



Covariate selection for association screening in multiphenotype genetic studies

Hugues Aschard¹⁻³ , Vincent Guillemot¹, Bjarni Vilhjalmsson⁴, Chirag J Patel⁵ , David Skurnik⁶⁻⁹, Chun J Ye¹⁰, Brian Wolpin¹¹, Peter Kraft^{2,3,12,14} & Noah Zaitlen^{13,14}

Testing for associations in big data faces the problem of multiple comparisons, wherein true signals are difficult to detect on the background of all associations queried. This difficulty is particularly salient in human genetic association studies, in which phenotypic variation is often driven by numerous variants of small effect. The current strategy to improve power to identify these weak associations consists of applying standard marginal statistical approaches and increasing study sample sizes. Although successful, this approach does not leverage the environmental and genetic factors shared among the multiple phenotypes collected in contemporary cohorts. Here we developed covariates for multiphenotype studies (CMS), an approach that improves power when correlated phenotypes are measured on the same samples. Our analyses of real and simulated data provide direct evidence that correlated phenotypes can be used to achieve increases in power to levels often surpassing the power gained by a twofold increase in sample size.

Performing agnostic searches for associations between pairs of variables in large-scale data, using either common statistical techniques or machine-learning algorithms, faces the problem of multiple comparisons. This problem is particularly present in genetic association studies, in which contemporary cohorts have access to millions of genetic variants as well as a broad range of clinical factors and biomarkers for each individual. With billions of candidate associations, identifying a true association of small magnitude is extremely challenging. Standard analysis approaches currently consist of examining the data in one dimension (i.e., testing a single outcome with each of the millions of candidate genetic predictors) and applying univariate statistical tests—the commonly named genome-wide association study (GWAS) approach^{1,2}. To increase power, GWAS relies on increasing the sample size to reach the multiple-comparisons-adjusted significance level. The largest studies to date, including hundreds of

thousands of individuals across dozens of cohorts, have decreased the limit of detectable effect sizes. For example, researchers have reported genetic variants explaining less than 0.01% of the total variation in body mass index³.

In addition to the substantial financial costs of collecting and genotyping large cohorts, this brute-force approach has practical limits. More importantly, this approach does not leverage the large amount of additional phenotypic and genomic information measured in many studies. Joint analyses of multiple phenotypes with each predictor of interest (for example, multivariate analysis of variance (MANOVA) and MultiPhen)⁴⁻⁶ offer a gain in power but have three major drawbacks. First, a significant result can be interpreted only as an association with any one of the phenotypes. Although this information is useful for screening purposes, it is insufficient to identify specific genotype–phenotype associations⁶. Second, such analyses make the replication process difficult, because association signals in the discovery sample depend on many parameters including the phenotypic correlation and the effect of the genotype on each phenotype. Third, joint tests have lower power than do univariate tests when only a small proportion of the phenotypes are associated with the tested genetic variant. This lower power is a simple problem of dilution: a small number of true associations mixed with many null phenotypes decreases the power.

In this work, we developed covariates for multiphenotype studies (CMS), a method that improves association-test power in multiphenotype studies while providing the resolution of univariate tests. When testing for association between a genotype and a phenotype, CMS allows the other collected correlated phenotypes to serve as covariates. The core of the method is a principled approach to selecting a set of these covariates that are correlated with the phenotype but not with the genotype, thereby decreasing phenotypic variance independently of the genotype and concomitantly increasing power. Via application of CMS to simulated and real data, we found that CMS scales to thousands of phenotypes, produces gains in power equivalent to that

¹Centre de Bioinformatique, Biostatistique et Biologie Intégrative (C3BI), Institut Pasteur, Paris, France. ²Department of Epidemiology, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA. ³Program in Genetic Epidemiology and Statistical Genetics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA. ⁴Bioinformatics Research Centre, Aarhus University, Aarhus, Denmark. ⁵Department of Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA. ⁶Division of Infectious Diseases, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA. ⁷Massachusetts Technology and Analytics, Brookline, Massachusetts, USA. ⁸Department of Microbiology, Necker Hospital, University Paris-Descartes, Paris, France. ⁹Institut Necker–Enfants Malades, INSERM U1151–Equipe 11, Paris, France. ¹⁰Department of Epidemiology and Biostatistics, Institute of Human Genetics, San Francisco, California, USA. ¹¹Center for Gastrointestinal Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, Massachusetts, USA. ¹²Department of Biostatistics, Harvard TH Chan School of Public Health, Boston, Massachusetts, USA. ¹³Department of Medicine, University of California, San Francisco, San Francisco, California, USA. ¹⁴These authors jointly directed this work. Correspondence should be addressed to H.A. (hugues.aschard@pasteur.fr).

Received 14 July; accepted 21 September; published online 16 October 2017; doi:10.1038/ng.3975

resulting from a two- to threefold increase in sample size, and outperforms other recently proposed multiphenotype approaches with univariate resolution, including a Bayesian approach (multivariate Bayesian imputation-based association mapping (mvBIMBAM⁷)) and dimensionality-reduction approaches (principal component analysis⁸ and probabilistic estimation of expression residuals (PEER⁹)).

RESULTS

Covariates as a proxy for unmeasured causal factors

The objective of this work was to develop a method that keeps the resolution of univariate analysis in testing for association between outcome Y and candidate predictor X , but takes advantage of other available covariates $C = (C_1, C_2, \dots, C_m)$ to increase power. Consider the inclusion of covariates correlated with the outcome in a standard regression framework. This inclusion may increase the signal-to-noise ratio between the outcome and the candidate predictor when testing $Y = X + C_L$, where $C_L \subset C$. Selection of which covariates C_i are relevant to a specific association test is usually based on causal assumptions^{10,11}. Epidemiologists and statisticians commonly recommend inclusion of two types of covariates in testing for association between X and Y : (i) those that are potential causal factors of the outcome and independent of X and (ii) those that may confound the association signal between X and Y , i.e., variables such as principal components (PCs) of genotypes or covariates that capture undesired structures in the data that can lead to false associations¹². All other variables that vary with the outcome because of shared risk factors are usually ignored. However, those variables carry information about the outcome and more precisely about the risk factors of the outcome. Because they potentially share dependencies with the outcome, they can be used as proxies for unmeasured risk factors. As such, they can be incorporated in C_L to improve the detection of associations between X and Y . However, when these variables depend on the predictor X , using them as covariates can lead to both false-positive and false-negative results depending on the underlying causal structure of the data.

The presence of interdependent explanatory variables, also known as multicollinearity¹³, can induce bias in the estimation of the predictor's effect on the outcome. We have recently discussed this issue in the context of GWAS adjusting for heritable covariates¹⁴. To illustrate this collider bias, consider first the simple case of two independent covariates U_1 and U_2 that are true risk factors of Y . In testing for association between X and Y , adjusting for U_1 and U_2 can increase power, because the residual variance of Y after the adjustment is smaller while the effect of X is unchanged (Fig. 1a), i.e., the ratio of the outcome variance explained by X over the residual variance is larger after removal of the effects of U_1 and U_2 . However, in practice, true risk factors of the outcome are rarely known. Consider instead the more realistic scenario in which U_1 and U_2 are unknown, but a covariate C , which also depends on those risk factors, has been measured. Because of their shared etiology, Y and C display a positive correlation, and when X is not associated with C , adjusting Y for C increases the power to detect (Y, X) associations (Fig. 1b). Problems arise when C is associated with X . In that case, adjusting Y for C biases the estimation of the effect of X on Y , thereby decreasing the power when the effect of X is concordant between C and Y (Fig. 1c), and inducing a false signal when X is not associated with Y (Fig. 1d). The same principles apply when multiple covariates correlated with the outcome are included.

When none of the covariates depend on the predictor (Fig. 1a,b), their inclusion in a regression can decrease the variance of the outcome without confounding, thus the increasing statistical power while maintaining the correct null distribution. This gain in power can be

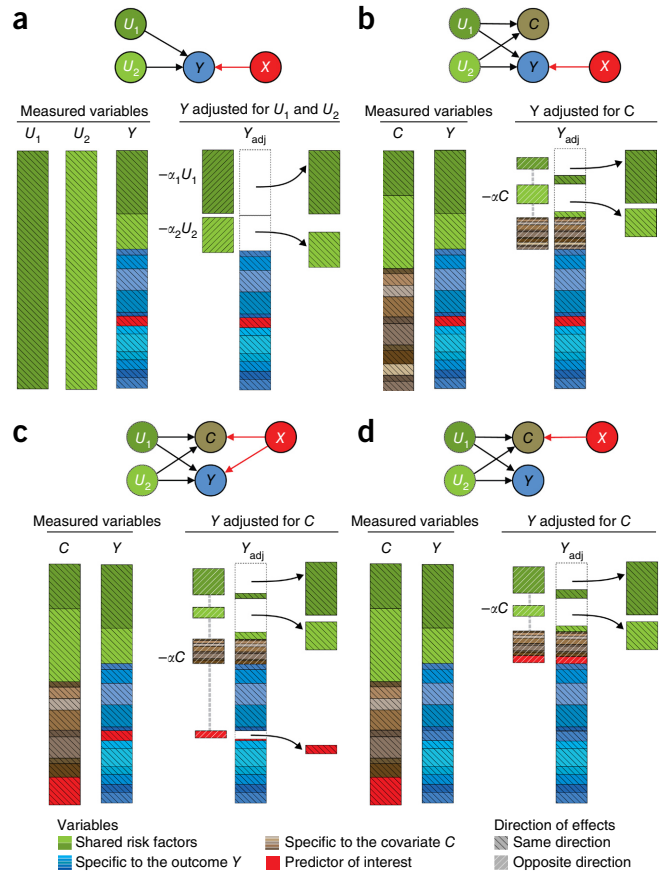


Figure 1 Variance components of adjusted variables. (a–d) Illustrations of the components of the variance of outcome Y before and after adjusting for other variables. The predictor of interest X is displayed in red. In **a**, the adjusting variables (U_1 and U_2) are true causal factors that have direct effects on Y ; therefore, adjusting Y for U_1 and U_2 (thus yielding Y_{adj}) decreases the variance of Y . In **b**, the true factors are not measured, but a variable C influenced by U_1 and U_2 , is measured. Adjusting Y for C decreases the residual variance of Y but also introduces a component of the variance specific to C . In **c**, the covariate shares factors with Y but is also influenced by X . When the effect of X on C is concordant with the effect of X on Y , a power loss may be induced. In **d**, Y is not associated with the predictor, and adjusting for C can induce a false-association signal by introducing the effect of X into the residual of Y .

easily described in terms of an equivalent sample-size increase. The noncentrality parameter (ncp) of the standard univariate chi-square test between X and Y is $ncp_{XY} = n \times \pi_X^2 / (\sigma_Y^2 - \pi_X^2)$, where n , σ_Y^2 and π_X^2 are the sample size, the total variance of the outcome Y , and the squared correlation between X and Y , respectively. When reducing σ_Y^2 by a factor γ through covariate adjustment, and assuming that the effect of X on Y is small, so that $\sigma_Y^2 - \pi_X^2 \approx \sigma_Y^2$, ncp_{XY} can be approximated by $n \times \pi_X^2 / (\sigma_Y^2 \times \gamma) = (n/\gamma) \times (\pi_X^2 / \sigma_Y^2)$. For example, when the covariates explain 30% of the variance of Y , the power of the adjusted test is equivalent to that when a sample size ~ 1.4 -fold larger (as compared with the unadjusted test) is analyzed. When covariates explain 80% of the phenotypic variance—a realistic proportion in some genetic data sets examined below—the power gain is equivalent to that resulting from a fivefold increase in sample size (Fig. 2a).

Selecting covariates for each outcome–predictor pair

The central problem that must be solved is how to select a subset of the available covariates to optimize power while preventing induction

of false-positive associations between the outcome and the predictor. To perform this selection, all covariates associated with the outcome should be included except those also associated with the predictor. A naive solution would consist of filtering out covariates on the basis of a P -value threshold from the association test between each covariate and the predictor (for example, removing predictors with a predictor–covariate association $P < 0.05$). However, unless the sample size were to be infinitely large, type I covariates (covariates associated with the predictor) would be included. Furthermore, such a filtering would also imply that some type II covariates (covariates not associated with the predictor) would be removed because they would incidentally pass the P -value threshold. Interestingly, removing type II covariates by using this approach not only results in a suboptimal test but also induces an inflated false-positive rate (Supplementary Fig. 1). In brief, when the outcome and the covariate are correlated, a low predictor–covariate P value implies a low predictor–outcome P value. As a result, the P -value distribution from the subset of predictor–outcome-unadjusted statistics (those for which the predictor–covariate P value is below the threshold) is enriched for low P value, while the complementary subset of predictor–outcome-adjusted statistics is expected to be uniform, thus resulting in an overall inflation of type I error for the approach (Supplementary Note and Supplementary Fig. 2).

In this work, we developed CMS, a computationally efficient heuristic to improve the selection of type II covariates while removing type I covariates. We present an overview of the approach, and complete details of the algorithm are provided in the Online Methods and the Supplementary Note.

Let $\hat{\delta}$ and $\hat{\beta}$ be the marginal estimated regression coefficients between X and C , and between X and Y (not adjusted for C), respectively, and let $\hat{\gamma}$ be the estimated correlation between Y and C . Naive P -value-based filtering, i.e., unconditional filtering on $\hat{\delta}$, assumes that under the null ($\delta = 0$), $\hat{\delta}$ is normally distributed with $\mathbb{E}(\hat{\delta}) = 0$ and variance $1/n$, where n is the sample size. The central advance of CMS is to additionally use the expected mean and variance of $\hat{\delta}$ conditional on $\hat{\beta}$ under a complete null model ($\delta = \beta = 0$). We show that these can be approximated as: $\mathbb{E}(\hat{\delta}|\hat{\beta}) \approx \hat{\beta}\hat{\gamma}$ and $\text{var}(\hat{\delta}|\hat{\beta}) \approx \text{var}(\hat{\delta} - \hat{\beta}\hat{\gamma}) = (1 - \hat{\gamma}^2)/n$ (Supplementary Note and Supplementary Fig. 3).

The bias observed from naive univariate P -value filtering (Supplementary Fig. 1) is induced by the misspecification of the expected mean and variance of the estimate of the predictor–covariate effect when the predictor is associated with neither the outcome nor the covariates. The $\hat{\delta}$ inclusion area for a P -value threshold of 5%—i.e., if $\hat{\delta}$ is outside the inclusion area, the covariate C is filtered out—based on the unconditional distribution is illustrated in Figure 3a. Using the distribution of $\hat{\delta}$ conditional on $\hat{\beta}$ to select covariates is also a poor solution resulting in a deflated test statistic for $\hat{\beta}$, owing to an overestimation of the standard error of $\hat{\beta}$ when adjusting for the selected covariates (Supplementary Table 1 and Supplementary Figure 4, which describe the simple case of a single covariate). The improvement from CMS is derived from defining the inclusion area as a combination of the unconditional and conditional distributions of $\hat{\delta}$ (Fig. 3b,c). This procedure solves the inflation observed in Supplementary Figure 1 and leads to a valid test under the complete null model with a variable number of available covariates (Supplementary Fig. 3 and Supplementary Table 1).

Finally, to decrease the risk of false positives, the algorithm scales inclusion areas on the basis of the total amount of the outcome’s variance explained by $C_{l \in L}$ and $\hat{\beta}$. To further improve the performance of filtering covariates, we also considered the omnibus association test between $C_{l \in L}$ and Y , which can be more effective when multiple covariates have small to moderate effects (Supplementary Note).

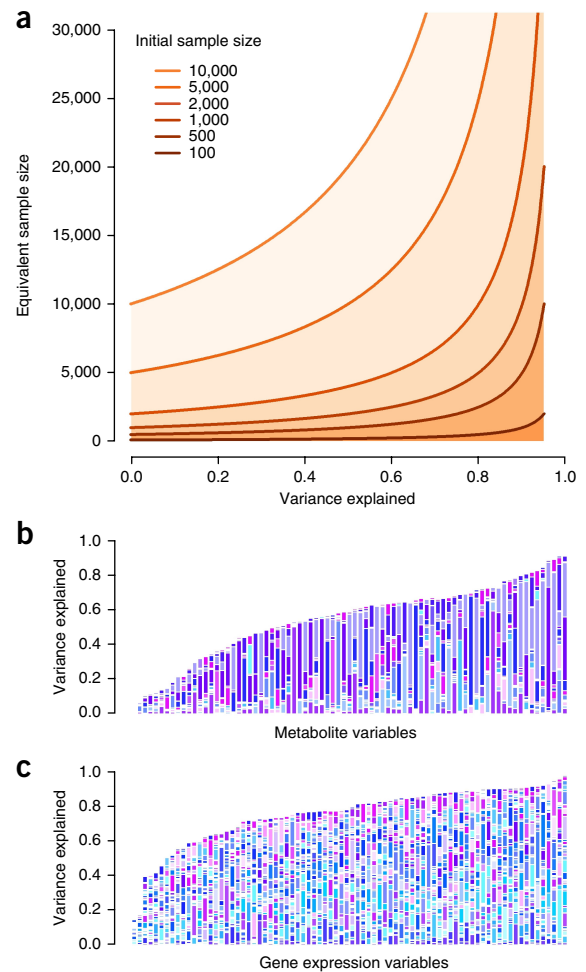


Figure 2 Examples of shared variance in real data and equivalent increases in sample size. (a) Equivalent increase in sample size as a function of the variance of the outcome explained by covariates, assuming initial sample sizes ranging from 100 to 10,000. (b,c) Distribution of variance explained by other variables for 79 metabolites from the PanScan study (b) and a random subsample of expression abundance estimates from 79 genes in the gEUVADIS study (c). The size of the bar corresponds to the total variance of each outcome explained by other available covariates, and the relative contributions of these covariates to each outcome are illustrated with different sets of random colors for each bar.

Simulated data analysis and method comparisons

We first assessed the performance of the proposed method through a simulation study in which we generated series of multiphenotype data sets over an extensive range of parameter settings (Online Methods and Supplementary Note). Each data set included n individuals genotyped at a SNP with the minor allele frequency (MAF) drawn uniformly from $[0.05, 0.5]$, a normally distributed phenotype Y , and $m = [10, 40, 80]$ correlated covariates $C = (C_1, C_2, \dots, C_m)$. Under the null, the SNP did not contribute to the phenotype, and under the alternate, the SNP contributed to the phenotype under an additive model. In some data sets, the SNP also contributed to a fraction $\pi = [0\%, 15\%, 35\%]$ of the covariates. These were the covariates that we sought to identify and filter out of the regression. We considered sample sizes (n) of 300, 2,000, and 6,000, and we varied r_C^2 , the variance of Y explained by C , from 25% to 75%. We varied the effect of the predictor on Y and C , when relevant, from almost undetectable (median $\chi^2 = 3$) to relatively large (median $\chi^2 = 20$). For each choice

of parameters, we generated 10,000 replicates and performed four association tests: (unadjusted) linear regression (LR), LR with covariates included based on P -value filtering at an α threshold of 0.1 (FT), CMS, and an oracle method including only the covariates not associated with the SNP (OPT), which was the optimal test regarding our goal. We considered a total of 432 scenarios, and the type I error rate of CMS was well calibrated across parameter ranges (Fig. 4 and Supplementary Tables 2–4). Notably, we did not consider strategies including all $C_{l=1\dots m}$ variables as covariates, or ‘reverse regression’ (MultiPhen)⁵, because these approaches substantially inflate the type I error rate (Supplementary Fig. 5).

We compared the performance of CMS with those of other recently proposed multiphenotype approaches including mvBIMBAM. The CMS approach was more than 100-fold faster than mvBIMBAM, and the two methods showed similar accuracy when they were compared with receiver-operating-characteristic curves (Supplementary Fig. 6). We also considered data-reduction techniques aimed at modeling hidden structure. For each data set, we tested the association between the primary outcome and the genotype while adding PCs or PEER factors. We observed increasing type I error rates when increasing the number of PCs or PEER factors in the model (Supplementary Fig. 7). Furthermore, at a fixed false-positive rate, when we applied CMS in addition to PEER factors, we found that CMS substantially increased the power above that gained from PEER (Supplementary Fig. 8 and Supplementary Note).

Real-data analysis

We first analyzed a set of 79 metabolites measured in 1,192 individuals genotyped at 668 candidate SNPs. We derived the correlation structure between these metabolites³ (Fig. 2b and Supplementary Fig. 9) and estimated the maximum gain in power that could be achieved by our approach in these data. The proportion of variance of each metabolite explained by the other metabolites varied between 1% and 91% (Fig. 2b). This proportion was higher than 50% for two-thirds of the metabolites and was equivalent to that resulting from a twofold increase in sample size. For 10% of the metabolites, other variables explained more than 80% of the variance and corresponded to a fivefold increase in sample size. In such cases, predictors explaining less than 1% of a metabolite’s variation can change from undetectable (power <1%) to fully detectable (power >80%) when CMS is applied.

We performed a systematic screening for the association between each SNP and each metabolite, using both a standard univariate linear regression adjusting for potential confounding factors and CMS to identify additional covariates. Overall, both tests showed correct P -value distribution ($\lambda_{GC} \sim 1$, (Supplementary Fig. 10a)). We focused on associations significant after Bonferroni correction ($P < 9.5 \times 10^{-7}$ corresponding to the 52,772 tests performed). The standard unadjusted approach (LR) detected five significant associations. In comparison, the CMS approach identified ten associations (Table 1), including four of the five associations identified by LR. In most cases, the P value of CMS was dramatically lower (1,000-fold smaller for rs780094 (alanine)). Comparing these results with those of four independent GWAS metabolite scans of larger sample size (study total $n = 8,330$ for Finnish¹⁵, 7,824 and 2,820 for Kooperative Gesundheitsforschung in der Region Augsburg (KORA) plus TwinsUK^{16,17}; and 2,076 for Framingham Heart Study (FHS)¹⁸ cohorts), we found that all metabolite–gene associations identified by only CMS replicated (Supplementary Table 5).

This analysis confirmed the power of CMS, highlighting its ability to identify variants with much smaller sample sizes than those

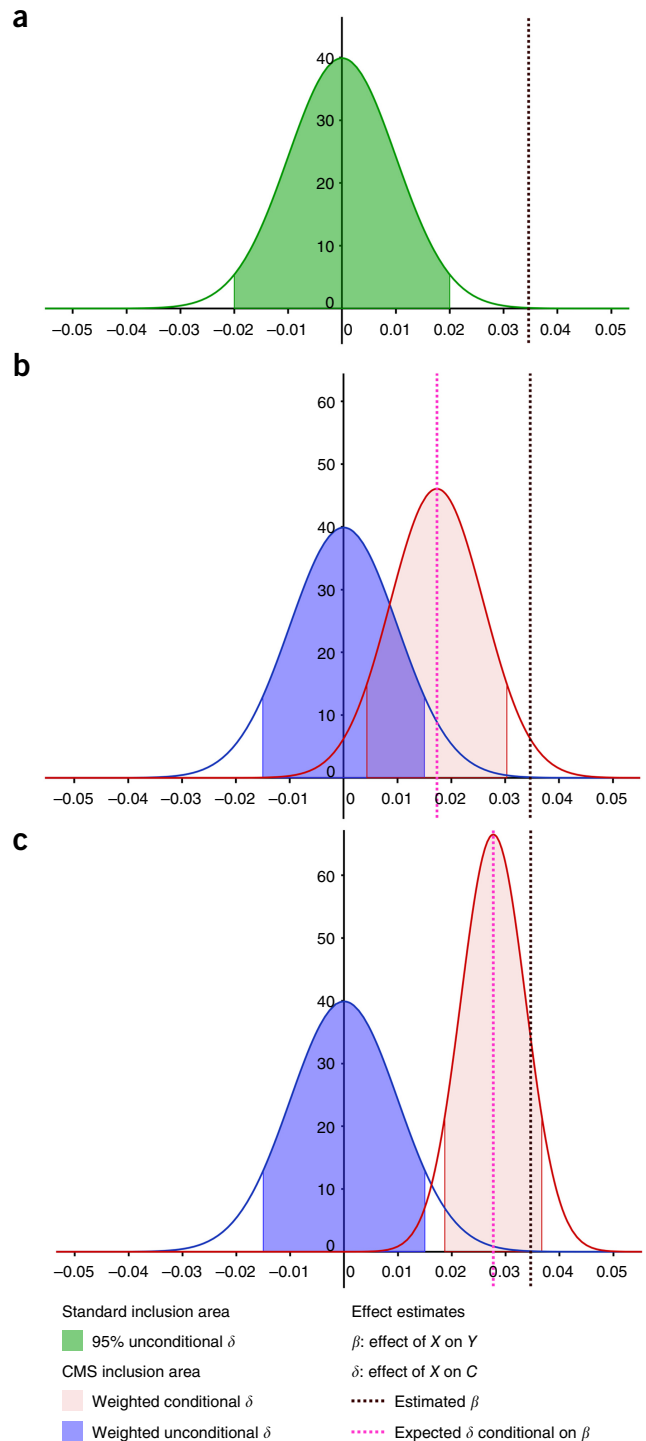


Figure 3 Conditional and unconditional distribution. Example of inclusion area based on the distribution of $\hat{\delta}$, the estimated effect between the predictor X and the covariate C under the null hypothesis of no association between X and C ($\delta = 0$) and no association between X and the outcome Y ($\beta = 0$). (a) Standard 95% confidence interval (green area) corresponding to $P < 0.05$ unconditional on $\hat{\beta}$. (b,c) Unconditional (blue curve) and conditional (pink curve) distribution of $\hat{\delta}$. CMS combines the two, setting an inclusion area (blue and pink shaded) while weighting both intervals by a factor depending on the correlation between Y and C , which equals 0.5 in b and 0.8 in c. Plots were drawn on the basis of the assumption that all variables are standardized, with a sample size of 10,000, an overall explained variance of Y of 0.7, $\hat{\beta} = 0.035$ and a multivariate test of association between all covariates and Y with a P value (P_{MUL}) of 0.3.

required in the standard unadjusted approach. Interestingly, the only association identified by the unadjusted analysis (lactose and GC, $P = 6.1 \times 10^{-7}$) and not confirmed by CMS ($P = 6.3 \times 10^{-6}$) was also the only one that did not replicate in the larger studies. Notably, in our analysis (Table 1), we followed an approach identical to that of the previous studies and did not adjust for either PCs or PEER factors⁹. However, adjusting did not qualitatively change the results. For example, we considered adjusting for 5, 10, and 20 PCs and obtained 11, 15, and 17 hits for CMS and 9, 11, and 5 hits for LR with PC covariates (Supplementary Table 6). The overall replication rate was lower when PCs were included, in agreement with a potential higher false-positive rate, as observed in our simulations.

We then considered genome-wide mapping of *cis*-expression quantitative trait loci (*cis*-eQTL) in RNA-seq data from the Genetic European Variation in Health and Disease (gEUVADIS) study. Gene expression is a particularly compelling benchmark, because the gold-standard analyses already use an adjustment strategy to account for hidden factors in eQTL GWAS^{9,19}. We used the PEER approach⁹ to derive hidden factors, because this method was applied in the original analysis²⁰. After stringent quality control, the data included 375 individuals of European ancestry with expression estimated on 13,484 genes, of which 11,675 had at least one SNP with a MAF $\geq 5\%$ within 50 kb of the start and end sites.

We observed that expression levels between genes were highly correlated (Fig. 2c), an ideal scenario for CMS. We first performed a standard *cis*-eQTL screening using LR, testing each SNP within 50 kb of each available gene for association with the overall normalized RNA level while adjusting for ten PEER factors, for a total of ~ 3.5 million tests. Then we applied CMS to identify, for each test, which other genes' RNA levels could be used as covariates in addition to the PEER factors. Both LR and CMS showed large numbers of highly significant associations (Supplementary Fig. 10b). For comparison purposes, we plotted the most significant SNP per gene obtained with the standard approach against those obtained with CMS (Fig. 5) and found that 2,725 genes had a least one SNP significant with both methods, whereas 56 genes were identified by only the standard approach. In contrast, 657 genes were found with only CMS, corresponding to a 22% increase in detection of *cis*-eQTL loci. This result indicated that by being gene/SNP specific, CMS is a priori able to recover substantial additional variance, thus allowing for increased power (Table 2 and Supplementary Table 7).

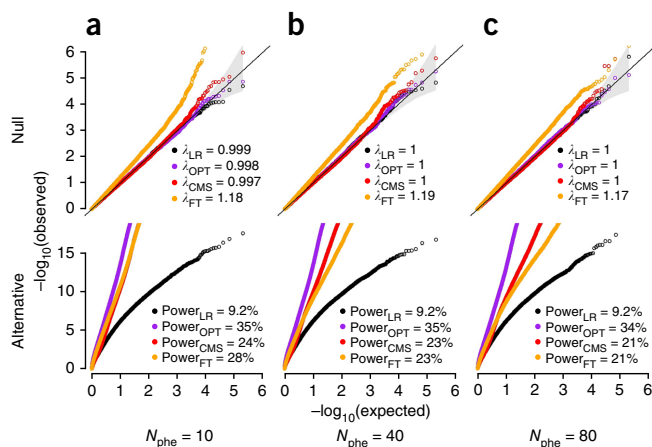


Figure 4 Power and robustness quantile-quantile plots under the null and alternate distributions of P values from a series of simulations. (a–c) Four statistical tests are compared: a standard marginal univariate test (LR); the optimally adjusted test (OPT), which includes as covariates only the outcomes not associated with the predictor; CMS; and a univariate test that includes as covariates all outcomes with a P value for association with the predictor above 0.1 (FT). Gray boxes show the genomic inflation factor λ_{GC} for the null models (top) and the estimated power at an α threshold of 5×10^{-7} (to correct for 100,000 tests) for the alternative model (bottom). Null models also include the 95% confidence interval of the $-\log_{10}(P$ values), displayed as a gray cone around the diagonal. Simulations were taken from 100,000 data sets including 10 (a), 40 (b), and 80 (c) outcomes (N_{phe}) under a null model (top), in which a predictor of interest is not associated with a primary outcome but is associated with 0%, 15%, or 35% of the other outcomes with probability 0.75, 0.2, or 0.05, respectively, and under the alternative (bottom), in which the predictor is associated with the primary outcome only. The variance of the primary outcome that could be explained by the other outcomes was randomly chosen from [25%, 50%, 75%] with equal probability.

To assess the validity of our results, we performed an *in silico* replication analysis, using two databases of known eQTLs^{21,22}. We found that 35% of the SNP-gene associations found by both LR and CMS replicated. For the subset of association found by only CMS, the replication rate was 20%, a value similar to the 22% from the LR-only replication. The replication rate was 6% for genes without a CMS or LR association. The replications were primarily in a lymphoblastoid cell line (LCL; Table 2), and the replication rate for our study

Table 1 Identified signals from the association test between 79 metabolites and 668 candidate SNPs

Chromosome	SNP	Gene	Outcome	P value			Known from study
				P_{LR}	P_{CMS}	SS_{incr}	
1	rs477992	<i>PHGDH</i>	Serine	6.2×10^{-5}	1.4×10^{-7}	2.15	KORA + TwinsUK ¹⁶ /FHS ¹⁸
2	rs2216405	Near <i>CPS1</i> , <i>LANCL1</i>	Glycine	4.1×10^{-26}	2.3×10^{-33}	1.56	KORA + TwinsUK ¹⁶ /FHS ¹⁸
			Serine	3.7×10^{-5}	6.4×10^{-10}	1.76	KORA + TwinsUK ¹⁶ /FHS ¹⁸
			Creatine	7.6×10^{-8}	4.8×10^{-9}	1.34	KORA + TwinsUK ¹⁶ /FHS ¹⁸
			Acetylglycine	2.2×10^{-8}	3.1×10^{-9}	1.44	KORA + TwinsUK ¹⁶
2	rs780094	<i>GCKR</i>	Alanine	6.1×10^{-5}	4.0×10^{-8}	2.06	KORA + TwinsUK ¹⁶ /FHS ¹⁸ /Finnish ¹⁵
4	rs1352844	<i>GC</i>	Lactose	6.1×10^{-7}	6.3×10^{-6}	2.06	
10	rs7094971	<i>SLC16A9</i>	Carnitine	2.9×10^{-10}	1.1×10^{-15}	2.01	KORA + TwinsUK ¹⁶ /FHS ¹⁸
			Acetylcarnitine	1.4×10^{-6}	9.4×10^{-13}	2.36	KORA + TwinsUK ¹⁶
12	rs2657879	<i>GLS2</i>	Glutamine	3.1×10^{-5}	4.2×10^{-10}	2.50	KORA + TwinsUK ¹⁶ /Finnish ¹⁵
16	rs6499165	<i>SLC7A6</i>	Lysine	2.6×10^{-5}	7.5×10^{-10}	3.00	KORA + TwinsUK ¹⁶

There were 79 metabolites tested for association with 668 SNPs, for a total of 52,104 tests. The P -value threshold accounting for multiple testing was 9.5×10^{-7} . Significant P values are indicated in bold. P_{LR} , P value for the standard unadjusted univariate test of each single phenotype with each single SNP; P_{CMS} , P value from the CMS algorithm; SS_{incr} , equivalent sample-size increase achieved after adjustment for covariates selected by the CMS algorithm. The sample sizes of the replication were 8,330, 7,824, and 2,076 for the Finnish¹⁵, KORA plus TwinsUK^{16,17}, and FHS¹⁸ studies, respectively.

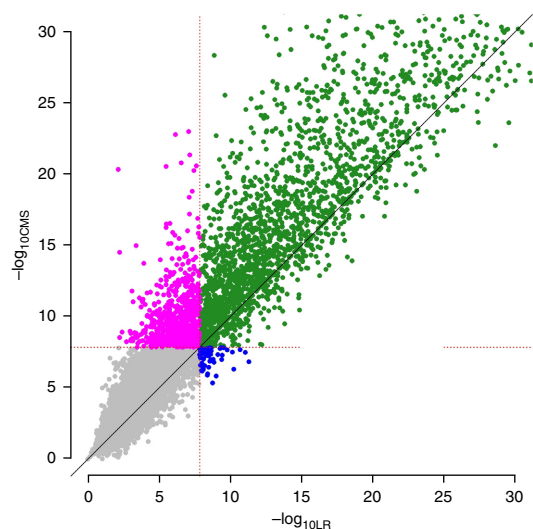


Figure 5 Analysis of the gEUVADIS data. Plot of $-\log_{10}(P$ values) of the most significant SNP per gene obtained by CMS (y axis) and LR (x axis) from genome-wide *cis*-eQTL mapping of 11,675 genes in 375 individuals from the gEUVADIS study. For illustration purposes, we truncated the plots at $-\log_{10}(P$ value) = 30. Both CMS and LR were adjusted for ten PEER factors, and the CMS analysis also included 0–50 additional covariates per SNP–gene pair tested. We considered a stringent significance threshold of 1.4×10^{-8} to account for the approximately 3.5 million tests and derived the number of genes showing at least one *cis*-eQTL with LR only (blue), CMS only (red), both approaches (turquoise), or neither approach (gray).

was within the same range as the replication rate in previous LCL studies (Supplementary Table 8), thus confirming that a substantial number of the additional associations identified by CMS probably corresponded to real signal (Online Methods). Additional GC correction of the *P* values by using inflation factors from a quasi-null experiment ($\lambda_{LR} = 1.01$, and $\lambda_{CMS} = 1.05$; Supplementary Fig. 11) did not qualitatively change the results.

DISCUSSION

Growing collections of high-dimensional data across myriad fields, driven in part by the ‘big-data revolution’ and the Precision Medicine Initiative, offer the potential to gain new insights and solve open problems. However, when mining for associations between collected variables, identifying signals within the noise remains challenging. Although univariate analysis offers precision, it fails to leverage the correlation structure between variables. In contrast, joint analyses

of multiple phenotypes increase power at the cost of decreased precision. Using both simulated and real data, we demonstrated that the proposed method, CMS, maintains the precision of univariate analysis but can still exploit global data structures to increase power. Indeed, in the data sets examined in this study, we observed up to a threefold increase in effective sample size in both the gene-expression and metabolite data as a result of the inclusion of relevant covariates (Supplementary Fig. 12).

CMS can be applied generally, but it is particularly well suited to the analysis of genetic data for several reasons. First, the genetic architectures of many complex phenotypes are consistent with a polygenic model with many genetic variants of small effect size that are difficult to detect with standard approaches²³. Second, many correlated phenotypes share genetic and environmental variance without complete genetic overlap²⁴. Third, the underlying structure of the genomic data is relatively well understood, and there is extensive literature describing the causal pathway from genotypes to phenotypes through direct and indirect effects on RNA, protein, and metabolites (Supplementary Fig. 13 and Supplementary Note). Finally, when the predictors of interest are genetic variants, there is less concern regarding potential confounding factors. The only well-established confounder of genetic data is population structure, and this confounding can be easily addressed through standard approaches¹². For other types of data, when the underlying structure of the data is unknown, the risk of introducing bias is high.

Several other groups have considered the problem of association testing in high-dimensional data while maintaining precision. In genetics, multivariate linear mixed models (mvLMMs) have demonstrated both precision and increases in power when correlated phenotypes are tested jointly. However, mvLMMs exploit only the genetic similarity of phenotypes and are not computationally efficient enough to handle dozens of phenotypes jointly⁴. CMS leverages both genetic and environmental correlations and can be easily adapted to hundreds or thousands of phenotypes, as demonstrated here. Instead, we compared CMS with other more related approaches, including the Bayesian approach mvBIMBAM, and adjustment for hidden factors inferred from either principal component analysis or PEER. We found that mvBIMBAM and CMS had very similar accuracy, as measured by the area under the curve, whereas mvBIMBAM was approximately 100-fold slower and was applicable to only a small number of phenotypes (fewer than ten). As for strategies that reconstruct hidden variables, we have found that they can induce false positives²⁵, and they are suboptimal in comparison to CMS. Indeed, the gEUVADIS analysis showed a 22% increase in the detection of eQTL when it was applied in addition to PEER-factor adjustment.

Table 2 Replication of association from the *cis*-eQTL screening in GEUVADIS

Approach	No. disc. ^a	SNP ^b	% rep. ^c	Replication per tissue									
				Fibr.	LCL	T cell	Brain	B cell	Mon.	Liv.	Adi.	Skin	Blood
LR and CMS	2,725	LR	34.7%	1	737	4	27	20	69	33	137	125	185
		CMS	35.9%	3	770	2	26	20	73	28	147	133	175
LR only	56	LR	21.8%	0	5	0	0	1	0	0	3	4	3
		CMS	24.1%	0	8	0	0	1	0	0	2	2	3
CMS only	657	LR	20.2%	1	79	0	1	6	7	3	14	7	35
		CMS	19.6%	1	78	0	1	5	6	2	16	11	35
None	8,237	LR	7.0%	1	185	1	2	9	38	10	61	53	245
		CMS	7.2%	1	199	2	4	8	46	7	81	62	258

^aNumber of SNP–gene associations with *P* values below the Bonferroni-corrected significance threshold. ^bSNP used for the replication analysis. ^cPercentage of SNP–gene association replicated, after removal of the discovery SNPs that could not be mapped. Fibr., fibroblast; mon., monocytes; liv., liver; adi., adipose. The per-tissue sum does not equal the number of hits times the percentage of replication, because a given association can be replicated in multiple tissues.

There are several caveats to our approach. First, the proposed heuristic is conservative by design to avoid false-association signals, and so all the available power gain is not achieved. Second, although all performed simulations showed strong robustness, this method remains a heuristic, like other methods^{9,19}. Ultimately, we recommend external replication to validate results and effect size, as is standard in genetic studies. Third, CMS is more computationally intensive than methods such as principal component analysis or PEER. Fourth, CMS assumes that the variables are measured and available on all samples. The current implementation includes a naive missing-data imputation, and simple-case-scenario simulations showed that this strategy has a minimal effect on the robustness of CMS (**Supplementary Fig. 14**). However increasingly advanced approaches have been developed²⁶. Fifth, although the principles that we leveraged are probably applicable to categorical and binary outcomes (logistic regression in ref. 27), our algorithm is currently applicable to only continuous outcomes. Sixth, for monogenic disorders, or phenotypes without intermediately measured endophenotypes, CMS is unlikely to result in power gains.

We focused on association screening and aimed at optimizing power and robustness. However, the selection of covariates performed by CMS might carry information about which covariates operate through specific SNPs. Future work will explore whether output from CMS can generate hypotheses on the underlying causal model. There are other additional improvements not specific to CMS that are worth exploring. In particular, when multiple phenotypes are considered as outcomes, then a multiple-testing-correction penalty must be selected to account for all tests across all phenotypes. In this study, we applied a Bonferroni correction, not accounting for the correlation between outcomes; this is a conservative correction, and more powerful approaches are possible²⁸.

Large-scale genomic data have the potential to answer important biological questions and improve public health. However, those data come with methodological challenges. Many questions, such as improving risk prediction or inferring causal relationships rely on the ability to identify associations between variables. In this study, we provide a comprehensive overview of how leveraging shared variance between variables can be used to fulfill this goal. Building on this principle, we developed the CMS algorithm, an innovative approach that can dramatically increase statistical power to detect weak associations.

METHODS

Methods, including statements of data availability and any associated accession codes and references, are available in the [online version of the paper](#).

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

H.A. and N.Z. were supported by NIH grant R03DE025665. H.A. was also supported by NIH grant R21HG007687, and N.Z. was also supported by NIH career development award K25HL121295 and NIH grant U01HG009080. C.J.P. was supported by NIH grant R00 ES023504.

AUTHOR CONTRIBUTIONS

H.A. conceived the approach and performed all real-data analyses. H.A., N.Z., B.V., C.J.P., D.S., and P.K. contributed substantially to improving the approach

and the study design. C.J.Y. contributed to the quality control and analysis of the gEUVADIS data. B.W. collected the metabolite data and contributed to quality control and analysis of the metabolite data. H.A. and N.Z. conceptualized and performed the simulation study. V.G. contributed to the simulation study. H.A. and N.Z. wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Reprints and permissions information is available online at <http://www.nature.com/reprints/index.html>. Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

- Stranger, B.E., Stahl, E.A. & Raj, T. Progress and promise of genome-wide association studies for human complex trait genetics. *Genetics* **187**, 367–383 (2011).
- Sham, P.C. & Purcell, S.M. Statistical power and significance testing in large-scale genetic studies. *Nat. Rev. Genet.* **15**, 335–346 (2014).
- Locke, A.E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518**, 197–206 (2015).
- Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nat. Methods* **11**, 407–409 (2014).
- O'Reilly, P.F. *et al.* MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS. *PLoS One* **7**, e34861 (2012).
- Aschard, H. *et al.* Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies. *Am. J. Hum. Genet.* **94**, 662–676 (2014).
- Stephens, M. A unified framework for association analysis with multiple related phenotypes. *PLoS One* **8**, e65245 (2013).
- Liang, L. *et al.* A cross-platform analysis of 14,177 expression quantitative trait loci derived from lymphoblastoid cell lines. *Genome Res.* **23**, 716–726 (2013).
- Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
- Greenland, S., Pearl, J. & Robins, J.M. Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48 (1999).
- Hernán, M.A., Hernández-Díaz, S., Werler, M.M. & Mitchell, A.A. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am. J. Epidemiol.* **155**, 176–184 (2002).
- Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
- Farrar, D.E. & Glauber, R.R. Multicollinearity in regression analysis: the problem revisited. *Rev. Econ. Stat.* **49**, 92–107 (1967).
- Aschard, H., Vilhjálmsson, B.J., Joshi, A.D., Price, A.L. & Kraft, P. Adjusting for heritable covariates can bias effect estimates in genome-wide association studies. *Am. J. Hum. Genet.* **96**, 329–339 (2015).
- Kettunen, J. *et al.* Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nat. Genet.* **44**, 269–276 (2012).
- Shin, S.Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
- Suhre, K. *et al.* Human metabolic individuality in biomedical and pharmaceutical research. *Nature* **477**, 54–60 (2011).
- Rhee, E.P. *et al.* A genome-wide association study of the human metabolome in a community-based cohort. *Cell Metab.* **18**, 130–143 (2013).
- Leek, J.T. & Storey, J.D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
- Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
- Yu, C.H., Pal, L.R. & Moul, J. Consensus genome-wide expression quantitative trait loci and their relationship with human complex trait disease. *OMICS* **20**, 400–414 (2016).
- Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
- Gibson, G. Rare and common variants: twenty arguments. *Nat. Rev. Genet.* **13**, 135–145 (2012).
- Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat. Genet.* **47**, 1236–1241 (2015).
- Dahl, A., Guillemot, V., Mefford, J., Aschard, H. & Zaitlen, N. Adjusting for principal components of molecular phenotypes induces replicating false positives. Preprint at <https://www.biorxiv.org/content/early/2017/03/26/120899> (2017).
- Dahl, A. *et al.* A multiple-phenotype imputation method for genetic studies. *Nat. Genet.* **48**, 466–472 (2016).
- Robinson, L.D. & Jewell, N.P. Some surprising results about covariate adjustment in logistic regression models. *Int. Stat. Rev.* **59**, 227–240 (1991).
- Peterson, C.B., Bogomolov, M., Benjamini, Y. & Sabatti, C. Many phenotypes without many false discoveries: error controlling strategies for multitrait association studies. *Genet. Epidemiol.* **40**, 45–56 (2016).

ONLINE METHODS

The CMS algorithm. We developed an algorithm to select relevant covariates when testing for association between a predictor X and an outcome Y . For a set of candidate covariates $C = (C_1, C_2, \dots, C_m)$, the filtering is applied on $\hat{\delta}_l$ and P_b , the estimated marginal effect of the predictor X on C_l and its associated P value, respectively. It uses four major features: (i) r_C^2 , the total amount of variance of Y explained by the C ; (ii) $(\hat{\gamma}_{lu}^2, \hat{\gamma}_l^2)$, the estimated effect of each $C_{l \in 1..m}$ on Y from univariate and joint models, respectively; (iii) $\hat{\beta}$, the estimated effect of X on Y from the marginal model $Y \sim \alpha + \beta X$; and (iv) P_{MUL} , the P value for the multivariate test of all $C_{l=1..m}$ and X , which is estimated with a standard multivariate approach (MANOVA).

Filtering is applied in two steps, using the aforementioned features and additional parameters described thereafter. Step 1 is an iterative procedure focusing on P_{MUL} . It consists of removing potential covariates until P_{MUL} reaches t_{MUL} , a P -value threshold set to 0.05 by default. This step is effective at removing combinations of covariates with strong to moderate effects but may potentially leave weakly associated covariates.

Step 2 is also iterative and uses covariates preselected at step 1. It consists of deriving two confidence intervals, $\Delta_{l,cond}$ and $\Delta_{l,un}$, for the expected distribution of $\hat{\delta}_l$ conditional on $\hat{\beta}$ under a complete null model ($\delta_l = 0$ and $\beta = 0$), and the unconditional distribution of $\hat{\delta}_l$, respectively. The unconditional distribution of $\hat{\delta}_l$ can be approximated as $\mathcal{N}(0, \sqrt{1/n})$, and the conditional distribution is

$$\mathcal{N}(\hat{\gamma}\hat{\beta}, \sqrt{(1-\hat{\gamma}^2)/n}),$$

where $\hat{\gamma}$ is the estimated correlation between Y and C (Supplementary Note). The inclusion area for each $\hat{\delta}_l$ is defined as the union of $\Delta_{l,cond}$ and $\Delta_{l,un}$, which are determined from the conditional and unconditional distributions, $r_C^2, (\hat{\gamma}_{lu}^2, \hat{\gamma}_l^2), \hat{\beta}$, and distribution-specific weights w_u and w_c , which we further introduced to improve power and robustness. Specifically,

$$\Delta_{l,un} = [\mu_{l,un} - \sigma_{l,un} \times w_u, \mu_{l,un} + \sigma_{l,un} \times w_u]$$

and

$$\Delta_{l,cond} = [\mu_{l,cond} - \sigma_{l,cond} \times w_c, \mu_{l,cond} + \sigma_{l,cond} \times w_c],$$

where

$$(\mu_{l,un}, \mu_{l,cond}) \text{ and } (\sigma_{l,un}, \sigma_{l,cond})$$

are the unconditional and conditional means and s.d., respectively.

The weights w_u and w_c are always less than two and shrink the size of the inclusion area. To obtain (w_u, w_c) , we first set an *ad hoc* stringency parameter

$$w_{ST} = 0.1 \times P_{MUL} \times (1 - r_C^2) \times (1 - \hat{\gamma}_{lu}^2) / \hat{\gamma}_l^2,$$

which decreases as $r_C^2, \hat{\gamma}_l$, and the $\hat{\gamma}_{lu}$ increase, thus making the inclusion area smaller, because the covariate C_l being considered explains more of the variance of Y . The purpose of this parameter is to decrease the risk of false positives, because bias is enhanced when the residual variance of the outcome is decreased¹⁴. This phenomenon is illustrated in Figure 3, in which the unconditioned inclusion area from CMS is smaller than that for the standard approach.

As $|\hat{\beta}|$ increases, the likelihood of the true β being null decreases, and we want w_c and the conditional interval $\Delta_{l,cond}$ to shrink to zero. We use a simple linear function for w_c with a transition that corresponds to the point where the 95% CI of the observed $\hat{\beta}$ and $\hat{\delta}_l | \delta_l = 0$ stop overlapping. When all variables are standardized, the former CI is approximately equal to

$$\hat{\beta} \pm 2/\sqrt{n_X}$$

whereas the latter equals

$$0 \pm 2/\sqrt{n}$$

Thus, the proposed transition point corresponds to $\hat{\beta} = 4/\sqrt{n}$. Expressed as chi squared, it equals:

$$\chi_{\beta}^2 = \hat{\beta}^2 \times n \sigma = 16$$

We set

$$w_c = \min(w_{ST}, f_c(\chi_{\beta}^2))$$

and

$$w_u = \min(w_{ST}, f_u(\chi_{\beta}^2))$$

where $f_c(\chi_{\beta}^2)$ and $f_u(\chi_{\beta}^2)$ vary between 0 and 2, and are defined to linearly scale with respect to this transition point (Supplementary Note).

Altering the transition point or scaling the inclusion interval can increase the risk of false positives or decrease power (Supplementary Figs. 15–17). We chose the CMS parameters conservatively to prevent false positives; however, alternative approaches such as cross-validation may identify parameters that increase the power of CMS while maintaining a calibrated null distribution. Interestingly, the omnibus association test between $C_{l \in L}$ and Y had very little effect on the overall performance (Supplementary Fig. 17) with the parameters used here.

Finally, because of multicollinearity, the estimated γ_l can vary substantially depending on which other covariates $C_{k \neq l}$ are already included in the model. As a result, γ_l cannot be estimated from a marginal model such as $Y \sim \gamma_l C_l$. To address this issue, we implemented the selection of covariates into an iterative loop in which $\hat{\gamma} = (\hat{\gamma}_1 \dots \hat{\gamma}_m)$ terms are reestimated from a joint model each time a candidate covariate is excluded. The complete CMS algorithm is provided in the Supplementary Note.

Simulations. We simulated series of genetic and phenotypic data sets under a variety of genetic models to interrogate the properties of the proposed test. Each data set included n individuals genotyped at a SNP, a normally distributed phenotype Y , and $m = [10, 40, 80]$ correlated covariates $C = (C_1, C_2, \dots, C_m)$. Genotypes g for each of the individuals were generated by summing two samples from a binomial distribution with probability uniformly drawn in $[0.05, 0.5]$ and then normalized to have mean 0 and variance 1. Under the null, the SNP does not contribute to the phenotype, and under the alternate, the SNP contributes to the phenotype under an additive model. In some data sets, the SNP also contributes to a fraction $\pi = [0\%, 15\%, 35\%]$ of the covariates. Those were the covariates that we sought to identify and filter out of the regression. The remaining variance for each phenotype, which represents the remaining genetic and environmental variance, was drawn from a $m+1$ -dimensional multivariate normal distribution with mean 0 and variance σ_C . In instances in which this matrix was not positive definite, we used the Higham algorithm²⁹ to find the closest positive definite matrix. The diagonal of the covariance matrix was specified as 1 minus the effect of g (if relevant) such that the total variance of each phenotype had an expected value of 1.

We considered sample sizes (n) of 300, 2,000, and 6,000, and we varied r_C^2 , the variance of Y explained by C , from 25% to 75%. We varied the effect of the predictor on Y and C , when relevant, from almost undetectable (median $\chi^2 = 3$) to relatively large (median $\chi^2 = 20$). For each choice of parameters, we generated 10,000 replicates and performed four association tests: (unadjusted) LR, LR with covariates included on the basis of P -value filtering at an α threshold of 0.1 (FT), CMS, and an oracle method including only the covariates not associated with the SNP (OPT), the optimal test regarding our goal. For each null model, we derived the genomic inflation factor³⁰ λ_{GC} , whereas for the alternative model, we estimated power at an α threshold of 5×10^{-7} to account for the 100,000 tests performed. All tests were two sided. Results for each of the 432 scenarios considered are presented in Supplementary Figures 18–44.

To comprehensively summarize the performance of the different tests across these scenarios, we randomly sampled subsets of the simulations to mimic real-data sets while focusing on a sample size of 2,000 individuals and a total of 100,000 SNPs tested. For null models, we assumed that two-thirds (66%) of the genotypes were under the complete null (not associated with any covariate, $\pi = 0$), whereas 27% were associated with a small proportion of the covariates ($\pi = 0.15$), and the remaining 7% were highly pleiotropic ($\pi = 0.35$).

We compared the performances of CMS against those of other recently proposed multiphenotype approaches, including mvBIMBAM, a Bayesian approach to classifying the outcome as directly associated, indirectly associated, or unassociated with the predictor. The main advantage of the mvBIMBAM approach is that it proposes a formal theoretical framework that, similarly to structural equation modeling, explores a wide range of underlying causal models. However, there is a large computational cost, and the approach is currently limited to the analysis of a relatively small number of traits (fewer than ten). We therefore performed our comparison by using small-scale simulated data (ten phenotypes).

Other potential alternatives to CMS are data-reduction techniques for modeling hidden structure. These methods have been widely used for the analysis of molecular phenotypic data, with a primary goal of removing confounding effects^{8,9,19}. We examined principal component analysis because it has been widely used and is still one of the most commonly used approaches⁸, and a more complex factor-analysis-inspired method (PEER), which has outperformed similar methods⁹. We simulated series of large multivariate data sets under a null model, in which a genotype is associated with multiple variables but not the primary outcome of interest (i.e., in the presence of type I covariates). For each data set, we tested the association between the primary outcome and the genotype while adding PCs or PEER factors (**Supplementary Fig. 7**) and found an increasing type I error rates after increasing the number of PCs or PEER factors in the model.

Previous studies have also shown that including fixed effects can improve power over dimensionality-reduction approaches that incorporate these same variables³¹, probably as a result of the shrink that is applied when these methods jointly fit effect sizes of multiple correlated variables. To investigate the power gains available to CMS when PCs/PEER factors are used, and assuming that type I error is controlled, we simulated data under an alternative model of true association but in the absence of type II covariates to avoid the aforementioned issue. We applied CMS in addition to a variable number of PEER factors and found that CMS can substantially increase the power above that gained from PEER (**Supplementary Fig. 8**).

Metabolite data. Circulating metabolites were profiled by liquid chromatography–tandem mass spectrometry (LC–MS) in prediagnostic plasma from 453 prospectively identified pancreatic cancer cases and 898 controls. The subjects were drawn from four US cohort studies: the Nurses' Health Study (NHS), Health Professionals Follow-up Study (HPFS), Physicians' Health Study (PHS) and Women's Health Initiative (WHI). Two controls were matched to each case on the basis of year of birth, cohort, smoking status, fasting status at the time of blood collection, and month/year of blood collection. Metabolites were measured in the laboratory of C. Clish at the Broad Institute by using the methods described in Wang *et al.*³² and Townsend *et al.*³³. A total of 133 known metabolites were measured; 50 were excluded from analysis because of poor reproducibility in samples with delayed processing ($n = 32$), CV >25% ($n = 13$), or undetectable levels for >10% subjects ($n = 5$). The remaining 83 metabolites showed good reproducibility in technical replicates or after delayed processing³³. Among those, 79 had no missing data and were considered further for analysis. Additional details of these data can be found in ref. 34. Genotypic data were also available for some of these participants. A subset of 645 individuals from NHS, HPFS, and PHS had genome-wide genotypes data as part of the PanScan study³⁵. Among the remaining participants, 547 have been genotyped for 668 SNPs chosen to tag genes in inflammation, vitamin D, and immunological pathways. To maximize sample size, we focused our analysis on these 668 SNPs, which were therefore available in a total of 1,192 individuals. The in-sample MAFs of these variants ranged from 1.1% to 50%. The metabolite levels were approximately Gaussian after adjustment for the confounding factors and were therefore not transformed further (**Supplementary Fig. 45**). We first applied standard linear regression testing of each SNP for association with each metabolite while adjusting for five potential confounding factors: pancreatic cancer case–control status, age at blood draw, fasting status, self-reported race, and sex. We then applied the CMS while also including the five confounding factors as covariates. All tests were two sided.

gEUVADIS data. The gEUVADIS data²⁰ consist of RNA-seq data for 464 LCL samples from five populations in the 1000 Genomes Project. Of these,

375 are of European ancestry (CEU, FIN, GBR, and TSI), and 89 are of African ancestry (YRI). In these analyses, we considered only the European-ancestry samples. Raw RNA-sequencing reads obtained from the European Nucleotide Archive were aligned to the transcriptome by using UCSC annotations matching hg19 coordinates. RNA-seq by expectation-maximization (RSEM)³⁶ was used to estimate the abundance of each annotated isoform, and total gene abundance was calculated as the sum of all isoform abundance values normalized to one million total counts or transcripts per million (TPM). For each population, TPM values were \log_2 transformed and median normalized to account for differences in sequencing depth in each sample. A total of 29,763 genes were initially available. We removed those that appeared to be duplicates or that had low expression (defined as $\log_2(\text{TPM}) < 2$ in all samples). After filtering, 13,484 genes remained. The genotype data were obtained from the 1000 Genomes Project Phase 1 data set. We restricted the analysis to the SNPs with a MAF $\geq 5\%$ that were within ± 50 kb from the gene tested for *cis* effects. A total of 11,175 genes had at least one SNP that matched those criteria. We performed standard *cis*-eQTL screening, first applying standard linear regression while adjusting for ten PEER factors. We then applied CMS while including the same PEER factors as covariates. All tests were two sided.

When running CMS, we performed prefiltering of the candidate covariates. More specifically, for each gene analyzed—referred to as the target gene—we restrained the number of candidate covariates (gene other than the target) to be evaluated. First, we aimed at avoiding genes whose expression was more likely to be associated with some of the SNPs tested because of a *cis* effect, because such genes were more likely to induce false signal. Thus, all genes in physical proximity to the target genes (≤ 1 Mb) were excluded. Second, we aimed at decreasing the number of candidate covariates (13,484 minus 1, a priori), because most of them were likely to be uninformative and because our simulation showed that for small sample size, CMS would have low robustness if the number of candidate covariates were too large. To do so, we performed an initial screening for association between the target and all other genes and used the top 50 showing the strongest squared correlation with the target.

We performed an *in silico* replication analysis by using two databases of known eQTLs. The first database included results from 15 publicly available studies (excluding the European gEUVADIS) from multiple tissues²¹, and a second one included eQTLs in whole-blood samples from a joint analysis of seven studies²². Summary statistics were not available for every SNP; instead, these databases listed all SNPs found at an FDR of 5% in each study. Therefore, we were not able to perform a standard replication study and instead compared the replication rate of CMS and LR in these databases. Notably, we expected a smaller replication rate for LR only and CMS only compared with those identified by both approaches, because the last group includes variants with the largest effects, whereas the first two correspond to associations of smaller magnitude. Finally, we performed a quasi-null experiment in which we tested for *trans* effects by using random SNPs from the genome, assuming that most of those would be under the null.

Variance explained in multiple regressions. We plotted the variance of a set of outcomes $Y = (Y_1, \dots, Y_K)$ that could be explained by covariates in the data, i.e., how much of the variance of Y_i could be explained by $Y_{j \neq i}$ (**Fig. 2b,c**). For illustration purposes, we also approximated the individual contribution of each $Y_{j \neq i}$ covariate. In brief, we standardized all variables and estimated γ_j^2 , the proportion of variance of the outcome explained by each $Y_{j \neq i}$ from the marginal models $Y_i \sim \gamma_j Y_{j \neq i}$, and r_{model}^2 , the total variance of Y_i explained by all $Y_{j \neq i}$ jointly, from the model

$$Y_i \sim \gamma Y_{j=1..K, j \neq i}$$

Then, we derived v_{ji} , an approximation of the relative contribution of each $Y_{j \neq i}$ to the variance of Y_i as follows:

$$v_{ji} = \frac{\gamma_j^2}{\sum_{k \neq i} \gamma_k^2} \times r_{\text{model}}^2$$

Notably, this is an arbitrary rescaling of the real contribution of the $Y_{j \neq i}$ variable. Indeed, the correlation between all $Y_{j \neq i}$ induces multicollinearity in the regression, and it follows that

$$\sum_{k \neq i} \gamma_k^2 \gg r_{\text{model}}^2.$$

Missing data. The current version of the algorithm includes a naive imputation strategy for missing data that consists of replacing missing values of candidate covariates with their mean values, thereby avoiding the sharp decrease in sample size that might arise if the proportion of missing values is too large. Notably, the inference was performed per predictor–outcome pair and only for the covariates, whereas we did not infer missing values for the outcome or the predictor tested. The imputation did not strongly affect the robustness of the test (**Supplementary Fig. 14**), although large-scale (i.e., $\geq 50\%$ of missing values) random missingness appeared to slightly deflate the test statistics from CMS.

Code availability. An implementation of the approach is freely available at <https://github.com/haschard/CMS/>.

Data availability. The gEUVADIS RNA-sequencing data, genotype data, variant annotations, splice scores, quantifications, and QTL results are freely

and openly available with no restrictions at <http://www.geuvadis.org/>. The metabolite data that support the findings of this study are available from the corresponding author upon reasonable request. A **Life Sciences Reporting Summary** for this paper is available.

29. Higham, N.J. Computing the nearest correlation matrix: a problem from finance. *IMA J. Numer. Anal.* **22**, 329–343 (2002).
30. Devlin, B., Roeder, K. & Wasserman, L. Genomic control, a new approach to genetic-based association studies. *Theor. Popul. Biol.* **60**, 155–166 (2001).
31. Liu, X., Huang, M., Fan, B., Buckler, E.S. & Zhang, Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. *PLoS Genet.* **12**, e1005767 (2016).
32. Wang, T.J. *et al.* Metabolite profiles and the risk of developing diabetes. *Nat. Med.* **17**, 448–453 (2011).
33. Townsend, M.K. *et al.* Reproducibility of metabolomic profiles among men and women in 2 large cohort studies. *Clin. Chem.* **59**, 1657–1667 (2013).
34. Mayers, J.R. *et al.* Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nat. Med.* **20**, 1193–1198 (2014).
35. Wolpin, B.M. *et al.* Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nat. Genet.* **46**, 994–1000 (2014).
36. Li, B. & Dewey, C.N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form is intended for publication with all accepted life science papers and provides structure for consistency and transparency in reporting. Every life science submission will use this form; some list items might not apply to an individual manuscript, but all fields must be completed for clarity.

For further information on the points included in this form, see [Reporting Life Sciences Research](#). For further information on Nature Research policies, including our [data availability policy](#), see [Authors & Referees](#) and the [Editorial Policy Checklist](#).

▶ Experimental design

1. Sample size

Describe how sample size was determined.

NA

2. Data exclusions

Describe any data exclusions.

NA

3. Replication

Describe whether the experimental findings were reliably reproduced.

We performed in-silico replications for the two real data analyses. Both replication analyses were successful.

4. Randomization

Describe how samples/organisms/participants were allocated into experimental groups.

NA

5. Blinding

Describe whether the investigators were blinded to group allocation during data collection and/or analysis.

NA

Note: all studies involving animals and/or human research participants must disclose whether blinding and randomization were used.

6. Statistical parameters

For all figures and tables that use statistical methods, confirm that the following items are present in relevant figure legends (or in the Methods section if additional space is needed).

n/a Confirmed

- The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement (animals, litters, cultures, etc.)
- A description of how samples were collected, noting whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
- A statement indicating how many times each experiment was replicated
- The statistical test(s) used and whether they are one- or two-sided (note: only common tests should be described solely by name; more complex techniques should be described in the Methods section)
- A description of any assumptions or corrections, such as an adjustment for multiple comparisons
- The test results (e.g. P values) given as exact values whenever possible and with confidence intervals noted
- A clear description of statistics including central tendency (e.g. median, mean) and variation (e.g. standard deviation, interquartile range)
- Clearly defined error bars

See the web collection on [statistics for biologists](#) for further resources and guidance.

► Software

Policy information about [availability of computer code](#)

7. Software

Describe the software used to analyze the data in this study.

An implementation of the approach is freely available at <https://github.com/haschard/CMS>

For manuscripts utilizing custom algorithms or software that are central to the paper but not yet described in the published literature, software must be made available to editors and reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). *Nature Methods* [guidance for providing algorithms and software for publication](#) provides further information on this topic.

► Materials and reagents

Policy information about [availability of materials](#)

8. Materials availability

Indicate whether there are restrictions on availability of unique materials or if these materials are only available for distribution by a for-profit company.

NA

9. Antibodies

Describe the antibodies used and how they were validated for use in the system under study (i.e. assay and species).

NA

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.

NA

b. Describe the method of cell line authentication used.

NA

c. Report whether the cell lines were tested for mycoplasma contamination.

NA

d. If any of the cell lines used are listed in the database of commonly misidentified cell lines maintained by [ICLAC](#), provide a scientific rationale for their use.

NA

► Animals and human research participants

Policy information about [studies involving animals](#); when reporting animal research, follow the [ARRIVE guidelines](#)

11. Description of research animals

Provide details on animals and/or animal-derived materials used in the study.

NA

Policy information about [studies involving human research participants](#)

12. Description of human research participants

Describe the covariate-relevant population characteristics of the human research participants.

NA