Laplace distributed traits

First, we focus on compound symmetry trait correlation structure with low correlation ($\rho = 0.2$). For variants with

MAC around 300 (i.e., MAF 5% for $n = 3000$), all methods seem to be well-calibrated (although `metaUSAT` and `PCO` may have slightly inflated type I errors) with or without INT when there is slight departure from multivariate normality (Fig. 1). For low-count variants, however, all methods (except `POM-LRT`) with or without INT display inflated type I errors. Type I error inflation is severe specially for very low MAC. Also notable is the behavior of `SHom`: it exhibits lesser inflation since it combines all the traits into a single trait, thereby increasing effective sample size and hence effective MAC. We observe similar behavior of all these methods for compound symmetry trait correlation structure with moderate and strong correlations (figures not shown), for block-diagonal trait correlation structure (Fig. S3), and for larger sample size of 10,000 (Fig. S15). Similarity of Fig. 1 (for $n = 3000$) and Fig. S15 (for $n = 10,000$) reaffirms that MAC is a sample-size invariant measure of test calibration. The MAC threshold after which a test is well calibrated, however, can be much higher for a cross-phenotype test (based on multivariate linear

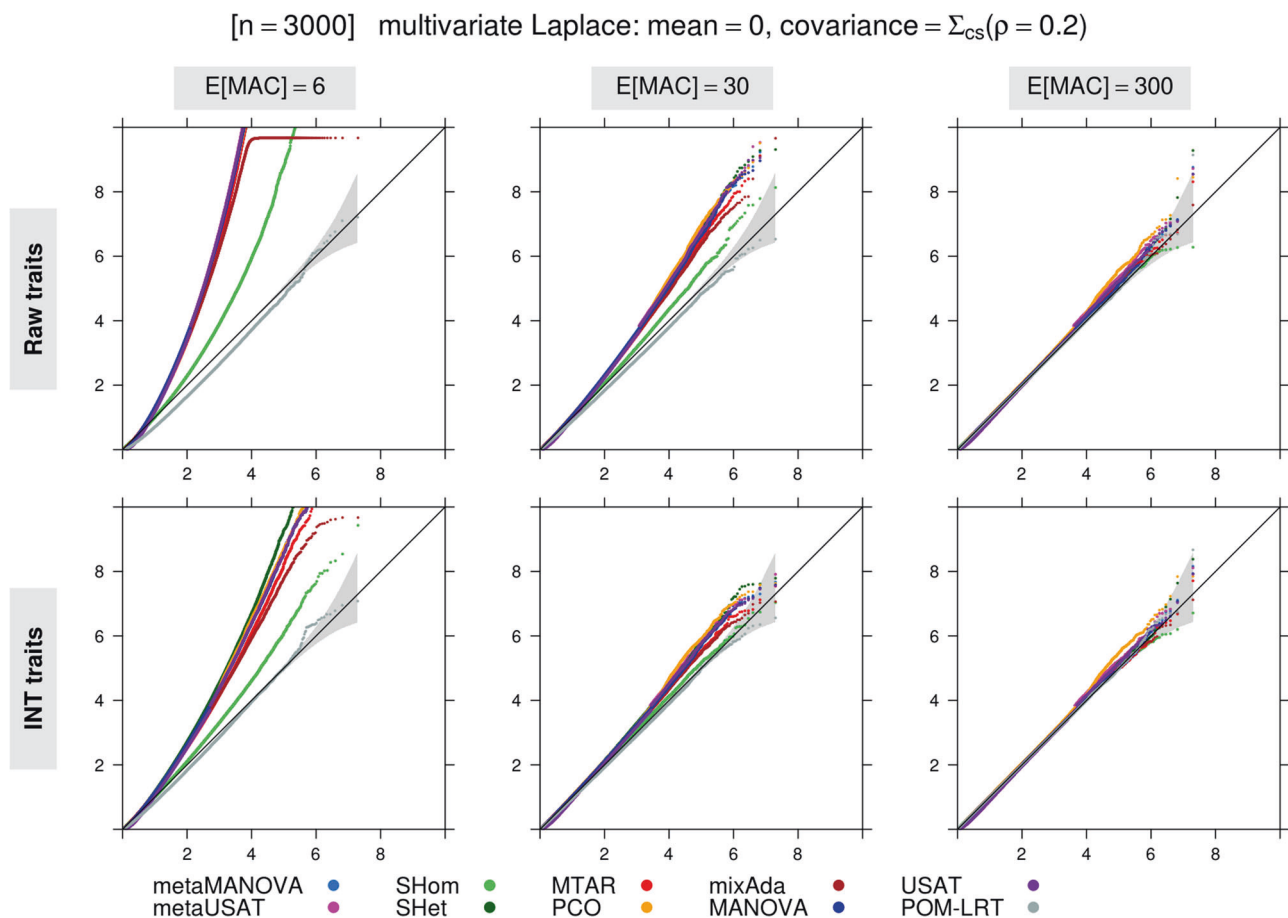


Fig. 1 Scenario 1: QQ plots for null data. Observed ($-\log_{10} p$ -values) are plotted on the y -axis and expected ($-\log_{10} p$ -values) on the x -axis. Each replicate has $n = 3000$ unrelated individuals with $K = 10$ multivariate Laplace distributed traits with pairwise trait correlations $\rho = 0.2$. Performance of cross-phenotype tests is based on 10 million such

replicates. Expected MAC 6, 30, and 300, respectively, correspond to MAF 0.1, 0.5, and 5% for sample size $n = 3000$. The gray shaded region represents a conservative 95% confidence interval for the expected distribution of p -values. P -values $\geq 10^{-10}$ are shown here

regression) than a single-trait test (based on normal linear regression) (see Fig. S4).

Scenario 2: multivariate t distributed traits

When the multivariate traits have thick tails, all the methods, except POM-LRT, suffer from severely inflated type I errors for all MAC values (6, 30, or 300) we studied (Fig. S5). The severity of inflation even at MAC around 300 is, however, not so evident from the values of genomic inflation factor (λ_{GC}) based on p -values (Table S1). Correction of marginal trait distribution using INT helps calibrate the methods for variants with MAC 300 or more. As expected, S_{Hom} exhibits lesser inflation in each scenario. Like before, these observations are corroborated for stronger correlations in a compound symmetry trait correlation structure (figures not shown), for block-diagonal trait correlation structure (Fig. S6), and for larger sample size of 10,000 (Fig. S16).

Scenario 3: multivariate mixture normal traits

In the presence of outliers in the data, most methods have near-nominal type I error rates when testing association with variants having MAC 300 or more (Fig. S7). While methods such as PCO, metaUSAT, S_{Het} , MANOVA, and MTAR may have slightly inflated type I errors for MAC around 300 when applied to raw traits, it can be corrected by applying these methods on inverse normalized traits. For low count variants, all the methods except POM-LRT grossly fail to maintain proper type I error even after using INT on the traits. False positive rate can be very high for very low MAC. Interestingly, we observe that POM-LRT becomes more conservative with decrease in MAC. As expected, S_{Hom} exhibits lesser inflation in each scenario. Consistent results are observed for compound symmetry trait correlation structure with moderate and strong correlations (figures not shown), for block-diagonal trait correlation structure (Fig. S8), and for larger sample size of 10,000 (Fig. S17).

Scenario 4: multivariate normal traits with heteroscedasticity

When the genotype at a variant predicts the variance and covariance of the traits, no method (with or without INT on traits) maintains appropriate type I error at the MAC values (6, 30, or 300) we considered (Fig. S9). Unlike the previous scenarios, here POM-LRT is poorly calibrated across all MAC values. As expected, the type I error control worsens for a low count variant. S_{Hom} exhibits lesser inflation in each scenario. We continue to observe similar behavior for compound symmetry trait correlation structure with moderate and strong correlations (figures not shown), for block-

diagonal trait correlation structure (Fig. S10), and for larger sample size of 10,000 (Fig. S18).

Scenario 5: multivariate normal traits

In this ideal situation, all the methods (with or without INT on traits) seem to be well calibrated for MAC 30 or more (Fig. S11). Some data-adaptive methods like USAT and PCO may exhibit slightly inflated type I errors at stringent significance levels. These observations are similar across different correlation structures (Fig. S12), and different sample sizes (Fig. S19) we considered. It is worth noting that although the methods based on summary data assume only asymptotic multivariate normality of the estimated effect sizes, the effective sample size needed for the asymptotics to kick in seems to depend on the underlying multivariate distribution of the traits. For instance, at expected MAC 30, the summary-data based methods are well calibrated when the traits are multivariate normal (Fig. S11) while they show inflation when the traits are multivariate Laplace (Fig. 1) or multivariate t distributed (Fig. S5) (note that individual traits were rank inverse-normalized in all scenarios). For a given MAC, the magnitude of inflation seems to increase with degree of deviation from multivariate normality of traits.

Due to proper calibration of tests, this simulation scenario gives us the opportunity to compare power of these methods (Fig. 2). We find that, in general, the multivariate methods are more powerful when a subset of the traits is associated compared with when all the traits are associated. Such a behavior of multivariate association analyses has been observed and explained before [11, 13]. This behavior is more pronounced when the pairwise trait correlation is stronger. Note that this observation is based on equal and positive genetic effects for the associated traits, and equal and positive pairwise trait correlations. The power of a cross-phenotype test depends on a complex interplay of not only the number, strength, and direction of genetic effects of truly associated traits but also the strength and direction of the pairwise trait correlations [11, 33]. The underlying association scenario changes from one variant to the next, and is not known *a priori* for any real dataset. Here, we observe that the data-adaptive approaches (metaUSAT, MTAR, PCO, mixAda, and USAT) exhibit similar statistical power across all scenarios we studied, and are at least as powerful as metaMANOVA, S_{Het} , MANOVA, and POM-LRT. In addition, similar to what previous studies [11, 22, 33] have shown, we demonstrated massive power gains achieved by a cross-phenotype analysis (e.g., POM-LRT) over multiple single-trait analyses (e.g., Nyholt-Šidák corrected minP [22]) under most scenarios of association (Fig. S13). This commonly used minP (or minimum p -value) approach selects the most significant p -value from

[n = 3000] multivariate normal: mean = 0, covariance = $\Sigma_{cs}(\rho)$

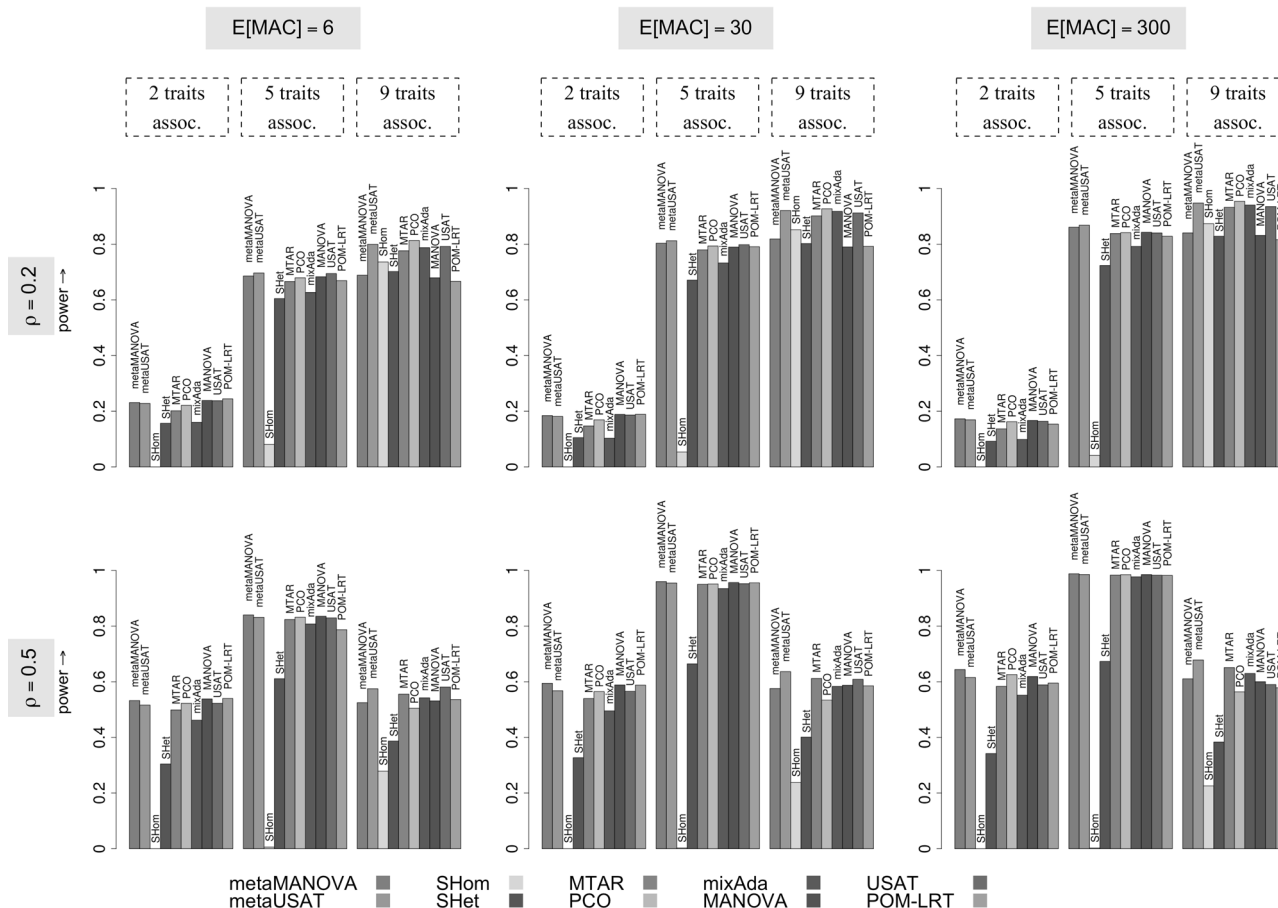


Fig. 2 Scenario 5: Power plots for non-null data (at level $\alpha = 5 = 10^{-8}$) with either first 2, 5, or 9 (out of 10) traits associated with the genetic variant. Each sample has $n = 3000$ unrelated individuals with $K = 10$ multivariate normal traits. The residual covariance matrix is

$\Sigma_{cs}(\rho)$ with $\rho = 0.2$ and 0.5 . Performance of cross-phenotype tests is based on 10,000 replicates. All plots are for raw traits only (the plots are nearly identical for inverse-normalized traits)

the single-trait association tests after correcting for multiple testing using Šidák correction [48], where the approximate number of independent tests are estimated using Nyholt’s approach [49]. The power of a given method is similar irrespective of whether the traits are inverse-normalized or not when the joint trait distribution is indeed multivariate normal (Fig. S14).

Application to Amino Acids Summary Data

Since our real data consists of only summary-level data, we could only apply the methods based on summary data. We analyzed the data using both \hat{R}_{pearson} (Fig. 3)—like we did in our simulation experiments—and \hat{R}_{LDscore} (Fig. S21). We obtained \hat{R}_{LDscore} by applying the `GCVR()` function in MTAR package [37] on the summary data and the pre-computed LD-scores [50] from 1000 Genomes European data. Results using \hat{R}_{pearson} and \hat{R}_{LDscore} are qualitatively similar; so we describe results using \hat{R}_{pearson} only. For

presenting results, we took the median of MACs across traits as the representative MAC for a particular variant. Further, due to over-representation of common variants (>10 million variants with median MAC 300 or more) compared with low-count variants (1.7 million variants with median MAC between 30 and 300), we have only presented the 1.2 million HapMap 3 common SNPs to make the two MAC groups comparable.

The cross-phenotype QQ plots stratified by MAC seem to show early departure from the null when compared with the single-trait QQ plots (Fig. 3a). Given the large effective sample size, it is possible for many common variants to show association signals. However, it seems rather unlikely that so many low-frequency/rare variants (variants with MAC between 30 and 300) are truly associated with at least one of the eight amino acids, the signals for which show up only in the multivariate tests. We plotted cross-phenotype association p -values against the most significant single-trait p -values to get a sense of the proportion of variants that are

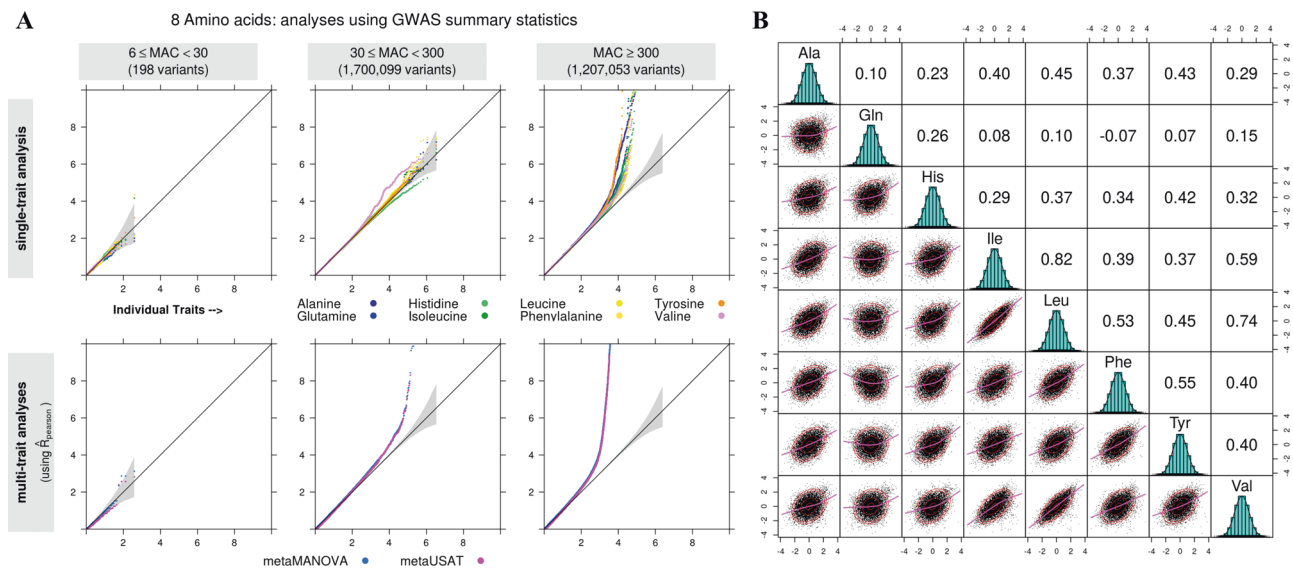


Fig. 3 a. QQ plots for single-trait (univariate) and cross-phenotype (multivariate) association tests of eight amino acid traits using summary statistics. Observed ($-\log_{10} p$ -values) are plotted on the y-axis and expected ($-\log_{10} p$ -values) on the x-axis. All cross-phenotype methods have similar performance, only metaMANOVA and metaUSAT are presented here for demonstration. The gray shaded region represents a conservative 95% confidence interval for the expected distribution of p -values. P -values $\geq 10^{-10}$ are shown here. **b.** The upper diagonal

shows Pearson's pairwise correlation coefficients of the 8 inverse normalized amino acid trait residuals from the METSIM study of >8500 Finnish men. The diagonal depicts the marginal distribution (histogram) of each amino acid. The lower diagonal depicts the scatter plots of pairwise distributions of the traits, where the red ellipses correspond to 10%, 50%, 90%, and 95% contours of standard bivariate normal distribution and the pink curves correspond to fitted local linear regression (LOESS) curves

detected only by cross-phenotype analysis (Fig. S22). For a picture of the amino acid trait distribution, we looked at individual-level amino acid data from a separate study of >8500 Finnish men, METSIM (Metabolic Syndrome in Men) [51]. Note that the summary statistics we analyzed consist of data from many Finnish cohorts [47]. When we looked at pairwise scatter plot of inverse normalized amino acid trait residuals (adjusted for age, age², BMI) [51] from METSIM (Fig. 3B), we found many points systematically distributed outside the 95% contour of a bivariate normal distribution. In addition, fitted local linear regression (LOESS) curves approximating pairwise trait relationships show evidence of non-linearity, which indicates possible deviation of the joint trait distribution from multivariate normality.

Discussion

Overview of this study

This article is an attempt toward identifying advantages and pitfalls of some of the currently-used single-variant cross-phenotype methods in GWAS of rare, low-frequency, and common variants when the basic assumption of multivariate normality is not satisfied. Methods based on individual-level data often assume multivariate normality of traits, while methods based on single-trait summary statistics

assume asymptotic multivariate normality of estimated effect sizes. We compared several popular and new individual-level-data based methods as well as summary-data based methods. Our simulation experiments indicate very poor control of type I error for all but one methods at the rare, low, and common MACs we studied. When the methods are applied on inverse-normalized traits, they continue to show inflated (sometimes severely inflated) type I error rates when the MAC of a genetic variant is low or rare. Although summary-data based methods assume only asymptotic multivariate normality of effect size estimates, the effective sample size at which they are well calibrated increases with deviation of the underlying trait distribution from multivariate normality (note that individual traits were inverse normalized). This is because only the univariate normality of effect size estimates are guaranteed when each trait or trait residual is inverse normalized. We think that for variants with large MAC, when the underlying joint trait distribution is not multivariate normal, the multivariate distribution of single-trait effect size estimates is asymptotically closer to multivariate normal than for variants with low MAC (see Supplementary S1). We found calibration of S_{Hom} to be better than most others because it is a burden-type test—it linearly combines the traits into a single weighted trait, thereby increasing the effective sample size and the effective MAC. Consequently, S_{Hom} has inadequate power to capture heterogeneous effects, which is more likely when analyzing multiple traits. We re-establish MAC (and

not MAF) as the key parameter determining calibration of tests. The MAC threshold after which a test is well calibrated can be much higher for a cross-phenotype test than a single-trait test. We emphasize that the genomic inflation factor (λ_{GC}) may fail to capture systematic bias of association tests, while QQ plot allows us to see the behavior and calibration of such tests across a wide spectrum of significance levels.

POM-LRT (or MultiPhen)—a method requiring individual-level data—is the only method that we found to display type I error that is either appropriate or slightly conservative in all but one scenarios we considered for variants with MAC 30 or more. The POM being based on ordinal regression of genotypes on traits does not require normality of the traits. In addition, statistical power of POM-LRT is comparable to other multivariate methods under multivariate normality of traits, which is consistent with findings from another simulation study modeling complex networks [52]. We also implemented Wald test in this ordinal regression framework (POM-Wald), and found it to be anti-conservative compared with POM-LRT for low-count variants under most simulation scenarios we considered (Figs. S15–S19). This Wald-vs-LRT behavior in POM is opposite to what other researchers [16, 53] found when testing association of rare variants in case-control GWAS using logistic regression model for disease status. Exploring this aspect in more detail is beyond the scope of this article.

As a proof-of-principle example, we performed multivariate association test of eight amino acids. Only single-trait summary statistics from European samples (including Finnish samples) were available to us. We found an excess of statistically significant low-count variants from cross-phenotype (multivariate) analysis—irrespective of the summary-data based method used—compared with single-trait (univariate) analysis. We think it is due to deviations from multivariate normality. We went on to look at pairwise relationships between amino acid traits using individual-level amino acid trait data from METSIM (a study of Finnish men), and found some patterns of non-linearity that indicate possible violation of multivariate normality assumption.

Recommendations

Based on our findings, we recommend in general extra caution when applying cross-phenotype association tests in GWAS with low-frequency or rare variants due to possible violation of multivariate normality assumption. However, we found that robust association testing is still possible for variants with $MAC > 30$ by application of the POM-LRT method, which uses reverse regression modeling (implemented in R program <https://github.com/RayDebashree/mvtests>). We

recommend inverse normalizing each trait residual after accounting for important covariates, and then using the rank-normalized trait residuals to test for genetic association when individual-level data are available. The POM-LRT method, in its current form, cannot handle summary-level data. If only summary-level data are available, one could apply a variety of alternative methods for cross-phenotype association tests, but results may not be robust for genetic variants with MAC below 300. Our recommendation is based on an MAC threshold (instead of an MAF threshold as is commonly used) because we found consistent type I error calibration of methods when the MAC is kept constant. We, additionally, emphasize use of QQ plots, instead of just the genomic inflation factor, to assess calibration of tests at genome-wide levels.

Practical issues with the recommended cross-phenotype method

In a reverse regression framework like POM-LRT, it is, however, unclear how to meaningfully adjust for sample relatedness and population structure. Furthermore, POM-LRT requires observed genotypes or ‘hard-call’ genotypes for imputed variants. It cannot readily incorporate imputation dosage like the usual multivariate linear regression approaches, which may lead to decreased statistical power. Wu and Pankow [54] proposed imputation-score-weighted multinomial regression approach with robust GEE covariance estimates to extend the multi-trait reverse regression model for observed genotypes to imputed SNPs. We did not evaluate the performance of this method though. Another caveat of POM-LRT is its requirement of individual-level data. Restrictions on data sharing necessitate use of summary data. Summary statistics come adjusted for relatedness and population structure, and makes it straightforward and computationally easier to apply multivariate methods on genome-wide summary data. Unfortunately, for summary data on low-count variants, there is no multivariate method that we can recommend when there is concern about the validity of multivariate normality assumption.

Other practical concerns

Inducing approximate multivariate normality of traits

As one of the reviewers pointed out, ‘univariate normality does not imply multivariate normality’ begs the question: can the trait data be more intelligently transformed to induce approximate multivariate normality? If individual-level data were available, one approach can be to identify potential ‘multivariate outliers’ that might be contributing to the breakdown of multivariate normality assumption, and check

sensitivity of cross-phenotype analyses to inclusion/exclusion of such outliers [13]. Detecting outliers in a multi-dimensional space is a challenging problem. We briefly explored two outlier detection approaches under Scenario 2 with multivariate t distributed traits. First approach detects multivariate outliers using sample Mahalanobis Distance (MD) [1, 13] and excludes individuals with significant sample MD at, say, 5% level. Second approach detects univariate outliers for each trait and performs winsorization to limit the influence of outlying trait values in one dimension [14]. We observed some attenuation of inflation—more so when potential multivariate outliers are removed—for all cross-phenotype methods across all MACs (Fig. S24).

Ties in the trait data

It is possible for some traits to exhibit ties (e.g., zero value of some blood measure for multiple individuals). First, our recommendation of inverse normalizing single-trait residuals after necessary covariate adjustments is very likely to break the ties in the traits. Second, the problem of ties can persist in the trait residuals if very few confounders are adjusted (especially if the confounders are not continuous). We briefly explored if INT on tied trait data can affect the calibration of the cross-phenotype tests. In Scenario 2 with multivariate t traits, we artificially created ties in the first five traits to ensure that an average of 10% or 50% individuals have ties for each trait. The resultant joint trait distribution is right skewed with fat tails and many ties. We, then, applied INT on each trait (including the traits with ties). The QQ plots (Fig. S25), when compared with those without ties (Fig. S5), did not reveal any noticeable effect of INT on ties on the calibration of the methods across different MACs.

Study limitations and caveats

This empirical study is not without limitations. First, neither the methods nor the simulation scenarios we considered are exhaustive. We chose a handful of Frequentist methods for our study, none of which is optimized to specifically detect pleiotropic variants (a variant that is associated with at least two traits). Second, our simulation framework is very simple and does not reflect the underlying complex genetic architecture of biological traits. Our simulation study uses ten traits and we have not examined high dimensional traits as is common in neuroimaging and NMR metabolomics. We briefly explored, using 5–30 correlated traits, if our MAC recommendations are dependent on the number of traits being tested. We found that for variants with MAC between 30 and 300, POM-LRT becomes somewhat conservative while the other methods become more inflated

with increase in the number of traits when the underlying trait distribution is not multivariate normal (Figs. S26–S28). Our MAC threshold recommendations may be used as long as the number of traits is between 5 and 30. Our extensive simulations and recommendations are based on continuous traits only. So, we considered an additional limited simulation study comparing performance of summary-level methods on binary traits. In Scenario 5 with multivariate normal traits, we dichotomized each trait at 0 (or 4.23) to ensure, on an average, a 1:1 (or 1:10) case-control distribution for all ten traits. We, then, analyzed each binary trait using logistic regression model and used the resultant GWAS summary statistics to implement cross-phenotype association tests. All summary-level methods seem well-calibrated for binary traits with balanced case-control distribution when MAC is 30 or more (slight deflation is observed for MAC 30), while they exhibit inflation, even at MAC 300, when the case-control distribution gets skewed (Fig. S29). Another limitation is that our simulations do not involve any confounders. We have assumed unrelated individuals without any cryptic relatedness or population structure. Our simplistic power simulations indicate that most multivariate methods have similar statistical power, especially the data-adaptive ones. More sophisticated simulations will probably bring out their differences [52]. Nonetheless, it is important to bear in mind that the aim of the current empirical study is not to determine which method gives better power under what scenario of association when the traits are indeed multivariate normal. Rather, we undertake the first attempt to study how these popular cross-phenotype methods fare when the assumption of multivariate normality fails, especially when testing association with a low-count variant.

Web resources

The URLs for software, codes, and data used in this article are as follows:

metaUSAT v1.17 (implements metaMANOVA, metaUSAT): <https://github.com/RayDebashree/metaUSAT>
 CPASSOC v1.01 (implements S_{Hom} and S_{Het}): <http://hal.case.edu/~xxz10/zhu-web/CPASSOC/>
 MTAR v0.1.0: <https://github.com/baolinwu/MTAR>
 MPAT v1.0 (implements PCO and mixAda): <https://content.sph.harvard.edu/xlin/software.html#mpat>
 USAT v1.21 (implements MANOVA and USAT): <https://github.com/RayDebashree/USAT>
 mvtests v0.3 (implements POM-LRT and Nyholt-Šidák corrected minP): <https://github.com/RayDebashree/mvtests>
 Amino Acids summary data: http://www.computationalmedicine.fi/data#NMR_GWAS

LD scores from 1000 Genomes European data: https://data.broadinstitute.org/alkesgroup/LDSCORE/eur_w_ld_chr.tar.bz2

List of HapMap3 SNPs: https://data.broadinstitute.org/alkesgroup/LDSCORE/hapmap3_snps.tgz

QQ plot code: https://genome.sph.umich.edu/wiki/Code_Sample:_Generating_QQ_Plots_in_R

Manhattan plot code: https://genome.sph.umich.edu/wiki/Code_Sample:_Generating_Manhattan_Plots_in_R

Acknowledgements This research was supported in part by the NIH for the Environmental influences of Child Health Outcomes Data Analysis Center (U24OD023382). It was carried out using computing cluster—the Joint High Performance Computing Exchange—at the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health. We thank Dr Michael Boehnke and Dr Markku Laakso for kindly providing us access to individual level phenotype data on the METSIM amino acid traits. We are grateful to the reviewers for their constructive feedback that immensely helped us improve this article. Finally, DR is thankful to Dr Matthew Stephens for a stimulating conversation on multivariate analyses in GWAS a few years back that sowed the seeds for this article.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

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

References

- Shim H, Chasman DI, Smith JD, Mora S, Ridker PM, Nickerson DA, et al. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS ONE*. 2015;10:e0120758.
- Heid IM, Winkler TW. A multitrait GWAS sheds light on insulin resistance. *Nat Genet*. 2016;49:7–8.
- Liang J, Le TH, Edwards DRV, Tayo BO, Gaulton KJ, Smith JA, et al. Single-trait and multi-trait genome-wide association analyses identify novel loci for blood pressure in African-ancestry populations. *PLoS Genet*. 2017;13:e1006728.
- Shen X, Klaric L, Sharapov S, Mangino M, Ning Z, Wu D, et al. Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation. *Nat Commun*. 2017;8:447. <https://www.nature.com/articles/s41467-017-00453-3>
- Hill WD, Marioni RE, Maghazian O, Ritchie SJ, Hagenaars SP, McIntosh AM, et al. A combined analysis of genetically correlated traits identifies 187 loci and a role for neurogenesis and myelination in intelligence. *Mol Psychiatry*. 2019;24:169–81.
- Jia X, Yang Y, Chen Y, Cheng Z, Du Y, Xia Z, et al. Multivariate analysis of genome-wide data to identify potential pleiotropic genes for five major psychiatric disorders using MetaCCA. *J Affect Disord*. 2019;242:234–43.
- Siewert KM, Voight BF. Bivariate genome-wide association scan identifies 6 novel loci associated with lipid levels and coronary artery disease. *Circ Genom Precis Med*. 2018;11:e002239.
- Inouye M, Ripatti S, Kettunen J, Lyytikäinen LP, Oksala N, Laurila PP, et al. Novel Loci for metabolic networks and multi-tissue expression studies reveal genes for atherosclerosis. *PLoS Genet*. 2012;8:e1002907.
- Marttinen P, Pirinen M, Sarin AP, Gillberg J, Kettunen J, Surakka I, et al. Assessing multivariate gene-metabolome associations with rare variants using Bayesian reduced rank regression. *Bioinformatics*. 2014;30:2026–34.
- Valcarcel B, Ebbels TM, Kangas AJ, Soininen P, Elliot P, Ala-Korpela M, et al. Genome metabolome integrated network analysis to uncover connections between genetic variants and complex traits: an application to obesity. *J R Soc Interface*. 2014;11:20130908.
- Ray D, Pankow JS, Basu S. USAT: a unified score-based association test for multiple phenotype-genotype analysis. *Genet Epidemiol*. 2016;40:20–34.
- Johnson RA, Wichern DW. Applied multivariate statistical analysis. Upper Saddle River, NJ: Prentice hall; 2002.
- Stephens M. A unified framework for association analysis with multiple related phenotypes. *PLoS ONE*. 2013;8:e65245.
- Auer PL, Reiner AP, Leal SM. The effect of phenotypic outliers and non-normality on rare-variant association testing. *Eur J Hum Genet*. 2016;24:1188–94.
- Beasley TM, Erickson S, Allison DB. Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav Genet*. 2009;39:580–95.
- Ma C, Blackwell T, Boehnke M, Scott LJ. GoT2D investigators. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet Epidemiol*. 2013;37:539–50.
- Hackinger S, Zeggini E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol*. 2017;7:170125. <https://doi.org/10.1098/rsob.170125>.
- Muller KE, Peterson BL. Practical methods for computing power in testing the multivariate general linear hypothesis. *Comput Stat Data Anal*. 1984;2:143–58.
- Yang Q, Wang Y. Methods for analyzing multivariate phenotypes in genetic association studies. *J Probab Stat*. 2012;2012:652569.
- Zhou X, Stephens M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat Methods*. 2014;11:407–9.
- Ferreira MA, Purcell SM. A multivariate test of association. *Bioinformatics*. 2008;25:132–3.
- O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FC, Elliott P, Jarvelin MR, et al. MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS ONE*. 2012;7:e34861.
- Majumdar A, Witte JS, Ghosh S. Semiparametric allelic tests for mapping multiple phenotypes: binomial regression and mahalanobis distance. *Genet Epidemiol*. 2015;39:635–50.
- Wu B, Pankow JS. Sequence kernel association test of multiple continuous phenotypes. *Genet Epidemiol*. 2016;40:91–100.
- Kaakinen M, Magi R, Fischer K, Heikkinen J, Jarvelin MR, Morris AP, et al. A rare-variant test for high-dimensional data. *Eur J Hum Genet*. 2017;25:988–94.
- Kim J, Pan W, for the Alzheimer's disease neuroimaging initiative. Adaptive testing for multiple traits in a proportional odds model with applications to detect SNP-brain network associations. *Genet Epidemiol*. 2017;41:259–77.
- Ray D, Basu S. A novel association test for multiple secondary phenotypes from a case-control GWAS. *Genet Epidemiol*. 2017;41:413–26.
- Kim J, Bai Y, Pan W. An adaptive association test for multiple phenotypes with GWAS summary statistics. *Genet Epidemiol*. 2015;39:651–63.
- Deng X, Wang B, Fisher V, Peloso G, Cupples A, Liu CT. Genome-wide association study for multiple phenotype analysis. *BMC Proc*. 2018;12(Suppl 9):55. <https://doi.org/10.1186/s12919-018-0135-8>

30. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet.* 2010;42:348–54.
31. Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, et al. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet.* 2015;96:21–36.
32. Yang Q, Wu H, Guo CY, Fox CS. Analyze multivariate phenotypes in genetic association studies by combining univariate association tests. *Genet Epidemiol.* 2010;34:444–54.
33. Ray D, Boehnke M. Methods for meta-analysis of multiple traits using GWAS summary statistics. *Genet Epidemiol.* 2018;42:134–45.
34. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh PR, et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet.* 2015;47:1236–41.
35. Turley P, Walters RK, Maghziyan O, Okbay A, Lee JJ, Fontana MA, et al. Multi-trait analysis of genome-wide association summary statistics using MTAG. *Nat Genet.* 2018;50:229–37.
36. Qi G, Chatterjee N. Heritability informed power optimization (HIPO) leads to enhanced detection of genetic associations across multiple traits. *PLoS Genet.* 2018;14:e1007549.
37. Guo B, Wu B. Integrate multiple traits to detect novel trait-gene association using GWAS summary data with an adaptive test approach. *Bioinformatics.* 2019;35:2251–7. <https://doi.org/10.1093/bioinformatics/bty961>
38. Xu X, Tian L, Wei LJ. Combining dependent tests for linkage or association across multiple phenotypic traits. *Biostatistics.* 2003;4:223–9.
39. Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. *Bioinformatics.* 2016;32:1981–9.
40. O'Brien PC. Procedures for comparing samples with multiple endpoints. *Biometrics.* 1984;40:1079–87.
41. Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P, 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.* 2012;90:821–35.
42. Liu Z, Lin X. Multiple phenotype association tests using summary statistics in genome-wide association studies. *Biometrics.* 2018;74:165–75.
43. Liu Z, Lin XA. Geometric perspective on the power of principal component association tests in multiple phenotype studies. *J Am Stat Assoc.* 2019;114:975–90. <https://doi.org/10.1080/01621459.2018.1513363>.
44. Dimou NL, Pantavou KG, Braliou GG, Bagos PG. Multivariate methods for meta-analysis of genetic association studies. *Methods Mol Biol.* 2018;1793:157–82.
45. R CoreTeam. R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2017. <https://www.R-project.org>.
46. Basu S, Zhang Y, Ray D, Miller MB, Iacono WG, McGue M. A rapid gene-based genome-wide association test with multivariate traits. *Hum Hered.* 2013;76:53–63.
47. Kettunen J, Demirkan A, Wurtz P, Draisma HH, Haller T, Rawal R, et al. Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of LPA. *Nat Commun.* 2016;7:11122.
48. Sidák Z. On multivariate normal probabilities of rectangles: their dependence on correlations. *Annals Math Stat.* 1968;39:1425–34.
49. Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet.* 2004;74:765–9.
50. Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J. Schizophrenia Working Group of the Psychiatric Genomics Consortium et al. LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet.* 2015;47:291–5.
51. Teslovich TM, Kim DS, Yin X, Stancakova A, Jackson AU, Wielscher M, et al. Identification of seven novel loci associated with amino acid levels using single-variant and gene-based tests in 8545 Finnish men from the METSIM study. *Hum Mol Genet.* 2018;27:1664–74.
52. Porter HF, O'Reilly PF. Multivariate simulation framework reveals performance of multi-trait GWAS methods. *Sci Rep.* 2017;7:38837.
53. Xing G, Lin CY, Wooding SP, Xing C. Blindly using Wald's test can miss rare disease-causal variants in case-control association studies. *Ann Hum Genet.* 2012;76:168–77.
54. Wu B, Pankow JS. Genome-wide association test of multiple continuous traits using imputed SNPs. *Stat Interface.* 2017;10:379–86.