Partitioning gene-based variance of complex traits by gene score regression

Yue Li*, Si Yi Li, Wenmin Zhang, Tianyi Liu
School of Computer Science, McGill University

Abstract

Motivation

Understanding the biological mechanism of complex phenotypes is challenging due to the lack of efficient approaches that can associate the vast majority of the genome-wide association studies (GWAS) loci in the non-coding regions of the human genome with relevant genes and ultimately the downstream pathways. Transcriptome-wide association studies (TWAS) provides a way to associate genes with phenotype of interest by correlating GWAS summary statistics with expression quantitative trait loci (eQTL) summary statistics obtained from a reference panel. However, genes that are correlated by the predicted gene expression may exhibit high TWAS statistic even though they are the causal genes for the trait. Existing gene set enrichment analysis assume independence of genes and may therefore lead to false discoveries.

Results

We propose a novel statistical method called Gene Score Regression (GSR). The rationale of GSR is based on the insight that genes that are highly correlated with the causal genes in the causal gene set or pathways will exhibit high marginal TWAS statistic. Consequently, by regressing on the genes' marginal statistic using the sum of the gene-gene correlation scores in each gene set, we can assess the amount of phenotypic variance explained by the predicted expression of the genes in that gene set. Our approach does not only operates on summary statistics without requiring individual genotype and phenotype but can also work with observed

*Correspondance: yueli@cs.mcgill.ca

gene expression and phenotype data without the need of genotype. Based on simulation, GSR demonstrates superior power and better controlled Type I Error rate over the existing methods.

We applied GSR to investigate 28 complex traits using diverse genome-wide and knowledge-based gene sets such as REACTOME and KEGG from MSigDB and also 205 cell-type-specific gene sets derived from observed gene expression. The significant gene sets detected by GSR are by and large consistent with the known biology of the traits. We also demonstrated the utility of using GSR on cancer data where only gene expression and tumor/normal tissue labels are available. Overall, GSR is not only accurate but efficient method usually taking less than 5 minutes to perform the full analysis on one phenotype, thereby presenting as a significantly useful novel contribution to the analytical bioinformatic pipeline.

Availability

The GSR software is available at GitHub at `https://github.com/li-lab-mcgill/GSR`.

1 Introduction

Genome-wide association studies (GWAS) have been broadly successful in associating complex traits and diseases with genetic loci and estimating heritabilities across large number of individuals [1–4]. Over the past decade, GWAS have provided a large repository of genetic associations for hundreds of polygenic phenotypes in terms of the summary statistics associations [5], which are the estimated effect sizes and their standard errors for each single nucleotide polymorphisms (SNPs) analysed in a GWAS [6]. This enables a facet of summary-based approaches such as partitioning heritability [7], inferring causal SNPs [8], and pathway enrichment analysis for complex traits [9]. Despite its success, linking these genetic associations with biological mechanisms has been a challenge. One main reason is due to the fact that the majority of the GWAS loci are in the non-coding regions of the human genome. On the other hand, transcriptome-wide association studies (TWAS) [10–12] offers a systematic way to integrate GWAS and the reference genotype-gene expression datasets such as GTEx [13] via the expression quantitative loci (eQTL).

In TWAS, we can regress on the expression changes using the genotype information from the reference cohort (e.g., GTEx), apply such model to predict gene expression in the GWAS cohort, and then correlate the predicted gene expression with the GWAS phenotype to prioritize risk genes [10]. Moreover, when individual-level genotype and gene expression are not available, we can derive TWAS statistic using only the marginal summary statistic of SNPs for the GWAS and those of the gene expression [11]. The rationale behind TWAS is that the genetic correlation between a GWAS trait and gene expression implies that the co-localization of GWAS and eQTL SNPs can lead to phenotypic changes via the mediating genes.

However, as depicted in Figure 1a-c, TWAS is often confounded by the gene-gene correlation of the genetically predicted gene expression due to the SNP-SNP correlation or commonly known as the linkage disequilibrium (LD) [12]. Consequently, relying on the TWAS statistic may lead to false positive discoveries of causal genes and pathways. One approach to address this problem is to fine-map causal genes by inferring the posterior probabilities of configurations of each gene being causal in a defined GWAS loci and then perform gene set enrichment using the credible gene sets of prioritized genes [14]. However, this approach is computationally expensive, is only restricted to GWAS loci, and is sensitive to the arbitrary thresholds used for deciding the credible gene set and the maximum number of causal gene per locus.

Another method called PASCAL [9] projects SNP signals onto genes while correcting for LD and then performs pathway enrichments as the aggregated transformed gene scores, which asymptotically follows a chi-squared distribution. However, PASCAL does not leverage the eQTL information for each SNP thereby assuming that *apriori* each SNP have equal effect on the gene. Stratified LD score regression (LDSC) offers a principle way to partition the SNP heritability into functional categories that are defined based on tissue or cell-type specific epigenomic regions [7] or eQTL regions of the genes that are expressed specifically in one tissue against all other tissue samples [15]. Although LDSC is able to obtain biologically meaningful tissue-specific enrichments, it operates at the SNP level, making it difficult to use in assessing the enrichments of pathway and gene sets.

Moreover, neither PASCAL nor LDSC is able to integrate the observed gene expression data that are broadly available across diverse studies of diseases including cancers, which are available from the The Cancer Genome Atlas (TCGA). Although the expression-based method namely gene enrichment analysis (GSEA) is often used with the observed gene expression and phenotypes [16], it does not account for the gene-gene correlation, which is distinct from TWAS-induced correlation but is rather due to the sharing of transcriptional regulatory network among genes. For example, if the expression of the disease-causing gene in the causal pathway is highly correlated with the expression of non-causal genes in the non-causal pathways, GSEA will likely produce false positives for the non-causal pathways. In this paper, we describe a novel, powerful, and unified method for gene set enrichments to facilitate investigating the underlying mechanisms of complex traits and cancers.

2 Results

2.1 Gene scores correlate with TWAS statistics in polygenic complex traits

We defined *gene score* for each gene as its sum of squared Pearson correlation with all of the genes. We calculated TWAS marginal statistic as the product of GWAS summary statistic and eQTL weights derived from the GTEx whole blood samples. To assess the impact of gene-gene correlation on TWAS statistic, we correlated the gene scores with the TWAS marginal statistics for 28 complex traits. We observed high correlation in many traits with the top trait being Schizophrenia with correlation as high as 0.76 (Figure 1d). Overall, most traits have gene-score correlation with the marginal TWAS statistic above 0.4. This implies a pervasive confounding impacts on the downstream analysis using the TWAS statistic (Figure 1e) when using existing approaches that mostly assume independence of genes.

To address this challenge, we describe a novel method called Gene Score Regression (GSR). The rationale of our approach is that genes that are highly correlated with the causal genes in the causal gene set or pathways will exhibit high marginal TWAS statistic. Consequently, by regressing on the genes' marginal statistic using the sum of the gene-gene correlation scores in each gene set, we can assess the amount of phenotypic variance explained by the predicted expression of the genes in that gene set. We then calculate the statistical significance of each gene set based on the z-score of the linear regression coefficients in the GSR model. Details are described in Section 4.1.

2.2 GSR demonstrates improved power in pathway enrichments

To evaluate our proposed approach, we performed a realistic simulation using reference genotype from 1000 Genome European population [17] and pathway information from MSigDb [16] (Section 4.3). We compared our approach with 3 existing methods, namely PASCAL [9], LDSC [7], and FOCUS [14] outlined in Table 1. Details for running each program are described in Section 4.5. Compared to PASCAL and LDSC, GSR demonstrates superior sensitivity in detecting causal pathways with improved statistical power as well as competitive specificity in controlling false positives.

Notably, the FOCUS-predicted 75%, 90%, 99% credible gene sets are also significantly enriched for causal pathways. This is not surprising given that FOCUS can accurately fine-map causal genes in the well-defined simulation settings. However, FOCUS is at least 20 times slower than GSR. For the simulated data, FOCUS took 30 minutes to fine-map all of the genes in GWAS loci whereas GSR took under 3 minute to test for pathway enrichments on the same machine. Also, because GSR operates at genome-wide level, no threshold is needed to decide what genes to be included whereas FOCUS needs user-defined threshold for constructing the credible gene set for the subsequent hypergeometric enrichment test.

We then varied four different settings (Supplementary Figure S1): (a) number of causal SNP per gene; (2) SNP-gene heritabilities; (3) gene-phenotype variance explained; (4) overlapping causal pathway. We focused our comparison with PASCAL because it directly tests for pathway associations and has been demonstrated to outperform other relevant enrichment methods [9]. At all settings, our approach demonstrates an improved power in detect the causal pathway compared to PASCAL. Notably, our model is able to detect causal pathways even when the proportion of variance explained by the gene expression is low. In contrast, a lot of causal pathways are not deemed significant by PASCAL based on the p-value threshold of 0.001, which was set based on the Bonferroni q-value $<$ 0.1 after correcting for multiple testing on approximately 100 pathways tested per simulation.

2.3 Improved power in pathway enrichment when using the observed gene expression

One unique feature of GSR is the ability to run not on only the summary statistics but also the observed gene expression, where the gene-gene expression correlation is directly estimated from the in-sample gene expression (Section 4.2). To evaluate the accuracy of this application, we simulated gene expression and phenotype for 1000 individuals, which were provided as input to GSR for pathway enrichment analysis. As a comparison, we applied GSR to the summary statistics generated from the same dataset.

Same as the simulation above, the SNP-expression weights were estimated from a separate set of 500 reference individuals whereas the SNP-phenotype associations were estimated from only 1000 individuals. Notably, the sample size for the GWAS cohort is much smaller than the previous application to mimic the real data where usually fewer than 1000 individuals have both the RNA-seq and phenotype available (e.g., TCGA). Additionally, we applied GSEA [16] to the same dataset with the observed gene expression.

We observed an improved power of GSR when using the observed gene expression (GSR_obsExprs) over GSR using the summary statistics (GSR_sumstat) (Figure 3). Here, GSEA also performs well in this application outperforming GSR_sumstat but fall behind GSR_obsExprs. We also compared the performances of GSR_obsExprs with GSEA on various settings in the simulations and obtained consistent conclusion (Supplementary Figure S2).

2.4 Gene set enrichments in complex traits

We then applied our approach to investigate pathway enrichments of 28 complex GWAS traits using their publicly available summary statistics and the precomputed tissue-specific genotype-expression weights based on the whole blood GTEx samples. We listed the top 10 enrichments over gene sets from MSigDB and Gene Ontology terms of the lipid trait High Density Lipoprotein (HDL) and the autoimmune trait Lupus in Table 2.

The significantly enriched gene sets are biologically meaningful. The enriched gene sets for HDL predominantly involve lipid metabolism whereas for Lupus are enriched for interferon signalling pathways, which is immunological hallmark. We also compared the enrichments between GSR and hypergeometric tests over the FOCUS 90% credible gene sets for the two traits (Figure 4). We observed improved significance of the commonly known associated gene sets compared to FOCUS-Hypergeometric approach.

Additionally, we applied GSR to test cell-type-specific enrichments using 205 cell types, 48 of which were derived from GTEx and 157 cell types were derived from Franke lab dataset [15]. We observed biologically meaningful cell types among the complex traits. In particular, Schizophrenia is highly enriched for CNS cell types, Lupus is enriched for immune cell types, Crohn's disease is enriched for immune and cells in digestive tracts, and coronary artery disease is enriched for heart-specific cell types. Lastly, we correlated traits based on their gene set enrichments and observed meaningful phenotypic cluster, highlighting common biology among the related phenotypes (Supplementary FigureS3).

2.5 Application on observed gene expression

We then applied GSR to the uniformly processed TCGA+GTEx datasets for three well powered cancers namely breast cancer (BRCA), thyroid cancer (THCA), and prostate cancer (PRAD) [18]. We tested the enrichments of each tumor type for the 186 oncogenic gene sets and the more general 1050 gene sets from BIOCARTA, KEGG, and REACTOME. Overall, we observed a significantly higher enrichments for the oncogenic signatures compared to the more general gene sets across all 3 tumour types Figure 6. As a comparison, we also ran GSEA and observed qualitatively similar enrichments Figure S4.

3 Discussion

In this paper, we describe GSR, an efficient method to test for gene set or pathway enrichments using either the summary statistics information or the observed gene expression and phenotype information. We demonstrate robust and powerful detection of causal pathways in extensive simulation using our proposed method compared with several state-of-the-art methods. When applying for 28 complex traits, we also obtained biologically meaningful enrichments for relevant gene sets and pathways. One unique feature of our approach is that it can leverage the observed individual-level gene expression that are broadly available to calculate more accurate in-sample gene-gene correlation. Indeed, we observed more accurate detection of causal pathway for modest sample size (1000 individuals) where the phenotype and gene expression are available compared to GSR operating only on summary statistics.

Our approach is easy to implement and very efficient to run with standard python libraries. In particular, GSR took only 3-5 minutes running on the full summary statistics and under 5 minutes on the full gene expression data with 1000 samples and 20,000 genes to test for enrichments of over 4000 gene sets. Together, we envision that GSR will be a valuable tool for the bioinformatic community and statistical genetic community as a fast way to investigate the functional implications of complex polygenic traits and cancers.

4 Method

4.1 Gene score regression on genetically predicted gene expression

Assuming gene expression $\mathbf{A} \in \mathbb{R}^{N\times G}$ for gene g from a reference panel (e.g., GTEx) of N_{ref} individuals are a linear combination of the genotype $\mathbf{X} \in \mathbb{R}^{N\times P}$:

$$\mathbf{A}_g^{\text{ref}} = \mathbf{X}^{\text{ref}}\mathbf{W}_g + \epsilon_g \tag{1}$$

The estimate $\hat{\mathbf{W}}_g$ can be obtained by either ordinary least square, best linear unbiased predictor (BLUP), LASSO, or elastic net.

Also, assuming that the phenotype for N_{gwas} individuals is a linear combination of the gene expression for N_{gwas} individuals.

$$\mathbf{y} = \mathbf{A}^{\text{gwas}}\boldsymbol{\alpha} + \epsilon \tag{2}$$

To obtain the marginal effect estimate of each gene, we can regress each gene separately on the phenotype:

$$\mathbf{y} = \mathbf{A}_g^{\text{gwas}}\alpha_g + \epsilon_{y,g} \tag{3}$$

The ordinary least squared solution of gene effect size α_g is then:

$$\hat{\alpha}_g = (\mathbf{A}_g^{\text{gwas}\top}\mathbf{A}_g^{\text{gwas}})^{-1}\mathbf{A}_g^{\text{gwas}}\mathbf{y} \tag{4}$$

In GWAS, the gene expression $\mathbf{A}_g^{\text{gwas}}$ are not available, and only the genotype $\mathbf{X}^{\text{gwas}}$ and phenotype $\mathbf{y}$ are measured. In this regard, we predict them as $\hat{\mathbf{A}}_g^{\text{gwas}} \equiv \hat{\mathbf{A}}_g = \mathbf{X}\hat{\mathbf{W}}_g$, where the linear weights $\hat{\mathbf{W}}_g$ are estimated based on the above reference genotype-expression cohort.

For the following derivation, we assume that both the phenotype $\mathbf{y}$ and genotype matrices $\mathbf{X}$ are standardized such that $\mathbf{y} = \sum_i y_i$, $\frac{1}{N}\mathbf{y}^\top\mathbf{y} = 1$, $\sum_i X_{ij} = 0$ and $\frac{1}{N}\mathbf{X}_j^\top\mathbf{X}_j = 1$.

Substituting $\mathbf{A}_g$ in the OLS with the predicted expression $\hat{\mathbf{A}}_g$ gives

$$\begin{aligned}\hat{\alpha}_g &= (\mathbf{A}_g^\top\mathbf{A}_g)^{-1}\mathbf{A}_g\mathbf{y} \\ &\approx (\hat{\mathbf{A}}_g^\top\hat{\mathbf{A}}_g)^{-1}\hat{\mathbf{A}}_g\mathbf{y} \\ &= (\hat{\mathbf{W}}_g^\top\mathbf{X}^\top\mathbf{X}\mathbf{W}_g)^{-1}\hat{\mathbf{W}}_g^\top\mathbf{X}^\top\mathbf{y} \\ &= (\hat{\mathbf{W}}_g^\top N_{\text{gwas}}\mathbf{\Sigma}_{\text{gwas}}\hat{\mathbf{W}}_g)^{-1}\hat{\mathbf{W}}_g^\top\mathbf{X}^\top\mathbf{y} \qquad (5)\end{aligned}$$

where $\mathbf{\Sigma}_{\text{gwas}} = \frac{1}{N}\mathbf{X}_{\text{gwas}}^\top\mathbf{X}_{\text{gwas}}$ is the SNP-SNP Pearson correlation and $\mathbf{X}^\top\mathbf{y} = \mathbf{z}$ is the summary statistics or z-scores when both the phenotype and genotype are standardized. Because the individual-level genotype $\mathbf{X}_{\text{gwas}}$ is easily accessible, we approximate this correlation using 1000 Genome reference (1KG) genotype $\mathbf{X}_{\text{gwas}}$ as follows:

$$\begin{aligned}\mathbf{\Sigma}_{\text{gwas}} &= \frac{1}{N_{\text{gwas}}}\mathbf{X}_{\text{gwas}}^\top\mathbf{X}_{\text{gwas}} \\ &\approx \frac{1}{N_{\text{1KG}}}\mathbf{X}_{\text{1KG}}^\top\mathbf{X}_{\text{1KG}} \equiv \mathbf{\Sigma}_{\text{1KG}}\end{aligned}$$

Substituting $\mathbf{\Sigma}_{\text{gwas}}$ with $\mathbf{\Sigma}_{\text{1KG}}$ in (5) gives:

$$\begin{aligned}\hat{\alpha}_g &= (\hat{\mathbf{W}}_g^\top\frac{N_{\text{gwas}}}{N_{\text{1KG}}}\mathbf{X}_{\text{1KG}}^\top\mathbf{X}_{\text{1KG}}\hat{\mathbf{W}}_g)^{-1}\hat{\mathbf{W}}_g^\top\mathbf{X}^\top\mathbf{y} \\ &= \frac{N_{\text{1KG}}}{N_{\text{gwas}}}(\hat{\mathbf{W}}_g^\top\mathbf{X}_{\text{1KG}}^\top\mathbf{X}_{\text{1KG}}\hat{\mathbf{W}}_g)^{-1}\hat{\mathbf{A}}_g^\top\mathbf{y} \\ &= \frac{N_{\text{1KG}}}{N_{\text{gwas}}}(\hat{\mathbf{A}}_{\text{g,1KG}}^\top\hat{\mathbf{A}}_{\text{g,1KG}})^{-1}\hat{\mathbf{A}}_g^\top\mathbf{y}\end{aligned}$$

Here we denote the predicted gene expression g in the 1KG cohort as $\hat{\mathbf{A}}_{\text{g,1KG}} = \mathbf{X}_{\text{1KG}}\hat{\mathbf{W}}_g$. Because we have access to $\mathbf{X}_{\text{1KG}}$ reference genotype, we can directly compute $\hat{\mathbf{A}}_{\text{g,1KG}}$. Importantly, we standardize the predicted expression $\hat{\mathbf{A}}_{\text{g,1KG}}$ such that $\frac{1}{N_{\text{1KG}}}\hat{\mathbf{A}}_{\text{g,1KG}}^\top\hat{\mathbf{A}}_{\text{g,1KG}} = 1$.

Therefore, we can simplify the above equation as

$$\begin{aligned}\hat{\alpha}_g &= \frac{N_{\text{1KG}}}{N_{\text{gwas}}}(\hat{\mathbf{A}}_{\text{g,1KG}}^\top\hat{\mathbf{A}}_{\text{g,1KG}})^{-1}\hat{\mathbf{A}}_g^\top\mathbf{y} \\ &= \frac{N_{\text{1KG}}}{N_{\text{gwas}}}\frac{1}{N_{\text{1KG}}}\hat{\mathbf{A}}_g^\top\mathbf{y} \\ &= \frac{1}{N_{\text{gwas}}}\hat{\mathbf{A}}_g^\top\mathbf{y} \qquad (6)\end{aligned}$$

Substituting $\mathbf{y}$ in (6) by its multiple regression formula (2) and replacing the true expression $\mathbf{A}^{\text{gwas}}$ by the predicted counterpart $\hat{\mathbf{A}}^{\text{gwas}}$ lead to the following derivation:

$$\begin{aligned}
\hat{\alpha}_g &= \frac{1}{N_{\text{gwas}}}\hat{\mathbf{A}}_g^\top(\hat{\mathbf{A}}\boldsymbol{\alpha} + \epsilon) \\
&= \frac{1}{N_{\text{gwas}}}\sum_k \hat{\mathbf{A}}_g^\top\hat{\mathbf{A}}_k\alpha_k + \frac{1}{N_{\text{gwas}}}\hat{\mathbf{A}}_g^\top\epsilon \\
&= \frac{1}{N_{\text{gwas}}}\sum_k \hat{\mathbf{W}}_g^\top\mathbf{X}_{\text{gwas}}^\top\mathbf{X}_{\text{gwas}}\hat{\mathbf{W}}_k\alpha_k + \epsilon' \\
&= \frac{1}{N_{\text{gwas}}}\sum_k \hat{\mathbf{W}}_g^\top(\frac{N_{\text{gwas}}}{N_{\text{1KG}}}\mathbf{X}_{\text{1KG}}^\top\mathbf{X}_{\text{1KG}})\hat{\mathbf{W}}_k\alpha_k + \epsilon' \\
&= \sum_k \frac{1}{N_{\text{1KG}}}\hat{\mathbf{W}}_g^\top\mathbf{X}_{\text{1KG}}^\top\mathbf{X}_{\text{1KG}}\hat{\mathbf{W}}_k\alpha_k + \epsilon' \\
&= \sum_k \frac{1}{N_{\text{1KG}}}\hat{\mathbf{A}}_{\text{g,1KG}}^\top\hat{\mathbf{A}}_{\text{k,1KG}}\alpha_k + \epsilon'
\end{aligned}$$

Since $\hat{\mathbf{A}}_{\text{g,1KG}}$ is standardized, we observe that the Pearson correlation of the predicted gene expression between gene g and gene k is $r_{gk} = \frac{1}{N_{\text{1KG}}}\hat{\mathbf{A}}_{\text{g,1KG}}^\top\hat{\mathbf{A}}_{\text{k,1KG}}$. Therefore,

$$\hat{\alpha}_g = \sum_k r_{gk}\alpha_k + \epsilon' \tag{7}$$

Consider a chi-squared variable as $\chi_g^2 \equiv N\hat{\alpha}_g^2$ and its expectation is then

$$\begin{aligned}
\mathrm{E}[\chi^2] &= \mathrm{E}[N_{\text{gwas}}\hat{\alpha}_g^2] \\
&= N_{\text{gwas}}\mathrm{E}\left[\left(\sum_k r_{gk}\alpha_k + \epsilon'\right)^2\right] \\
&= N_{\text{gwas}}\sum_k \mathrm{E}[r_{gk}^2]\mathrm{E}[\alpha_k^2] + \mathrm{E}[\epsilon'^2]
\end{aligned} \tag{8}$$

Here, we assume all of the random variables are independent. Therefore, the expectations of all of the cross terms in the squared of sums become zero. Also, we assume that $\mathrm{E}[\alpha] = 0$ and

$\mathrm{E}[\epsilon] = 0$. Therefore, $\mathrm{E}[\alpha_k^2] = \mathrm{Var}[\alpha_k]$, and

$$\begin{aligned}
\mathrm{Var}[\epsilon'] &= \mathrm{Var}\left[\frac{1}{N_{\text{gwas}}}\hat{\mathbf{A}}_g^\top \epsilon\right] \\
&= \frac{1}{N_{\text{gwas}}^2}\hat{\mathbf{A}}_g^\top \mathrm{Var}[\epsilon]\hat{\mathbf{A}}_g \\
&= \frac{1}{N_{\text{gwas}}^2}\hat{\mathbf{A}}_g^\top \hat{\mathbf{A}}_g \sigma_\epsilon^2 \\
&= \frac{1}{N_{\text{gwas}}^2}(\hat{\mathbf{W}}_g^\top \mathbf{X}_{\text{gwas}}^\top \mathbf{X}_{\text{gwas}} \hat{\mathbf{W}}_g)\sigma_\epsilon^2 \\
&= \frac{1}{N_{\text{gwas}}^2}(\hat{\mathbf{W}}_g^\top (\frac{N_{\text{gwas}}}{N_{\text{1KG}}} \mathbf{X}_{\text{1KG}}^\top \mathbf{X}_{\text{1KG}}) \hat{\mathbf{W}}_g)\sigma_\epsilon^2 \\
&= \frac{1}{N_{\text{gwas}}}(\frac{1}{N_{\text{1KG}}} \hat{\mathbf{W}}_g^\top \mathbf{X}_{\text{1KG}}^\top \mathbf{X}_{\text{1KG}} \hat{\mathbf{W}}_g)\sigma_\epsilon^2 \\
&= \frac{1}{N_{\text{gwas}}}(\frac{1}{N_{\text{1KG}}} \hat{\mathbf{A}}_{\text{g,1KG}}^\top \hat{\mathbf{A}}_{\text{g,1KG}})\sigma_\epsilon^2 \\
&= \frac{1}{N_{\text{gwas}}}\sigma_\epsilon^2
\end{aligned}$$

The last equation assumes that the predicted gene expression of 1KG cohort is standardized such that $\frac{1}{N_{\text{1KG}}}\hat{\mathbf{A}}_{\text{g,1KG}}^\top \hat{\mathbf{A}}_{\text{k,1KG}} = 1$.

Substituting $\mathrm{E}[\alpha_k^2]$ and $\mathrm{E}[\epsilon'^2]$ in (8) with the above variance terms gives:

$$\begin{aligned}
\mathrm{E}[\chi^2] &= N_{\text{gwas}}\left(\sum_k \mathrm{E}[r_{gk}^2]\mathrm{Var}[\alpha_k] + \frac{1}{N_{\text{gwas}}}\sigma_\epsilon^2\right) \\
&= N_{\text{gwas}}\sum_k \mathrm{E}[r_{gk}^2]\mathrm{Var}[\alpha_k] + \sigma_\epsilon^2 \qquad (9)
\end{aligned}$$

Suppose there are C gene sets $\mathcal{C}_1, \ldots, \mathcal{C}_C \subset \{1, \ldots, G\}$. We refine the variance of gene effect $\mathrm{Var}[\alpha_k]$ by the sum of variance components corresponding to the gene set that gene k belongs:

$$\mathrm{Var}[\alpha_k] = \sum_{c:k\in C_c} \tau_c \qquad (10)$$

where $\tau_c = \frac{\sum_{j\in C_c}\mathrm{Var}[\alpha_j]}{|C_c|}$ for $|C_c|$ genes in the gene set.

Therefore, we have

$$\begin{aligned}\mathrm{E}[\chi^2] &= N_{\textsf{gwas}} \sum_k \mathrm{E}[r_{gk}^2] \sum_{c:k\in C_c} \tau_c + \sigma_\epsilon^2 \\ &= N_{\textsf{gwas}} \sum_c \tau_c \sum_{k\in\mathcal{C}_c} \mathrm{E}[r_{gk}^2] + \sigma_\epsilon^2\end{aligned}$$

Assuming the observed correlation is its expectation minus a small constant $1/N_{\textsf{gwas}}$, then $r_{gk}^2 = \mathrm{E}[r_{gk}^2] - 1/N_{\textsf{gwas}}$, then $\mathrm{E}[r_{gk}^2] = r_{gk}^2 + 1/N_{\textsf{gwas}}$.

Therefore,

$$\begin{aligned}\mathrm{E}[\chi^2] &= N_{\textsf{gwas}} \sum_c \tau_c \sum_{k\in\mathcal{C}_c} \mathrm{E}[r_{gk}^2] + \sigma_\epsilon^2 \\ &= N_{\textsf{gwas}} \sum_c \tau_c \sum_{k\in\mathcal{C}_c} (r_{gk}^2 + 1/N_{\textsf{gwas}}) + \sigma_\epsilon^2 \\ &= N_{\textsf{gwas}} \sum_c \tau_c \sum_{k\in\mathcal{C}_c} r_{gk}^2 + N_{\textsf{gwas}} \sum_c \tau_c \sum_{k\in\mathcal{C}_c} 1/N_{\textsf{gwas}} + \sigma_\epsilon^2 \\ &= N_{\textsf{gwas}} \sum_c \tau_c \sum_{k\in\mathcal{C}_c} r_{gk}^2 + \sum_c \tau_c + \sigma_\epsilon^2 \\ &= N_{\textsf{gwas}} \sum_c \tau_c l(g,c) + \sum_c \tau_c + \sigma_\epsilon^2\end{aligned}$$

where we define

$$l(g,c) = \sum_{k\in\mathcal{C}_c} r_{gk}^2$$

as the *gene score* and $\sum_c \tau_c + \sigma_\epsilon^2$ is the total variance of the phenotype $\mathbf{y}$. Because the phenotype is standardized with variance equal to 1, we have $\mathrm{Var}[y] = \sum_c \tau_c + \sigma_\epsilon^2 = 1$.

Therefore, we arrive at the main equation:

$$\mathrm{E}[\chi_g^2] = N_{\textsf{gwas}} \sum_c \tau_c l(g,c) + 1 \tag{11}$$

To perform gene score regression (GSR), we regress the gene scores on the observed chi-squared:

$$N_{\textsf{gwas}} \hat{\alpha}^2{}_g \sim N_{\textsf{gwas}} \sum_c \tau_c l(g,c) + 1 \tag{12}$$

□

In practice, many gene sets are not disjoint and tend be correlated due to the sharing of common genes. In practice, we regress one gene set at a time along with a “dummy” gene set that include the union of all of the gene sets. We also include an intercept in the regression model to properly control non-gene-set biases.

Our approach is inspired by LD score regression (LDSC) [7, 15]. However, LDSC operates at SNP-level rather than gene-level and does not take into account the gene expression weights in the regression.

4.2 Gene score regression on observed gene expression data

When the real gene expression $\mathbf{A}$ (*not* genotype-predicted expression $\hat{\mathbf{A}}$) and the phenotype $(\mathbf{y})$ for N_{real} individuals are available for the same cohort (e.g., The Cancer Genome Atlas or TCGA), our method stays mostly the same. In particular, we assume that the phenotype and gene expression are all standardized to have zero mean and standard deviation equal to 1 across samples. Based on the same linear model $\mathbf{y} = \mathbf{A}\boldsymbol{\alpha} + \epsilon$, it is easy to see that the GSR formula has the same form:

$$N_{\text{real}}\hat{\alpha}^2{}_g \sim N_{\text{real}}\sum_c \tau_c l(g,c) + 1 \tag{13}$$

where

- $\hat{\alpha}_g = \mathbf{A}_g^\top \mathbf{y}$
- $l(g,c) = \sum_{k\in\mathcal{C}_c} r_{gk}^2$, and $r_{gk} = \frac{1}{N_{\text{real}}}\mathbf{A}_g^\top \mathbf{A}_k$
- τ_c is the regression coefficient for gene set c

Despite the same regression form, the biological implication here is quite different from the above genotype-derived transcriptomic approach. Here the gene-gene correlation r_{gk} is not directly due to LD (unless the genes share the same eQTL). Rather they reflect the gene co-expression program. For example, two genes may be regulated by a common set of transcriptional regulators such as transcription factors or microRNAs such that their expression are correlated. If one gene is a causal gene and the other is not, we will see inflated summary statistic for the non-causal gene, thereby confounding the detection for causal pathways. Nonetheless, despite different biological implication, our GSR approach is applicable in this case because it leverages such expression correlation to for calculating the pathway enrichments.

4.3 Simulation

4.3.1 Simulation step 1: simulate gene expression

1. To simulate individual genotype, we first partitioned genotype data for 489 individuals of European ancestry in 1000 Genome [17] into independent LD blocks as defined by LDetect [19];

2. We standardized the simulated genotype $\mathbf{X}_{ref}$;

3. We then randomly sampled 100 LD blocks and use only those 100 LD blocks for the subsequent simulation;

4. To simulate genotype $\mathbf{X}_{\text{ref}}$ for each of the $N_{\text{ref}} = 500$ reference individuals, we randomly sampled each LD block from the 489 Europeans in 1000 Genome data and concatenate these LD blocks to create a simulated genotype;

5. We randomly sampled k in-cis causal SNPs per gene within $\pm$ 500 kb around the gene, where $k = 1$ (default). We also experimented different number of causal SNPs $k \in \{2, 3, \text{all in-cis SNPs}\}$

6. We sampled SNP-gene weights $\mathbf{W}_g \sim \mathcal{N}(0, h_g^2/k)$ and gene heritability $h_g^2 = 0.1$ (default). We also experimented different gene heritability $h_g^2 = \{0.2, 0.3, 0.4, 0.5\}$

7. We then simulated gene expression $\mathbf{A}_{g,ref} = \mathbf{X}_{ref}\mathbf{W}_g + \epsilon$, where

   $$\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$$

   where

   $$\sigma_\epsilon^2 = \frac{1}{N_{\text{ref}}}||\mathbf{X}_{\text{ref}}\mathbf{W}_g||^2(\frac{1}{h_g^2} - 1)\mathbf{I}_{N_{\text{ref}}}$$

   such that the simulated noise conforms the heritability setting:

   $$\frac{1 - h_g^2}{h_g^2} = \frac{\sigma_\epsilon^2}{||\mathbf{X}_{\text{ref}}\mathbf{W}_g||^2/N_{\text{ref}}}$$

8. Apply LASSO regression $\mathbf{A}_{g,\text{ref}} \sim \bar{\mathbf{X}}\mathbf{W}_g$ to get $\hat{\mathbf{W}}_g$ for each gene

4.3.2 Simulation step 2: simulate GWAS phenotype

1. We simulated N_{gwas}=50,000 GWAS individuals by the 100 predefined LD blocks among the 489 Europeans in 1000 Genome data;

2. We standardized the simulated genotype $\mathbf{X}_{\text{gwas}}$;

3. We then sampled a causal pathway $\mathcal{C}_c$ from MSigDB such that all of the $G_c \equiv |\mathcal{C}_c|$ genes in $\mathcal{C}_c$ are causal genes for the phenotype

4. We processed the pathways based on the causal pathway in two ways:

   (a) For unique causal pathway by definition, we removed the causal genes from other non-causal pathways. If the resulting non-causal pathway contain fewer than 5 genes, we removed those pathways;

   (b) For a more realistic setting, we also allowed other pathways to overlap with the causal pathway;

5. Sample gene-phenotype effect $\boldsymbol{\alpha} \sim \mathcal{N}(0, \sigma_\alpha^2/G_c\mathbf{I}_{G_c})$, and phenotypic variance explained by gene expression $\sigma_\alpha^2 = 0.1$ (default). We also experimented different variance $\sigma_\alpha^2 \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$;

6. Simulate gene expression $\mathbf{A}_c$ as in step 1 for the N_{gwas} individuals, and standardize it to obtain $\bar{\mathbf{A}}_c$

7. Simulate phenotype using causal gene expression: $\mathbf{y} = \bar{\mathbf{A}}_c\boldsymbol{\alpha} + \epsilon_y$ where

$$\epsilon_y \sim \mathcal{N}(0, \sigma_{\epsilon_y}^2)$$

   where

$$\epsilon_y = \frac{1}{N_{\text{gwas}}}||\bar{\mathbf{A}}_c\boldsymbol{\alpha}||^2(\frac{1}{\sigma_\alpha^2} - 1)\mathbf{I}_{N_{gwas}}$$

   such that the simulated noise conforms the predefined variance-explained:

$$\frac{1-\sigma_\alpha^2}{\sigma_\alpha^2} = \frac{\sigma_{\epsilon_y}^2}{||\bar{\mathbf{A}}_c\boldsymbol{\alpha}||^2/N_{\text{gwas}}}$$

8. Compute GWAS summary statistic z-score $\mathbf{z} = \mathbf{X}_{\text{gwas}}^\top\mathbf{y}$

We performed 10 simulations per setting. Unless mentioned otherwise, while we were experimenting various settings, we kept the other settings at their default values:

1. $k = 1$ causal SNP per gene

2. Gene expression variance explained per causal SNP $h_g^2 = 0.1/k$

3. Phenotypic variance explained per gene $\sigma_\alpha^2 = 0.1$

4. One causal pathway. All of the genes in the causal pathway are causal for the phenotype. Other non-causal pathways do not contain any gene in the causal pathway

4.4 Data sets

GSR requires gene expression weights pre-estimated from a reference panel. To this end, we downloaded publicly available expression weight data from the TWAS/FUSION website (`http://gusevlab.org/projects/fusion/`) [11]. Reference LD was estimated in 1000 Genomes using 489 European individuals [17]. The GWAS summary statistics were downloaded from public database `https://data.broadinstitute.org/alkesgroup/sumstats_formatted/` [7].

The uniformly processed (normalized + batch-effect corrected gene) gene expression datasets from TCGA and GTEx were obtained from `https://figshare.com/articles/Data_record_3/5330593` [18]. Gene expression and phenotype were standardized before provided to the GSR software.

Gene sets were downloaded from the MSigDb website `http://software.broadinstitute.org/gsea/msigdb/index.jsp`. Here we combined BIOCARTA, KEGG and REACTOME to create a 1050 gene sets. We also downloaded the 4436 GO biological process terms as additional gene sets as well as the 189 gene sets involving the oncogenic signatures for the cancer analysis. Franke lab cell-type-specific gene expression dataset were obtained from `https://data.broadinstitute.org/mpg/depict/depict_download/tissue_expression`.

4.5 Running existing methods

4.5.1 PASCAL

PASCAL was downloaded from `https://www2.unil.ch/cbg/index.php?title=Pascal` [9]. We run the software by following the manual with default setting.

4.5.2 LDSC

Stratified LD score regression software was downloaded from `https://github.com/bulik/ldsc` [15]. Because LDSC operates on SNP level, we will need to obtain SNP-level annotation for each pathway. For each pathway, we computed the LD scores over all chromosomes by taking the SNPs within ± 500 kb of the genes in the pathway. We experimented the options of running LDSC with and without the 53 baseline annotations on our simulated data. We found that LDSC running without the 53 baseline worked better in our case. One possible reason is because the baseline annotations cover genome-wide SNPs whereas there are much fewer SNPs in the simulated pathways.

4.5.3 FOCUS

We obtained FOCUS [14] from `https://github.com/bogdanlab/focus`. We used FOCUS to infer the posterior probabilities of the each gene being causal to the phenotype across all of the LD blocks. We then took the 90% credible gene set as follows. We first summed all of the posterior over all of genes. We then sorted the genes by the decreasing order of their FOCUS-posteriors. We kept adding the top ranked the gene into the 90% credible gene until their posterior sum is equal or greater than the 90% of the total posterior. We used the 90% credible gene set for hypergeometric test for each pathway to compute the p-values. We also tried other credible sets ranging form 75% (including the fewest genes) to 99% (including the most genes).

4.5.4 GSEA

We obtain the GSEA software from `http://software.broadinstitute.org/gsea/index.jsp` [16]. We used the command-line version of GSEA to test for gene set enrichments using the observed gene expression and phenotype data.

4.6 Gene score regression software availability

GSR is implemented in Python using numpy and other standard Python libraries. The software is available here `https://github.com/li-lab-mcgill/GSR`.

5 Acknowledgment

The research is supported by NSERC Discovery Grant (RGPIN-2019-0621), FRQNT New Career Grant (FRQ-NT NC-268592), Canada First Research Excellence Fund (CFREF) Healthy Brains, Healthy Life (HBHL) New Investigator fund at McGill University and Montreal Neurologic Institute (MNI).

6 Tables

Software	Citatation	GWAS sumstat	TWAS sumstat	Gene expression	Running time
PASCAL	[9]	x			10 min
LDSC	[15]	x			$>$24 h†
FOCUS	[14]	x	x		$>$24 h
GSEA	[16]			x	10 min
GSR	proposed	x	x	x	3 min

Table 1: Comparison of existing methods with our proposed method GSR. † For custom gene sets, the main computation time for LDSC is calculating the LD score for all of the 1000 Genome SNPs.

trait	gs	p.value	t.value
HDL	REACTOME_CHYLOMICRON_MEDIATED_LIPID_TRANSPORT	3.23e-151	26.98
HDL	GO_NEUTRAL_LIPID_CATABOLIC_PROCESS	1.70e-132	25.14
HDL	GO_TRIGLYCERIDE_CATABOLIC_PROCESS	1.70e-132	25.14
HDL	GO_MACROMOLECULAR_COMPLEX_REMODELING	5.82e-131	24.98
HDL	GO_ACYLGLYCEROL_HOMEOSTASIS	1.11e-113	23.17
HDL	REACTOME_LIPOPROTEIN_METABOLISM	3.99e-112	23.00
HDL	REACTOME_LIPID_DIGESTION_MOBILIZATION_AND_TRANSPOR	5.93e-96	21.17
HDL	GO_PROTEIN_LIPID_COMPLEX_SUBUNIT_ORGANIZATION	2.58e-90	20.51
HDL	REACTOME_PACKAGING_OF_TELOMERE_ENDS	2.65e-42	13.74
HDL	KEGG_PPAR_SIGNALING_PATHWAY	1.72e-32	11.94
Lupus	GO_POSITIVE_REGULATION_OF_INTERFERON_ALPHA_PRODUCT	7.06e-150	26.85
Lupus	GO_POSITIVE_REGULATION_OF_INTERFERON_BETA_PRODUCTI	1.46e-101	21.82
Lupus	GO_RESPONSE_TO_MURAMYL_DIPEPTIDE1	2.26e-96	21.22
Lupus	GO_POSITIVE_REGULATION_OF_INTERLEUKIN_12_PRODUCTIO	1.39e-78	19.05
Lupus	GO_REGULATION_OF_INTERFERON_ALPHA_PRODUCTION	1.38e-72	18.27
Lupus	REACTOME_INTERFERON_GAMMA_SIGNALING	2.90e-65	17.28
Lupus	REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	1.30e-29	11.36
Lupus	KEGG_TOLL_LIKE_RECEPTOR_SIGNALING_PATHWAY	1.70e-24	10.26
Lupus	REACTOME_INTERFERON_SIGNALING	1.18e-13	7.44
Lupus	REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	5.11e-07	5.03

Table 2: Gene set enrichments on select traits. The enrichments were applied to 3 gene sets: (1) 4360 GO Biology Processes terms from MSigDb; (2) combined 1051 BIOCARTA, KEGG, REACTOME gene sets from MSigDb; (3) 207 gene sets each derived from GTEx and Franke tissue/cell-specifically expressed genes.

7 Figures

bioRxiv preprint doi: https://doi.org/10.1101/2020.01.08.899260; this version posted January 9, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

Figure 1

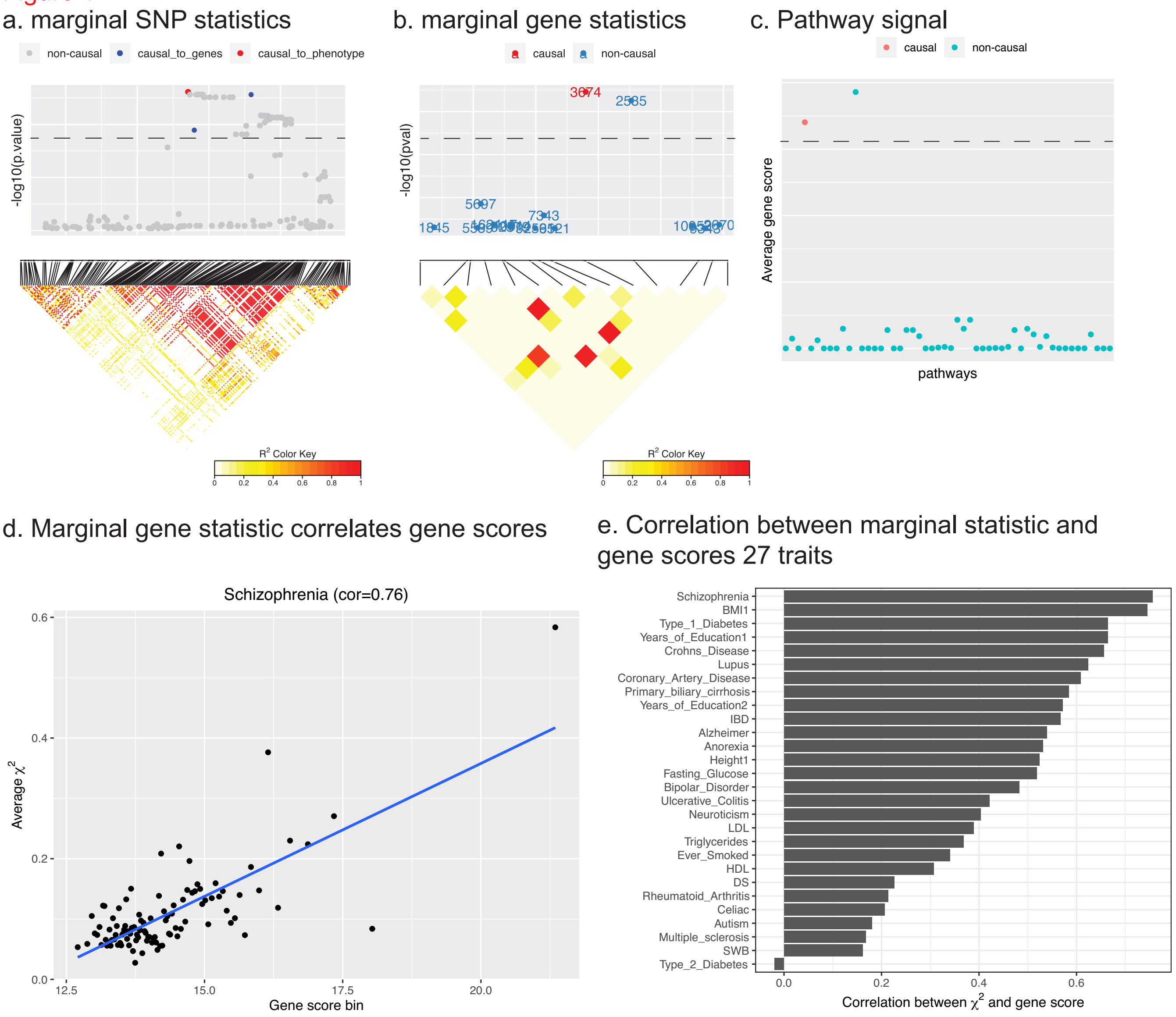

Figure 1: Problem overview. A hypothetical example that illustrates the confounding issue when using the genetically predicted transcriptome to assess the pathway enrichments for a target phenotype. (a) A GWAS locus. The y-axis is -logP values for SNP association for the target phenotype. The causal SNPs that are causal for the causal gene are in red. The SNPs that are causal for a non-causal gene are in blue. The rest of the SNPs are in grey. SNPs exhibit correlated signals due to the linkage disequilibrium (LD) as displayed by the upper triangle of the SNP-SNP Pearson correlation matrix; (b) Marginal gene association with the phenotype. The gene-gene correlation are partly induced by the SNP-SNP correlation (i.e., LD) and partly due to intrinsic co-regulatory expression program. (c) Pathway associations based on averaged gene associations. (d) Gene score were correlated with chi-squared of the TWAS marginal statistic for Schizophrenia. We binned genes by their gene scores. For each bin, we calculated the average gene scores and chi-squared (χ^2). (e) The same procedure in (d) was performed for distinct complex traits. The barplot display the Pearson correlation between the gene scores and TWAS marginal statistic for each trait.

Figure 2

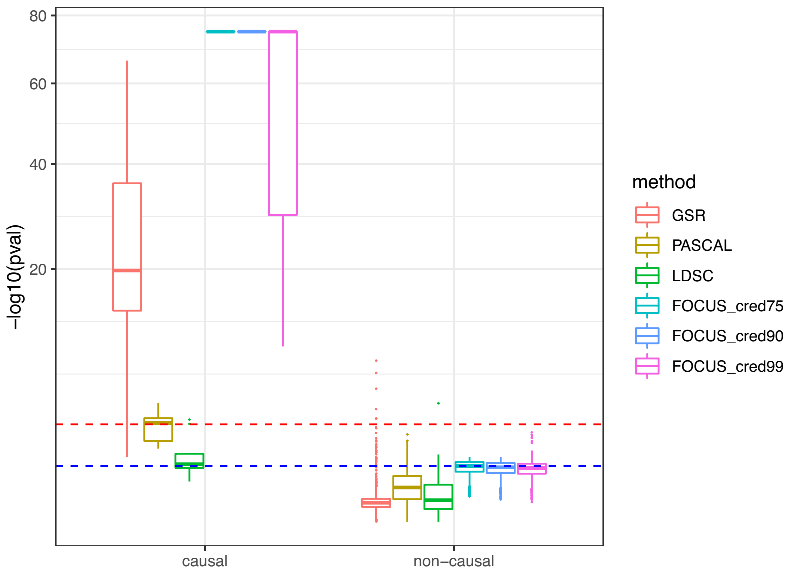

Figure 2: Evaluation of power and robustness in detecting causal pathways. We ran GSR along with 3 published methods namely PASCAL [9], LDSC [7] and FOCUS [14] with 10 simulation runs. For each method, the enrichment score for causal pathways and non-causal pathways are displayed. We experimented FOCUS with 75%, 90%, and 99% credible sets for the pathway enrichments. For the ease of comparison, we plotted the y-axis at the squared root of the -log p-values.

Figure 3

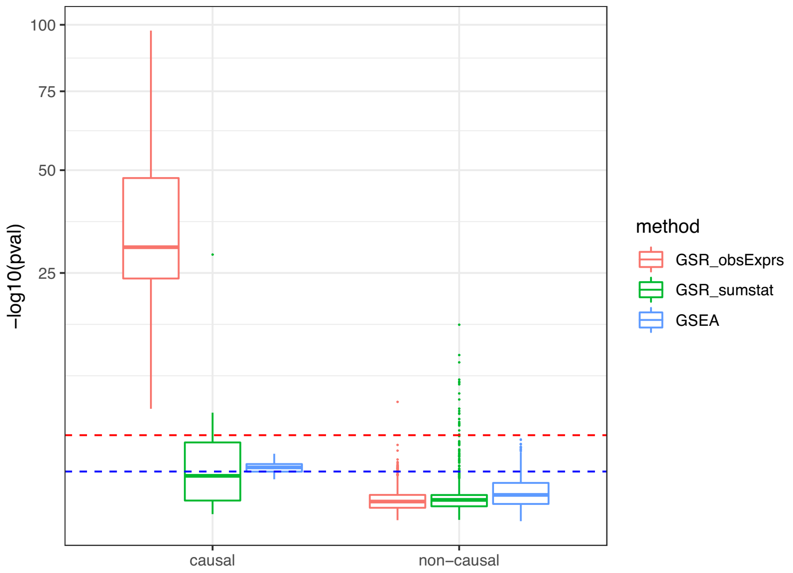

Figure 3: Comparison with GSEA [16] on pathway enrichment using observed gene expression.

Figure 4

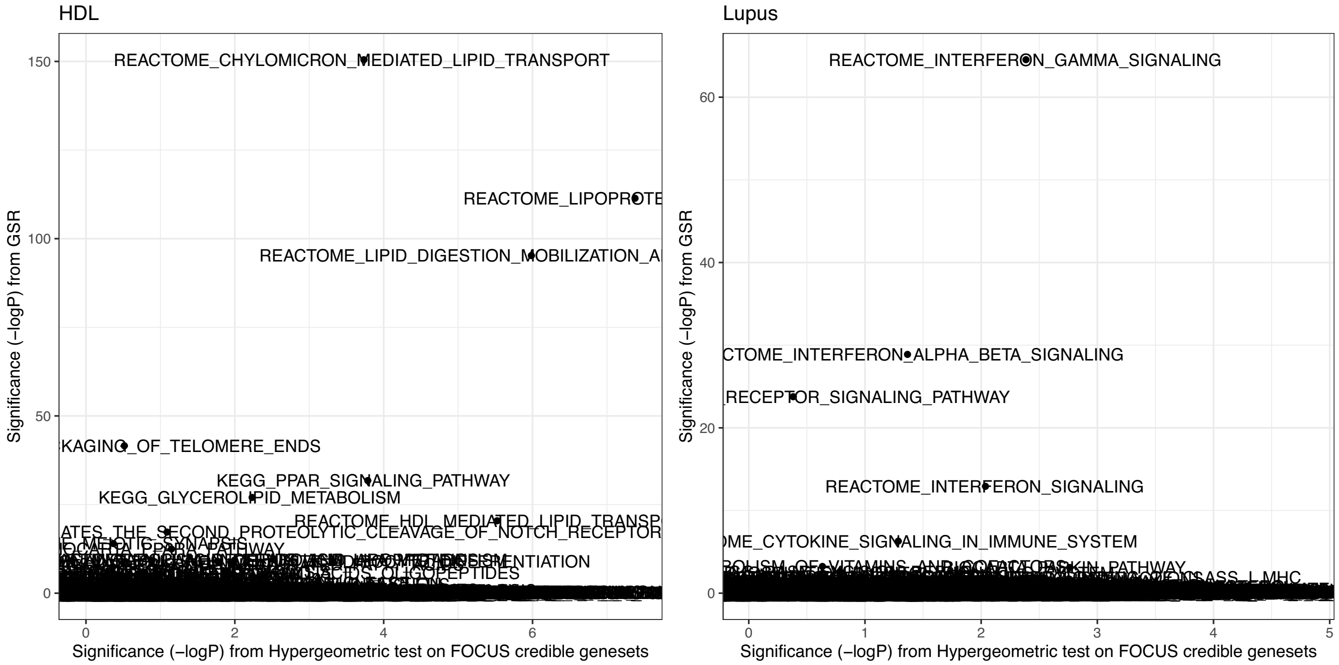

Figure 4: Comparison of gene set enrichments on HDL and Lupus traits. We compared the significance of enrichments for the MSigDb gene sets (i.e., the combined 1,050 BIOCARTA, KEGG, REACTOME gene sets) as -log10 p-values calculated by GSR and Hypergeometric test on 90% credible gene sets obtained from FOCUS.

Figure 5

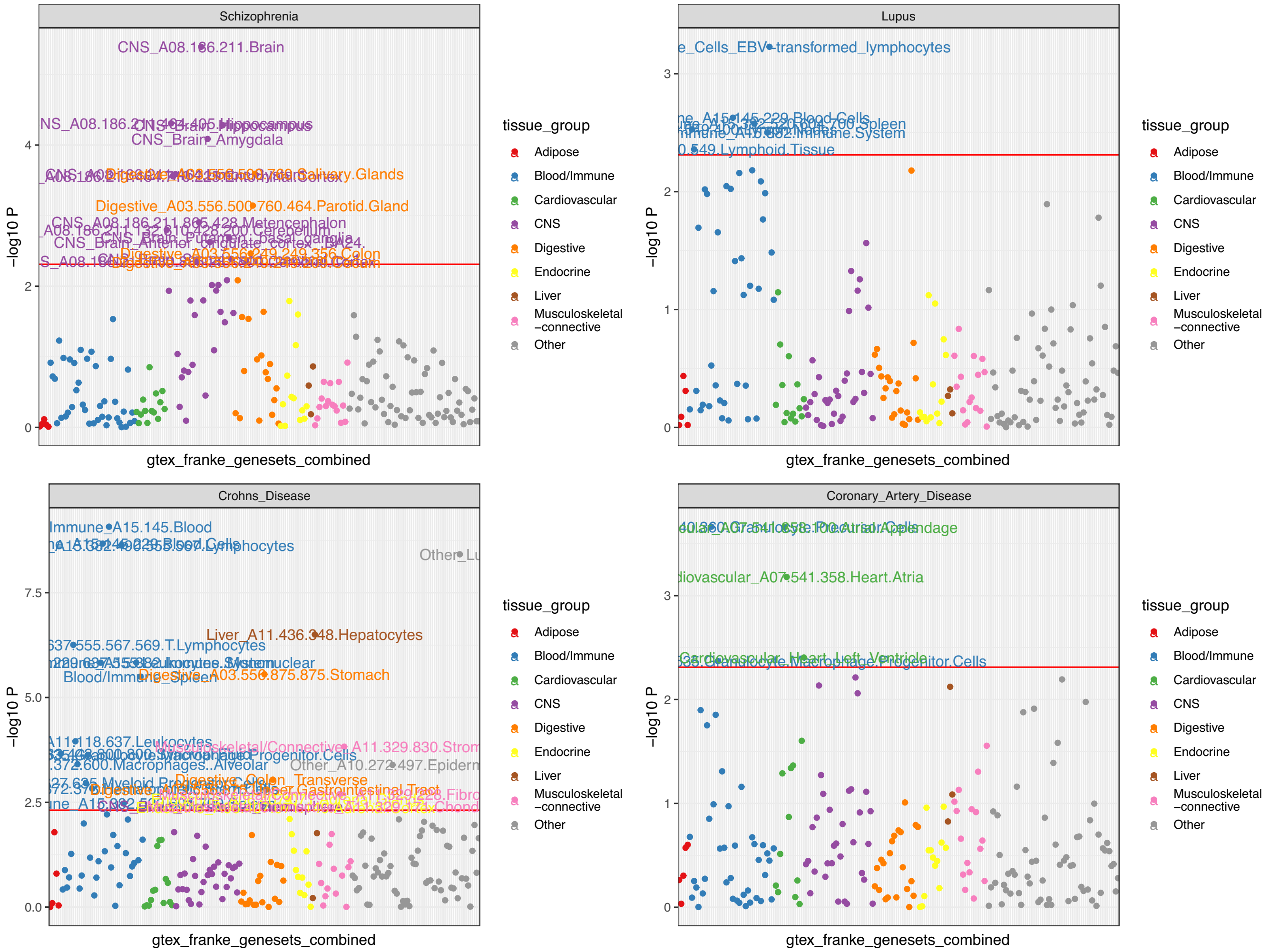

Figure 5: Enrichment of specific cell types. GSR was applied to each complex traits using 205 cell-type-specific gene sets. The red line indicate Bonferroni corrected q-value $<$ 0.05 per trait.

Figure 6

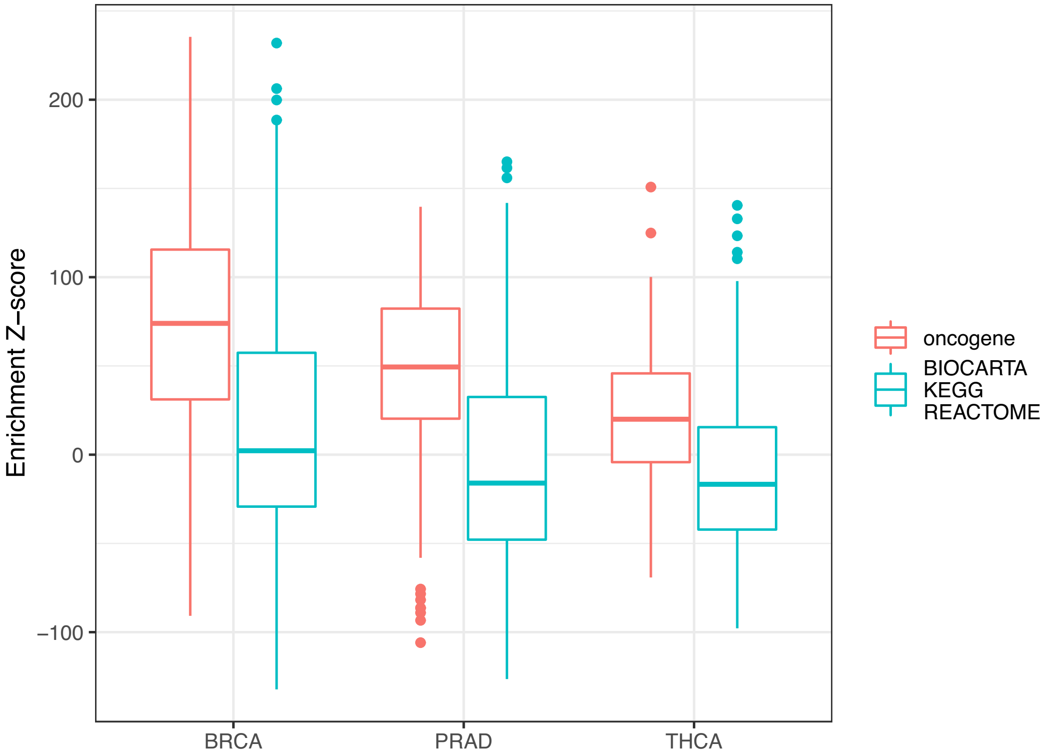

Figure 6: Gene set enrichment for TCGA tumor samples. We ran GSR on the processed gene expression data from breast cancer, thyroid cancer, and prostate tumor samples versus normal samples from TCGA and GTEx. We tested the enrichments for oncogenic signatures and the more general 1050 gene sets combining BIOCARTA, KEGG, REACTOME from MSigDB. We then separately plotted the Z-scores for oncogenic gene sets and the general gene sets across the 3 cancer types.

References

[1] Paul R Burton, David G Clayton, Lon R Cardon, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, June 2007.

[2] Peter M Visscher, Naomi R Wray, Qian Zhang, et al. 10 Years of GWAS Discovery: Biology, Function, and Translation. *AJHG*, 101(1):5–22, July 2017.

[3] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[4] Naomi R Wray, Michael E Goddard, and Peter M Visscher. Prediction of individual genetic risk of complex disease. *Current Opinion in Genetics & Development*, 18(3):257–263, June 2008.

[5] Jacqueline MacArthur, Emily Bowler, Maria Cerezo, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1):D896–D901, January 2017.

[6] Bogdan Pasaniuc and Alkes L Price. Dissecting the genetics of complex traits using summary association statistics. *Nature Publishing Group*, 18(2):117–127, November 2016.

[7] Hilary K Finucane, Brendan Bulik-Sullivan, Alexander Gusev, et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature Genetics*, 47(11):1228–1235, November 2015.

[8] Yue Li and Manolis Kellis. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Research*, July 2016.

[9] David Lamparter, Daniel Marbach, Rico Rueedi, Zoltán Kutalik, and Sven Bergmann. Fast and Rigorous Computation of Gene and Pathway Scores from SNP-Based Summary Statistics. *PLoS Computational Biology*, 12(1):e1004714–20, January 2016.

[10] GTEx Consortium, Eric R Gamazon, Heather E Wheeler, et al. A gene-based association method for mapping traits using reference transcriptome data. *Nature Genetics*, 47(9):1091–1098, August 2015.

[11] Alexander Gusev, Arthur Ko, Huwenbo Shi, et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics*, 48(3):245–252, February 2016.

[12] Michael Wainberg, Nasa Sinnott-Armstrong, Nicholas Mancuso, et al. Opportunities and challenges for transcriptome- wide association studies. *Nature Genetics*, 51(4):1–10, March 2019.

[13] Alexis Battle, Christopher D Brown, Barbara E Engelhardt, and Stephen B Montgomery. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204–213, October 2017.

[14] Nicholas Mancuso, Malika K Freund, Ruth Johnson, et al. Probabilistic fine-mapping of transcriptome-wide association studies. *Nature Genetics*, 51(4):1–12, March 2019.

[15] Hilary K Finucane, Yakir A Reshef, Verneri Anttila, et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nature Genetics*, 50(4):1–14, April 2018.

[16] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, October 2005.

[17] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526(7571):68–74, September 2015.

[18] Qingguo Wang, Joshua Armenia, Chao Zhang, et al. Data Descriptor: Unifying cancer and normal RNA sequencing data from different sources. *Scientific Data*, 5:1–8, April 2018.

[19] Tomaz Berisa and Joseph K Pickrell. Approximately independent linkage disequilibrium blocks in human populations. *Bioinformatics (Oxford, England)*, September 2015.