Leveraging omics data to boost the power of genome-wide association studies

Zhaotong Lin, Katherine A. Knutson, and Wei Pan

Summary

Genome-wide association studies (GWASs) have successfully identified many genetic variants and risk loci for complex traits and common diseases in the last 15 years. However, these identified variants, in general, can explain only a small to moderate proportion of the heritability, thus the task of improving GWAS power for more discoveries remains both critical and challenging. In addition to the usual but costly or even infeasible route of continuing to increase the sample size, many approaches have been proposed to incorporate functional annotations to prioritize SNPs but with only limited success. Here, by taking advantage of increasing availability of various types of omics data, we propose a new and orthogonal approach by integrating individual-level omics data with GWASs. The premise is that since omics data reflect both genetic and environmental (such as diet and other lifestyle) effects on individuals, they can be used to account for (otherwise unexplained) variations among individuals in GWAS analysis, leading to more precise/efficient estimation and thus higher power. As a concrete example, we propose boosting GWAS power by adjusting for metabolomics data in GWAS analysis. We applied the method to the UK Biobank subcohort of n = 90,000 individuals with both GWAS and metabolomics data. The analysis of 7 quantitative traits and one binary trait demonstrated clear power gains. For example, the new method (after adjusting for metabolomics data) identified 13 new loci for diastolic blood pressure that were all missed by the standard GWAS, and most or all of the 13 new signals were validated in two much larger GWAS datasets (n = 340,000 and 700,000); the improved estimation efficiency was equivalent to a 38.4% gain of GWAS sample size. The proposed method is both simple and promising and broadly applicable to integrating GWASs with other omics data.

Introduction

Genome-wide association studies (GWASs) have become a powerful and common tool to identify genetic variants and loci associated with complex traits and complex diseases. As of April 7, 2022, the GWAS Catalog had collected 5,690 publications and 372,752 associations, and these findings have provided much insight into the genetic architecture of complex diseases/trait with potential clinical applications. For example, GWAS findings can inform new drug development, drug repurposing, and preventing adverse effects. Risk variants can be used to identify individuals at high risk of certain diseases for whom early prevention or treatment can be implemented. Despite the unprecedented success of GWASs, the detected GWAS variants usually cannot explain more than a modest proportion of the trait heritability, largely due to polygenicity and small effect sizes, limiting the power of even biobank-scale GWASs with large sample sizes in hundreds of thousands. Hence, lack of sufficient GWAS power for weakly associated variants is still a major bottleneck for more new discoveries, and increasing the power of GWASs remains both extremely important and challenging. The most obvious and common route is to continue increasing the sample size, which, however, is not only costly and difficult but also ultimately limited by a finite or even small number of patients with disease. Alternatively, a rich class of approaches have been proposed to incorporate various functional annotations of the genome into GWASs to prioritize Single nucleotide polymorphisms (SNPs) and thus boost the power; however, their applications in practice are not common, presumably due to their limited success.

In this article, instead of (or in addition to) prioritizing SNPs, we take a completely new route to improve the statistical power of GWASs by using omics data to reduce unexplained/residual variations among individuals. Specifically, we investigate whether adjusting for individual-level metabolomics data in GWASs can improve statistical power. Metabolites are small molecules that are intermediate or end products of metabolism. An individual’s metabolic phenotypes are influenced not only by one’s genotypes but also by numerous environmental factors such as diet and other lifestyle variables. For example, metabolites are responsive to one’s diet, and the concentrations of many metabolites change after exercise. On the other hand, standard GWAS analysis only considers marginal SNP-trait associations with adjustment for few covariates, leading to possible confounding and/or large amounts of unexplained variations among the GWAS individuals, ultimately masking SNP-trait associations. We reason that since metabolites reflect both genetic and environmental effects, adjusting for them can account for some otherwise unexplained/residual variations among individuals in GWAS analysis and thus improve the power. This is analogous to using multiple regression to account for other unexplained variations. We applied the method to the UK Biobank subcohort of n = 90,000 individuals with both GWAS and metabolomics data. The analysis of 7 quantitative traits and one binary trait demonstrated clear power gains. For example, the new method (after adjusting for metabolomics data) identified 13 new loci for diastolic blood pressure that were all missed by the standard GWAS, and most or all of the 13 new signals were validated in two much larger GWAS datasets (n = 340,000 and 700,000); the improved estimation efficiency was equivalent to a 38.4% gain of GWAS sample size. The proposed method is both simple and promising and broadly applicable to integrating GWASs with other omics data.
for confounding and extra variations due to covariates, thus improving power over simple regression, though it may not work in some cases, as to be discussed in the discussion section. A common and plausible scenario is that the individuals are independently exposed to various environmental risk factors, which guarantees that adjusting for these risk factors will improve the statistical estimation efficiency and thus power, as in the well-known situation of adjusting for prognostic factors in randomized clinical trials. However, as often most environmental exposures are either difficult to measure or are unknown, metabolomics (or other omics) data can be used as their proxy to improve GWAS power, motivating our proposed method. Through an application to 7 quantitative traits and a binary trait (HTN) in the UK Biobank subcohort of n = 90,000 with both GWASs and metabolomics data available, many new signals were identified after adjustment for metabolomics data, which would be missed by a standard GWAS analysis. Importantly, the majority of these new loci were validated by other larger GWASs. For example, for body mass index (BMI), our new method identified 48 significant loci, including 31 novel ones, compared with 23 from the standard GWAS; our new method yielded much more precise estimates of the SNP-BMI association effect sizes, equivalent to an increase of 38.4% GWAS sample size over the standard GWAS. These results demonstrate that our proposed method, though simple, is promising with broad applications to other GWAS and metabolomics data.

So far, most GWASs with metabolomics data have been focused on uncovering genetic associations with metabolites, while others investigated the relationship between metabolites and diseases such as cardiovascular and kidney diseases. However, to our best knowledge, no previous research has leveraged metabolomics data to boost GWAS power, which is the goal of this work. Furthermore, using the same argument, we foresee that our proposed method and also opens a door for new and important applications of omics data. Finally, the proposed method is orthogonal to increasing sample size and to many existing approaches of integrating functional annotations of the genome. Hence, these approaches can be combined, though we will not pursue that here.

Material and methods

Discovery sample: UK biobank subcohort

UK Biobank (UKB) is a long-term population-based cohort study starting from 2006. It consists of about 500,000 individuals genotyped with around 800,000 SNPs and with a variety of phenotypic information. In this study, we will analyze 8 traits, including seven quantitative ones and a binary disease—self-reported hypertension (HTN). The 7 quantitative traits are BMI, standing height (Height), waist circumference (WC), hip circumference (HC), waist-to-hip ratio (WHR), systolic blood pressure (SBP), and diastolic blood pressure (DBP). 249 nuclear magnetic resonance (NMR) metabolic biomarkers are also available for a subcohort of around 120,000 randomly selected individuals, which will be used in our analysis. Targeted metabolic profiling was undertaken in baseline non-fasting plasma samples using high-throughput NMR spectroscopy. It simultaneously measured 249 metabolic biomarkers in a single assay, spanning multiple metabolic pathways, including lipoprotein lipids in 14 subclasses, fatty acids, and fatty acid compositions, as well as various low-molecular weight metabolites, such as amino acids, ketone bodies, and glycolysis metabolites. This dataset has been used by others, and further details have been described previously.

Validation GWAS summary data

For all of the 8 traits except for WHR, we used the UKB GWAS round 2 results published by the Neale lab in 2018 (http://www.nealelab.is/uk-biobank/) for (partial) validation, which used a larger sample of 361,194 UKB individuals in the analysis. For WHR, we used the GWAS of around 700,000 samples from a meta-analysis of the GIANT and UKB data. In addition, for BMI and Height, we also used the summary statistics of around 700,000 samples from a meta-analysis of the GIANT and UKB data. For WC and HC, we used the results published in 2015 from the GIANT consortium, with the sample size around 140,000. For SBP and DBP, we used the results published in 2018 of around 1 million individuals from the UKB and ICBP consortium.

Quality control and data pre-processing

In the quality control (QC) process, we included self-report White and unrelated individuals with the metabolomics data and the non-missing phenotype in the corresponding GWAS analysis. We further filtered out individuals who are outliers for heterozygosity or missing rate, those who have an abnormal number of sex chromosomes, and those who have unmatched self-reported sex and genetic sex. We also filtered out genetic variants with minor allele frequency less than 0.01, missing genotype rates larger than 0.05, or failing the Hardy-Weinberg equilibrium test at a p value threshold 1e-6. The top 10 genotype principal components (PCs) were calculated based on the remaining set of SNPs and individuals after QC. For missing metabolomics data after QC, we used mean imputation. Then, the top 20 metabolomics PCs were used for dimension reduction, as many of the 249 metabolites were highly correlated. Finally, the quantitative phenotypes were inverse-rank normalized.

Genome-wide association testing

For each of the 8 traits, we used two different GWAS models for single SNP-trait association analysis. First, we performed a standard linear regression between each of the 7 transformed quantitative traits (Y) and each SNP (G), adjusting for covariates (C) including sex, age, and top 10 genotype PCs: Y = β0G + CβC + εY. (Equation 1)

where εY is the random error term. For the second model, as our new method, we additionally adjusted for the top 20 metabolite PCs (M):

Y = β0G + CβC + MβM + εY. (Equation 2)

where εY is the random error term. Here, as usual, we chose using top 20 metabolite PCs based on aiming for both a large proportion...
of the total variation being explained and a relatively small number of PCs compared with the sample size; other choices may be more suitable for other data. We call the new method metabolomics-assisted GWAS (Ma-GWAS).

As for HTN, all steps described above for the quantitative traits remained the same except that logistic regression, instead of linear regression, was used. All GWAS analyses were performed in PLINK 2.0.44 Significant SNPs were identified at the usual p value threshold of 5e−8 and mapped to 1,703 independent linkage disequilibrium (LD) blocks,46 each of which was treated as an independent genomic locus.

One potential issue of mapping SNPs to the 1,703 independent LD blocks is that the SNPs located near and across the boundaries of two neighboring blocks might be in (weak) LD. Hence, as an independent genomic locus.

Removing the genetic components from metabolite PCs

An analysis with adjustment for metabolites takes account of both genetic and environmental effects on metabolite. With genome-wide genotypic data, in principle, we can adjust for genetic effects directly through joint modeling of genotypes (or their PCs), though technical challenges remain. On the other hand, it is often difficult or even unknown to measure all relevant environmental factors. Hence, it would be interesting to account for only environmental effects through metabolites. For this purpose, we propose removing the genetic component from metabolomics data. Specifically, for each trait, we first conducted GWASs for each of the 20 metabolite PCs (mPCs) adjusting for sex, age, and the top 10 genotype PCs. Both the mPCs and covariates were standardized (as in the GWAS for the traits). Then, we obtained the genetic component of each (standardized) mPC by fitting a polygenic risk score with continuous shrinkage prior (PRS-CS) model with the corresponding mPC as the trait:48 we applied PRS-CS with the LD reference panel constructed using the UKB EUR samples and the default parameters provided by the software (https://github.com/gettan107/PRScs/). The genetic component was calculated as the sum of the genetic markers across the genome, weighted by the posterior SNP effect sizes calculated by PRS-CS. We obtained the environmental component as the residual after subtracting the estimated genetic component by PRS-CS from the (standardized) mPC. Finally, we conducted GWASs for each trait, adjusting for sex, age, the top 10 genotype PCs and the environmental components of the top 20 mPCs, denoted as model $M_2$.

It is noted that the genetic components estimated by PRS-CS (or any other modeling method) may not completely capture all genetic effects while at the same time partially reflecting some genetically correlated environmental effects. Hence, our proposed method is just an approximation.

Simulation setups

In the simulation, we assumed that the metabolites only captured environmental effects, and we simulated the traits based on the UKB real data and some fitted model $M_2$ for BMI, SBP, and HTN, respectively, corresponding to three different scenarios. As shown later in Table 4, adjusting for the mPCs (with genetic components removed) in the model additionally explained a relatively large proportion of variance in BMI but only a moderate one for SBP. We also considered HTN as an example of binary trait. We chose three SNPs that were genome-wide significantly associated with BMI, SBP, and HTN in $M_2$, but not in $M_0$, respectively. Specifically, for quantitative traits, we generated the traits as follows:

$$Y = \beta \cdot g + C_1 \cdot b_1 + C_2 \cdot b_2 + e. \quad \text{(Equation 3)}$$

where $g$ was sampled from the UKB individuals' genotypes and standardized, $C_1$ was sampled from the UKB standardized covariates adjusted in $M_0$ (including the first 10 genotype PCs, sex, and age), $C_2$ was sampled from the UKB top 20 standardized mPCs (with the genetic components removed), and $b_1$ and $b_2$ were the
standardized estimates from the null model of $M_2$, i.e., $Y = \text{sex + age + genotype PCs + mPCs(no genetic comp)}$ with both the outcome and predictors standardized. $\epsilon$ is generated from $\mathcal{N}(0, \sigma^2)$, where $\sigma^2 = 1 - \text{var} (\beta g + C_1 b_1 + C_2 b_2)$. For the binary trait, we generated the binary outcome as follows:

$$
\alpha = \beta g + C_1 b_1 + C_2 b_2,
$$

$$
p = \frac{\exp (\alpha)}{1 + \exp (\alpha)},
$$

and

$$
Y_i \sim \text{Bernoulli}(p_i),
$$

where $g, b_1, b_2, C_1$, and $C_2$ were generated the same way as before, $p$ is the vector of probabilities, and the outcome $Y_i$ was generated from the Bernoulli distribution with corresponding probability $p$. The GWAS association effect $\beta$ was assigned a zero value and a non-zero value, where the non-zero value was determined by the corresponding standardized estimates in $M_2$ GWASs for the three traits. The sample sizes were also assigned according to the real data analysis, which were 90,000 for BMI and HTN and 85,000 for SBP.

Figure 1. Manhattan plots for BMI GWAS results
Top 2: the standard GWAS and proposed new method (without and with adjustment for metabolomics data) for the discovery sample (i.e., UKB sub-cohort). Bottom 2: two validation GWAS datasets.
respectively, and the third row is the number of significant
by the standard GWAS (Table 2 shows the main results for the 8 traits. The first and
The new method identified more trait-associated loci
and mapped them to independent loci. We validated the
respectively. We identified significant SNPs with p
An adjusted
$R^2$.

The main motivation for adjusting for metabolites in
GWASs for more efficient estimation to improve power
is through reduction of residual variances. To investigate
this, we fitted models with only the usual covariates
including sex, age, and top genotype PCs and the new
models with sex, age, genotype PCs, and mPCs as predic-
tors. Table 4 shows the coefficients of determination,
R-squared ($R^2$), of the two models for each of the 8 traits.
It is confirmed that additionally adjusting for mPCs led to
a larger $R^2$ and thus accounting for more variations of a trait,
implying more efficient estimation by the new method.
Next, we directly compared the estimated effect sizes of
SNPs and their corresponding SEs from the standard
GWAS and our proposed new GWAS models as shown in
Equations 1 and 2. To save space, we only presented the

For each setup, the simulation was replicated 1,000 times. We report the mean and standard deviation (SD) of the $\beta$ estimates and the mean of the estimated standard errors (SEs).

**Results**

After QC, there were around 90,000 individuals and
580,000 SNPs left in the analysis. Table 1 shows the sample
sizes for the discovery and validation GWAS datasets.
For each of the 8 traits, we performed two genome-wide
association analyses without or with adjustment for the top 20
mPCs, corresponding to the standard GWAS and our proposed new Ma-GWAS, denoted as models $M_0$ and $M_1$, respectively. We identified significant SNPs with $p < 5e-8$ and mapped them to independent loci. We validated the
new signals using other validation GWAS summary data.

**The new method identified more trait-associated loci**

Table 2 shows the main results for the 8 traits. The first and
second rows are the numbers of significant loci identified by
the standard GWAS ($M_0$) and our new Ma-GWAS ($M_1$), respectively, and the third row is the number of significant loci identified by the new method ($M_1$) but missed by the standard GWAS ($M_0$). The fourth and fifth rows are the
numbers of those new signals (in the third row) that could
be validated in the UKB Neale lab data and other larger
GWAS data, respectively. For example, for trait BMI, the
standard GWAS $M_0$ identified 23 loci, while the new
method ($M_1$) with additional adjustment for metabolomics
data identified 48 loci. There were 31 new loci identified
only by $M_1$ but missed by $M_0$, compared with 6 loci identified
by $M_0$ but not by $M_1$. Furthermore, 17 of the 31 new
loci were validated in the BMI GWAS results conducted
by the Neale lab. In addition, with a larger sample size of
around 700,000, 26 of the 31 new loci were validated by
another validation GWAS. The Manhattan plots of the
4 sets of the GWAS results for BMI are shown in Figure 1,
while those for other traits are given in the supplemental
information (Figures S1–S7).

It is noted that most of the new loci were validated in the
larger GWAS. In particular, for Height, SBP, and DBP, all new
loci were validated. For DBP, while all 13 loci identified by
the standard GWAS $M_0$ were also identified by the new
method $M_1$, the new method $M_1$ identified 13 more new
loci, which were all validated by the other larger GWAS.

Table 3 shows the result of the numbers of new loci and
validated ones for the 8 traits based on the new definition
of genomic loci given by FUMA. It is confirmed again
that the proposed new method, by adjusting for metabolo-
mics data in GWAS analysis, identified additional new loci
for all 8 traits, and many of them were replicated in the
validation GWAS data, though the numbers of validated
new loci were slightly smaller than that given by the other
definition of genomic loci as shown in Table 2.

### Table 3. Results for the 8 traits (with loci defined by FUMA)

| Trait | BMI | Height | SBP | DBP | WC | HC | WHR | HTN |
|-------|-----|--------|-----|-----|----|----|-----|-----|
| New loci | 31 | 6 | 4 | 13 | 26 | 32 | 22 | 7 |
| Validated new loci in Neale lab UKB | 13 | 5 | 3 | 10 | 10 | 25 | NA | 5 |
| Validated new loci in other dataset | 21 | 6 | 3 | 11 | 4 | 7 | 11 | NA |

The new method improved estimation efficiency. From the top row to the bottom, the number of new loci identified by the proposed new method, and the numbers of new loci validated by the Neale lab results and by other GWAS results, respectively.

### Table 4. $R^2$ of the (null) standard GWAS model and our proposed new GWAS models with top 20 metabolite PCs (before and after removing the genetic components)

| Model | $R^2$ (null) | $R^2$ (Ma-GWAS) |
|-------|-------------|-----------------|
| $Y = \text{sex \& age + genotype PCs}$ | 0.011 | 0.528 |
| $Y = \text{sex \& age + genotype PCs + mPCs}$ | 0.293($**$) | 0.537($**$) |
| $Y = \text{sex \& age + genotype PCs + mPCs (no genetic comp)}$ | 0.305 | 0.536 |

An adjusted $R^2$ for each quantitative trait and Nagelkerke's $R^2$ for the binary trait were used.

*The $p$ value from the likelihood ratio test comparing the full model 1 (including mPCs) with the reduced model 0 (without mPCs) was highly significant (smaller than 2.2e−16).
representative results for BMI here, while the detailed results for the other 7 traits are in the supplemental information (see Figures S8–S13). The top panel in Figure 2 compares the SEs: each point represents one SNP, and we only kept SNPs with p values less than 1e−4 in either one of the two models. It is clear that the new method $M_1$ (adjusting for metabolomic PCs) yielded more precise estimates with smaller SEs. Specifically, the SEs from the new method were 0.85 of those from the standard GWAS, a reduction of 15%, equivalent to a 1/0.852−1 = 38.4% gain in sample size by the new method over the standard GWAS. For the six other quantitative traits, i.e., Height, SBP, DBP, WC, HC, and WHR, the reductions of their SEs by the new method were 1%, 3%, 4%, 17%, 9%, and 13%, equivalent to the GWAS sample size gains of 2%, 6.3%, 8.5%, 45.2%, 20.8%, and 32.1%, respectively. It is notable that the magnitudes of the reduced SEs by the new method were proportional to the increased $R^2$ explained by the mPCs for the quantitative traits as shown in Table 4.

The bottom panel in Figure 2 compares the estimated effect sizes for SNPs that were significant in either one of the two models (p < 5e−8). The colors of the points correspond to the SNPs that were significant only in $M_0$ (red), significant only in $M_1$ (green), and significant in both $M_0$ and $M_1$ (blue). The triangles correspond to the SNPs that were validated in either one of the two validation GWAS datasets, and the circles correspond to the SNPs that were not validated or did not appear in the validation GWAS data. Two patterns are observed for the new signals (i.e., the green points that were only significant in $M_1$). The first is the presence of a subset of the points close to the identity line (i.e., the 45° diagonal line), meaning that the two estimates from the two models were similar. These new signals were only identified by $M_1$ largely because of its more precise estimation, and most of these new signals were also validated by at least one larger GWAS dataset. The second pattern corresponds to the remaining (green) points farther away from the identity line, suggesting that their statistical significance was (partly) due to their larger effect sizes estimated by $M_1$. Similar observations held for the other quantitative traits except for Height. For height, almost all points were aligned on or around the diagonal line, i.e., the estimated effect sizes were similar from the two models, which, along with their closer SEs, suggests that height is likely less affected by factors measured or manifested by metabolites. For more details on the other quantitative traits, see Figures S10–S13.
For binary trait HTN, both the estimated effect sizes and their SEs from the new method were slightly larger than their counterparts from the standard GWAS. While the larger SEs for (binary) HTN may be surprising, given that our method always yielded smaller SEs for the quantitative traits, this finding is due to non-linear logistic regression (compared with linear regression). This is in agreement with the literature: adjusting for important covariates such as prognostic factors (corresponding to metabolomics data here) for a binary outcome with an independent and non-null treatment (corresponding to an SNP here) in logistic regression for a randomized clinical trial would always lead to both a larger treatment effect estimate and a larger SE, but at the end, the estimation efficiency, and thus the power, will always be improved, as confirmed by more genetic loci identified by the new method shown earlier.

By default, we used the sample of self-reported White individuals, for which, as suggested by a reviewer, population stratification might still exist after adjusting for genotype PCs. We provided the QQ-plots for the GWAS scans on the 8 traits in the supplemental information (Figures S14 and S15). Only Height had a much-inflated inflation factor of 1.56, while the other 7 traits had inflation factors between 1.1 and 1.22. We also note that the Neale lab UKB GWAS data for Height also had a highly inflated inflation factor of 1.8. To alleviate the issue of population stratification, we further restricted the sample to those self-identified as “White British” and having a similar genetic ancestry (based on UKB data field 22006) and conducted the same analysis. The inflation factors were reduced for all traits, in particular from 1.56 to 1.35 for Height. Similarly, the analysis using the more restricted sample of individuals also supported our findings that novel loci were identified after adjusting for mPCs. The QQ-plots and other detailed results are given in Figures S16 and S17.

Why top 20 mPCs?
We chose to include the top 20 mPCs in the proposed model mainly because they explained a high proportion of the total variation (about 95%) in the 249 metabolite biomarkers, while the number of top mPCs, 20 here, was small compared with the sample size of GWASs. We also performed the same analysis but with the top 5 and top 10 mPCs, which explained around 81% and 89% of the total variance, respectively. We found that the results were similar to that with the top 20 mPCs—multiple novel loci were identified compared with the standard GWAS, and most of the novel loci were also validated by other GWAS data. Detailed results are given in the supplemental information (Tables S1–S3). In addition, we looked at the top 10 metabolites contributing most to each of the top 20 mPCs. We found that out of the 18 metabolite classes present in the UKB data, the top 10 metabolites of the top 5, top 10, and top 20 mPCs covered 12, 16, and 18 metabolite classes, respectively. This suggests that including the top 20 PCs covered all classes of metabolites that might be able to explain more background genetic and environmental effects, while including only the top 5 or top 10 PCs would miss some of them.

Removing the genetic components from mPCs
To focus on adjusting for only the environmental effects reflected in metabolite, we removed the genetic component from each of the top 20 mPCs, then conducted the GWAS on each trait. As shown in Table 5, we can see a similar pattern (as in Table 2) that adjusting for the environmental components of metabolites in M2 identified new loci beyond those by the standard GWAS model M0, though the number of new loci identified by M2 was less than that by M1. Furthermore, most of the new loci identified by M2 were validated as well.

As shown in Table 4, interestingly, the R²’s in model 2 (after removing the genetic components in the top mPCs) were quite close to, though all except one were slightly larger than, those in model 1 (before removing the genetic components), suggesting that the mPCs largely and perhaps even mostly represented environmental effects on the traits, thus offering one explanation on why adjusting for mPCs improved power.

As before (i.e., without removing the genetic components from mPCs), we compared the estimated effect sizes of SNPs and their corresponding SEs from the standard GWAS M0 and our proposed new GWAS M2 for BMI, while the detailed results for the other 7 traits are in the supplemental information (Figures S18–S23). As shown in the top panel in Figure 3, the new method M2 (adjusting for

| Traits   | Top 20 mPCs | Top 5 mPCs | Top 10 mPCs | Top 20 mPCs Validated | Top 5 mPCs Validated | Top 10 mPCs Validated | Top 20 mPCs Validated | Top 5 mPCs Validated | Top 10 mPCs Validated |
|----------|-------------|------------|-------------|----------------------|---------------------|-----------------------|----------------------|---------------------|-----------------------|
| BMI      | 23          | 22         | 22          | 23                   | 22                  | 23                    | 22                   | 22                  | 23                    |
| Height   | 232         | 225        | 225         | 230                  | 225                 | 230                   | 230                  | 225                 | 230                   |
| SBP      | 16          | 15         | 15          | 17                   | 16                  | 17                    | 17                   | 16                  | 17                    |
| DBP      | 13          | 12         | 12          | 14                   | 13                  | 14                    | 14                   | 13                  | 14                    |
| WC       | 16          | 15         | 15          | 17                   | 16                  | 17                    | 17                   | 16                  | 17                    |
| HC       | 23          | 22         | 22          | 24                   | 23                  | 24                    | 24                   | 23                  | 24                    |
| WHR      | 15          | 14         | 14          | 16                   | 15                  | 16                    | 16                   | 15                  | 16                    |
| HTN      | 15          | 14         | 14          | 16                   | 15                  | 16                    | 16                   | 15                  | 16                    |
environmental components of mPCs) yielded more precise estimates with smaller SEs. Specifically, the SEs from the new method were 0.84 of those from the standard GWAS. For the other six quantitative traits, the results were similar as before: the reductions of their SEs by the new method ranged from 1% to 18%. On the other hand, as shown in the bottom panel in Figure 3, the estimated effect sizes for SNPs by the new method were more similar to those from the standard GWAS. For the binary trait HTN, both the SEs and the estimated effect sizes by the new method were slightly larger than those from the standard GWAS, as observed and explained before.

Simulation results
On each simulated dataset, we applied the standard GWAS that only adjusted for the usual covariates including genotype PCs, sex, and age (C1 in Equation 3) and our proposed Ma-GWAS that adjusted for mPCs (C2 in Equation 3) in addition to the usual covariates. The left column in Figure 4 shows the box plots of the estimated GWAS associations across 1,000 replicates under different simulation setups, and we also added the corresponding type I error (T1E) and power. We can see that the standard GWAS gave biased estimates due to ignoring the mPCs as confounders (because mPCs were correlated with both the genotypes and the traits) and sometimes inflated T1E, while Ma-GWAS gave unbiased estimates with well-controlled T1E under the true model. Importantly Ma-GWAS was more powerful than the standard GWAS. The right column in Figure 4 compares the estimated SEs of $\beta$ between the two methods. Here, we only show the plots for a non-zero $\beta$ for each trait, but the results were similar when $\beta = 0$ (see Table S5). In agreement with our real data analysis, Ma-GWAS gave more precise estimates with smaller SEs for the two quantitative traits. For BMI, the metabolites could capture a larger portion of environmental effects, so the ratio of the estimated SEs between the two methods was smaller around 84%; in comparison, for SBP, the ratio was larger around 97%. For (binary) HTN, similar to the real data analysis results, we again observed that both the SEs and the estimated effect sizes for non-zero $\beta$ by Ma-GWAS were larger than those from the standard GWAS. In summary, the simulation results confirmed that adjusting for metabolites gave more precise/efficient estimation and thus higher power as we observed in the real data analysis.
Discussion

We have investigated and confirmed that adjusting for metabolomics data in GWAS analysis can boost its power. In an application to 7 quantitative traits and one binary disease of UKB samples with metabolomics data, we found that adjusting for the top 20 metabolomics PCs in GWASs substantially improved the power over the standard GWAS; most of the newly detected trait-associated loci were validated by larger GWASs. The main reason is that metabolomics (and other omics) data reflect both genetic and environmental influences, thus using them to account for additional variations among individuals is expected to reduce the residual variance in GWAS, leading to more precise/efficient estimation of the effect sizes of the SNPs, thus improving power. In particular, across the 7 quantitative traits, the reduction of the SEs of the estimated SNP effect sizes ranged from 1% to 13%, equivalent to the GWAS sample size gains from 2.0% (for Height) to 45.2% (for WC). Furthermore, after regressing out the genetic components from the mPCs, we obtained similar GWAS results, suggesting that perhaps the estimation and power gains from adjusting for mPCs mainly came from accounting for environmental effects reflected in metabolomics data.

There are several directions for future study. First, since some metabolites may mediate the effects of some SNPs on a trait, adjusting for these metabolites may lead to reduced power while in some other situations leading to biased inference. Our analysis results suggest that the potential gain outweighs the possible loss in power, especially if we take the proposed approach as complementary to the standard GWAS approach. Alternatively, we may regress out genetic components from metabolomics data (e.g., mPCs, as done in our analysis) before adjusting for mPCs. Nevertheless, to better deal with the problem, it is important to have a better understanding of the causal mechanism underlying genetic variants, metabolites, and a trait of interest. Relatedly, perhaps there are certain types of metabolites more suitable to be adjusted for a given GWAS trait. Similarly, there might be certain categories of traits more likely to benefit more from the adjustment of (certain types of) metabolites in GWASs. This could be due to that some traits were more largely and directly influenced by genetic and/or environmental factors (e.g., those involved in or influencing metabolic pathways) that are also better captured by the available metabolomics data, as partially reflected by their larger $R^2$'s explained by metabolic data. Hence, an $R^2$ explained by metabolic data for a trait may be used to predict how useful the given metabolic data for the trait, thus possibly informing the choice of the metabolic data and/or the trait. More studies are needed. Second, our proposed method is orthogonal to many existing approaches of incorporating the use of functional annotations of the genome and thus can be applied on top of these methods. It would be interesting to see how such a combined approach would perform in real data applications. Third, in...
this study, to adjust for the hidden genetic and environmental effects manifested by metabolomics data, we simply included the top 20 PCs obtained from the 249 metabolites as fixed effects in the GWAS regression model. There could be other better approaches to integrate metabolomics (or other omics) data in GWAS. Lastly, more real data applications, possibly with various omics data, are warranted.

Data and code availability

To access the discovery sample (UKB individual-level data), one is required to submit an application to UKB for approval. All the validation GWAS summary datasets are publicly available. The complete GWAS summary statistics for the 8 traits analyzed in the paper have been made available via https://doi.org/10.5281/zenodo.7118030. The details of metabolomics biomarkers can be found at https://biobank.ndph.ox.ac.uk/ukb/label.cgi?id=220. The R code for the real data analysis and simulation is available at https://github.com/ZhaotongL/Ma-GWAS.

Supplemental information

Supplemental information can be found online at https://doi.org/10.1016/j.xhgg.2022.100144.

Acknowledgments

We thank the editors and reviewers for many helpful comments. This study was supported by NIH grants R01 AG065636, R01 AG069895, RF1 AG067924, U01 AG073079, R01 HL116720, and R01 GM126002 and by the Minnesota Supercomputing Institute at the University of Minnesota. Access to the UKB data was through application #35107.

Author contributions

Z.L. and W.P. co-developed the methods; K.A.K. downloaded the UKB data; Z.L. downloaded other data, implemented the methods and did all analyses; Z.L. drafted the paper, and K.A.K. and W.P. proofread it.

Declaration of interests

The authors declare no competing interests.

Received: June 24, 2022
Accepted: September 15, 2022

References

1. Wang, W., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. Nat. Rev. Genet. 6, 109–118.
2. Buriel, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., et al. (2019). The nhgri-ebi gwas catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. Nucleic Acids Res. 47, D1005–D1012.
3. Giacomini, K.M., Yee, S.W., Mushiroda, T., Weinshilboum, R.M., Ratain, M.J., and Kubo, M. (2017). Genome-wide association studies of drug response and toxicity: an opportunity for genome medicine. Nat. Rev. Drug Discov. 16, 1–70.
4. Chen, W., Stambolian, D., Edwards, A.O., Branhman, K.E., Othman, M., Jakobsdottir, J., Tosakulwong, N., Pericak-Vance, M.A., Campochiaro, P.A., Klein, M.L., et al. (2010). Genetic variants near timp3 and high-density lipoprotein-associated loci influence susceptibility to age-related macular degeneration. Proc. Natl. Acad. Sci. USA 107, 7401–7406.
5. Thorleifsson, G., Magnussson, K.P., Sulem, P., Walters, G.B., Gudbjartsson, D.F., Stefansson, H., Jonsson, T., Jonasdottir, A., Jonasdottir, A., Stefansdottir, G., et al. (2007). Common sequence variants in the lvol1 gene confer susceptibility to exfoliation glaucoma. Science 317, 1397–1400.
6. Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature 461, 747–753.
7. Chung, D., Yang, C., Li, C., Gelernter, J., and Zhao, H. (2014). Gpa: a statistical approach to prioritizing gwas results by integrating pleiotropy and annotation. PLoS Genet. 10, e1004787.
8. Gao, J., Espin-Garcia, O., Paterson, A.D., and Sun., L. (2021). Current methods integrating variant functional annotation scores have limited capacity to improve the power of genome-wide association studies. Preprint at bioRxiv. https://doi.org/10.1101/2020.11.25.396721.
9. Ho, Y.-Y., Baechler, E.C., Ortmann, W., Behrens, T.W., Graham, R.R., Bhangale, T.R., and Pan, W. (2014). Using gene expression to improve the power of genome-wide association analysis. Hum. Hered. 78, 94–103.
10. Ionita-Laza, I., McQueen, M.B., Laird, N.M., and Lange, C. (2007). Genomewide weighted hypothesis testing in family-based association studies, with an application to a 100k scan. Am. J. Hum. Genet. 81, 607–614.
11. Ionita-Laza, I., McCallum, K., Xu, B., and Buxbaum, J.D. (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nat. Genet. 48, 214–220.
12. Kichaev, G., Bhatia, G., Lob, P.-R., Gazal, S., Burch, K., Freund, M.K., Schoech, A., Pasaniuc, B., and Price, A.L. (2019). Leveraging polygenic functional enrichment to improve gwas power. Am. J. Hum. Genet. 104, 65–75.
13. Li, X., Li, Z., Zhou, H., Gaynor, S.M., Liu, Y., Chen, H., Sun, R., Dey, R., Arnett, D.K., Aslibekyan, S., et al. (2020). Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale. Nat. Genet. 52, 969–983.
14. Lu, Q., Yao, X., Hu, Y., and Zhao, H. (2016). Genowap: gwas signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics 32, 542–548.
15. Lu, Q., Powles, R.L., Wang, Q., He, B.J., and Zhao, H. (2016). Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. PLoS Genet. 12, e1005947.
16. Nicolae, D.L., Gamazon, E., Zhang, W., Duan, S., Dolan, M.E., and Cox, N.J. (2010). Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. PLoS Genet. 6, e1000888.
17. Roeder, K., Bacanu, S.-A., Wasserman, L., and Devlin, B. (2006). Using linkage genome scans to improve power of association in genome scans. Am. J. Hum. Genet. 78, 243–252.
18. Roeder, K., Devlin, B., and Wasserman, L. (2007). Improving power in genome-wide association studies: weights tip the scale. Genet. Epidemiol. 31, 741–747.

19. Schork, A.J., Thompson, W.K., Pham, P., Torkamani, A., Rodgers, J.C., Sullivan, P.F., Kelsoe, J.R., O’donovan, M.C., Furberg, H.; and Tobacco and Genetics Consortium (2013). All snps are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated snps. PLoS Genet. 9, e1003449.

20. Xue, H., Wu, C., and Fan, W. (2020). Leveraging existing gwas summary data of genetically correlated and uncorrelated traits to improve power for a new gwas. Genet. Epidemiol. 44, 717–732.

21. Zhong, H., Yang, X., Kaplan, L.M., Molony, C., and Schadt, E.E. (2010). Integrating pathway analysis and genetics of gene expression for genome-wide association studies. Am. J. Hum. Genet. 86, 581–591.

22. Akbaraly, T., Würtz, P., Singh-Manoux, A., Shipley, M.J., Haapasoksi, R., Lehto, M., Desrumaux, C., Kähönen, M., Lehtimäki, T., Mikkilä, V., et al. (2018). Association of circulating metabolites with healthy diet and risk of cardiovascular disease: analysis of two cohort studies. Sci. Rep. 8, 8620–8714.

23. Johnson, R.K., Vanderlinden, L., DeFelice, B.C., Keckis, K., Uusitalo, U., Flehn, O., Sontag, M., Crume, T., Beyerlein, A., Lernmark, Å., et al. (2019). Metabolite-related dietary patterns and the development of islet autoimmunity. Sci. Rep. 9, 14819–14911.

24. Schranzer, D., Kastenmüller, G., Schönfelder, M., Römisch-Margl, W., and Wackerhage, H. (2020). Metabolite concentration changes in humans after a bout of exercise: a systematic review of exercise metabolomics studies. Sports Med. Open 6, 11–17.

25. Hauck, W.W., Anderson, S., and Marcus, S.M. (1998). Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control. Clin. Trials 19, 249–256.

26. Kahan, B.C., Jairath, V., Dore, C.J., and Morris, T.P. (2014). The risks and rewards of covariate adjustment in randomized trials: an assessment of 12 outcomes from 8 studies. Trials 15, 139–147.

27. Gallois, A., Mefford, J., Ko, A., Vaysse, A., Julienne, H., Ala-Korpela, M., Laakso, M., Zaitlen, N., Pajukanta, P., and Ashhard, H. (2019). A comprehensive study of metabolite genetics reveals strong pleiotropy and heterogeneity across time and context. Nat. Commun. 10, 4788–4813.

28. Kettunen, J., Demirkan, A., Würtz, P., Draisna, H.H.M., Haller, T., Rawal, R., Vaarhorst, A., Kangas, A.J., Lyytikäinen, L.P., Pirinen, M., et al. (2016). Genome-wide study for circulating metabolites identifies 62 loci and reveals novel systemic effects of lpa. Nat. Commun. 7, 11122–11129.

29. Feofanova, E.V., Chen, H., Dai, Y., Jia, P., Grove, M.L., Morris, A.C., Qi, Q., Daviglus, M., Cai, J., North, K.E., et al. (2020). A genome-wide association study discovers 46 loci of the human metabolome in the hispanic community health study/study of latinos. Am. J. Hum. Genet. 107, 849–863.

30. Yin, X., Chan, L.S., Bose, D., Jackson, A.U., VanDeHaar, P., Locke, A.E., Fuchsberger, C., Stringham, H.M., Welch, R., Xu, K., et al. (2022). Genome-wide association studies of metabolites in Finnish men identify disease-relevant loci. Nat. Commun. 13, 1644–1714.

31. Wang, Z., Zhu, C., Nambi, V., Morrison, A.C., Folsom, A.R., Ballantyne, C.M., Boerwinkle, E., and Yu, B. (2019). Metabolomic pattern predicts incident coronary heart disease: findings from the atherosclerosis risk in communities study. Arterioscler. Thromb. Vasc. Biol. 39, 1475–1482.

32. Yu, B., Flexeder, C., McGarrah, R.W., III, Wyss, A., Morrison, A.C., North, K.E., Boerwinkle, E., Kastenmüller, G., Gieger, C., Suhre, K., et al. (2019). Metabolomics identifies novel blood biomarkers of pulmonary function and copd in the general population. Metabolites 9, 61.

33. Mi, K., Jiang, Y., Chen, J., Lv, D., Qian, Z., Sun, H., and Shang, D. (2020). Construction and analysis of human diseases and metabolites network. Front. Bioeng. Biotechnol. 8, 398.

34. Iliou, A., Mikros, E., Karaman, I., Elliott, F., Griffin, J.L., Tzoulaki, I., and Elliott, P. (2021). Metabolic phenotyping and cardiovascular disease: an overview of evidence from epidemiological settings. Heart 107, 1123–1129.

35. Steinbrenner, I., Schultheiss, U.T., Kotsis, F., Schlosser, P., Stockmann, H., Mohnyp, R.P., Schmid, M., Oefner, P.J., Eckardt, K-U., Kötgen, A., et al. (2021). Urine metabolite levels, adverse kidney outcomes, and mortality in ckd patients: a metabololome-wide association study. Am. J. Kidney Dis. 78, 669–677.e1.

36. Zhang, J., Dutta, D., Kötgen, A., Tin, A., Schlosser, P., Grams, M.E., Harvey, B., CKDGen Consortium, Boerwinkle, E., Corsh, J., and Chatterjee, N. (2022). Plasma proteome analyses in individuals of european and african ancestry identify csi-pqtl and models for proteome-wide association studies. Nat. Genet. 54, 593–602.

37. Bycroft, C., Freeman, C., Petkova, D., Band, G., Elliott, L.T., Sharp, K., Motyer, A., Vuckevic, D., Delaneau, O., O’Connell, J., et al. (2018). The UK biobank resource with deep phenotyping and genomic data. Nature 562, 203–209.

38. Würtz, P., Kangas, A.J., Soininen, P., Lawlor, D.A., Davey Smith, G., and Ala-Korpela, M. (2017). Quantitative nuclear magnetic resonance metabolomics in large-scale epidemiology: a primer on omic technologies. Am. J. Epidemiol. 186, 1084–1096.

39. Julkunen, H., Cichońska, A., Slagboom, P.E., Würzt, P.; and Nightingale Health UK Biobank Initiative (2021). Metabolic biomarker profiling for identification of susceptibility to severe pneumonia and covid-19 in the general population. Elife 10, e63033.

40. Bragg, F., Trichia, E., Aguilar-Ramirez, D., Bešević, J., Lewington, S., and Emberson, J. (2022). Predictive value of circulating NMR metabolic biomarkers for type 2 diabetes risk in the UK biobank study. BMC Med. 20, 159–212.

41. Pulit, S.L., Stoneman, C., Morris, A.P., Wood, A.R., Glastonbury, C.A., Tyrrell, J., Yengo, L., Ferreira, T., Marouli, E., Ji, Y., et al. (2019). Meta-analysis of genome-wide association studies for body fat distribution in 694 649 individuals of european ancestry. Hum. Mol. Genet. 28, 166–174.

42. Yengo, L., Sidorenko, J., Kemper, K.E., Zheng, Z., Wood, A.R., Weedon, M.N., Frayling, T.M., Hirschhorn, J., Yang, J., Visscher, P.M.; and GIANT Consortium (2018). Meta-analysis of genome-wide association studies for height and body mass index in 700000 individuals of european ancestry. Hum. Mol. Genet. 27, 3641–3649.

43. Shungin, D., Winkler, T.W., Croteau-Chonka, D.C., Ferreira, T., Locke, A.E., Magi, R., Strawbridge, R.J., Pers, T.H., Fischer, K., Anne, E.J., et al. (2015). New genetic loci link adipose and insulin biology to body fat distribution. Nature 518, 187–196.

44. Evangelou, E., Warren, H.R., Mosen-Ansorena, D., Mifsud, B., Pazzini, R., Gao, H., Ntritsos, G., Dimou, N., Cabrera, C.P.
Karaman, I., et al. (2018). Genetic analysis of over 1 million people identifies 535 new loci associated with blood pressure traits. Nat. Genet. 50, 1412–1425.

45. Chang, C.C., Chow, C.C., Tellier, L.C.A.M., Vattikuti, S., Purcell, S.M., and Lee, J.J. (2015). Second-generation plink: rising to the challenge of larger and richer datasets. GigaScience 4, 7–015.

46. Berisa, T., and Pickrell, J.K. (2016). Approximately independent linkage disequilibrium blocks in human populations. Bioinformatics 32, 283–285.

47. Watanabe, K., Taskesen, E., Van Bochoven, A., and Posthuma, D. (2017). Functional mapping and annotation of genetic associations with fuma. Nat. Commun. 8, 1826–1911.

48. Ge, T., Chen, C.-Y., Ni, Y., Feng, Y.C.A., and Smoller, J.W. (2019). Polygenic prediction via bayesian regression and continuous shrinkage priors. Nat. Commun. 10, 1776–1810.

49. Gail, M.H., Tan, W.Y., and Piantadosi, S. (1988). Tests for no treatment effect in randomized clinical trials. Biometrika 75, 57–64.

50. Robinson, L.D., and Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models. Int. Stat. Rev./Rev. Int. Stat. 59, 227–240.

51. Begg, M.D., and Lagakos, S. (1993). Loss in efficiency caused by omitting covariates and misspecifying exposure in logistic regression models. J. Am. Stat. Assoc. 88, 166–170.

52. Neuhaus, J.M. (1998). Estimation efficiency with omitted covariates in generalized linear models. J. Am. Stat. Assoc 93, 1124–1129.

53. Aschard, H., Vilhjálmsson, B.J., Joshi, A.D., Price, A.L., and Kraft, P. (2015). Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. Am. J. Hum. Genet. 96, 329–339.