Inferring relevant tissues and cell types for complex traits in genome-wide association studies

Rujin Wang[1], Dan-Yu Lin[1,2], Yuchao Jiang[1,2,3,*]

[1] Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina, Chapel Hill, NC 27599, USA.

[2] Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, NC 27599, USA.

[3] Department of Genetics, School of Medicine, University of North Carolina, Chapel Hill, NC 27599, USA.

[*] To whom correspondence should be addressed: yuchaoj@email.unc.edu.

Abstract

More than a decade of genome-wide association studies (GWASs) have identified genetic risk variants that are significantly associated with complex traits. Emerging evidence suggests that the function of trait-associated variants likely acts in a tissue- or cell-type-specific fashion. Yet, it remains challenging to prioritize trait-relevant tissues or cell types to elucidate disease etiology. Here, we present EPIC (cEll tyPe enrIChment), a statistical framework that relates large-scale GWAS summary statistics to cell-type-specific omics measurements from single-cell sequencing. We derive powerful gene-level test statistics for common and rare variants, separately and jointly, and adopt generalized least squares to prioritize trait-relevant tissues or cell types while accounting for the correlation structures both within and between genes. Using enrichment of loci associated with four lipid traits in the liver and enrichment of loci associated with three neurological disorders in the brain as ground truths, we show that EPIC outperforms existing methods. We extend our framework to single-cell transcriptomic data and identify cell types underlying type 2 diabetes and schizophrenia. The enrichment is replicated using independent GWAS and single-cell datasets and further validated using PubMed search and existing bulk case-control testing results.

Keywords: genome-wide association studies, single-cell gene expression, trait-relevant tissues and cell types, risk loci, enrichment, prioritization.

Introduction

Many years of genome-wide association studies (GWASs) have yielded genetic risk variants associated with complex traits and human diseases. Emerging evidence suggests that the function of trait-associated variants likely acts in a tissue- or cell-type-specific fashion[1-5]. Recent advances in transcriptomic sequencing, including bulk RNA sequencing (RNA-seq)[6] and single-cell RNA sequencing (scRNA-seq)[7-9], enable characterization of tissue- and cell-type-specific gene expression. Combining the tissue- or cell-type-specific gene expression profiles with GWAS summary statistics provides a better understanding of genetic regulatory effects with increased resolution[10-13]. Along this line of research, recent studies have identified specific brain cell types that underlie neuropsychiatric disorders, such as schizophrenia[14] and Parkinson's disease[15], revealing that scRNA-seq data can offer finer-resolution insights that help to elucidate disease etiology.

Several methods[16-20] have been developed to integrate tissue- or cell-type-specific gene expression profiles with GWAS summary statistics to prioritize trait-relevant tissues and cell types. One set of methods, including RolyPoly[16] and LDSC-SEG[18], develops models on the single-nucleotide polymorphism (SNP) level and derives SNP-wise annotations from the transcriptomic data. RolyPoly adopts a polygenic model, and the effect sizes of all SNPs associated with a gene have a covariance that is a linear combination of the gene expressions across all tissues or cell types. RolyPoly, therefore, captures the effect of the cell-type-specific gene expression on the covariance of GWAS effect sizes. LDSC-SEG also constructs SNP annotations from tissue- or cell-type-specific gene expressions and then carries out a one-sided test using the stratified LD score regression framework[18,21-23]. It tests whether trait heritability is enriched in regions surrounding genes that have the highest specific expression in a given tissue or cell type.

Another set of methods, such as CoCoNet[19] and MAGMA[14,15,17,24], does not devise the SNP-level framework. These methods first derive gene-level association statistics since this more naturally copes with the gene-level expression measurements; they then prioritize risk genes in a specific tissue/cell type. Specifically, CoCoNet models gene-level association statistics as a function of the tissue-specific adjacency matrices inferred from gene expression studies. While CoCoNet is the first method to evaluate the gene co-expression networks, its rank-based method does not allow hypothesis testing due to the strong correlation among gene co-expression patterns constructed from different tissues and cell types. Like CoCoNet, MAGMA[17] and MAGMA-based approaches[14,15,24] also begin by combining SNP-level GWAS summary statistics into gene-level statistics. This step is followed by a second "gene-property" analysis, where the tissue- and cell-type-specific gene expressions are regressed against the genes' GWAS test statistics. The various versions of the methods adopt different ways to select genes, transform the outcome and predictor variables, and include different sets of additional covariates[14,15,24]. While MAGMA-based methods have been successfully used in several studies[25-27], Yurko et al.[28] examined the statistical foundation of MAGMA, and they identified an issue: type I error rate is inflated because the method incorrectly uses the Brown's approximation when combining the SNP-level p-values. In addition to this problem, we noticed that the MAGMA’s implementation uses squared correlations between SNPs, which masks the true LD structure.

When modeling on the gene level, one needs to account for the gene-gene correlations. RolyPoly ignores proximal gene correlations but implements a block bootstrapping procedure as a correction. MAGMA approximates the gene-gene correlations as the correlations between the model sum of squares from the second-step gene-property analysis. However, the gene-gene correlation of the effect sizes should be a function of the LD scores (i.e., the correlations between the SNPs within the genes). CoCoNet does not take account of this either, instead using LD information only to calculate the gene-level effect sizes and assuming that gene-gene covariance is a function solely of gene co-expression. A statistically rigorous and computationally efficient method to derive the gene-gene correlation structure while incorporating the SNP-level LD information is needed.

These existing methods either focus on common variants (e.g., RolyPoly and LDSC-SEG) or do not differentiate between common and rare variants (e.g., MAGMA with only summary statistics) due to the limited statistical power for rare variants. While methods for rare-variant association analysis have been developed (e.g., sequence kernel association test[29] and burden test[30]), to our best knowledge, no methods are currently available to detect tissue and/or cell-type enrichment of GWAS risk loci using summary statistics for both common and rare variants.

Here, we propose EPIC, a statistical framework to identify trait-relevant tissues or cell types by integrating tissue- or cell-type-specific gene expression profiles and GWAS summary statistics. We adopt gene-based generalized least squares to identify enrichment of risk loci. For the prioritized tissues and cell types, EPIC further carries out a gene-specific influence analysis to identify significant genes. We demonstrate EPIC on multiple tissue-specific bulk RNA-seq and scRNA-seq datasets, along with GWAS summary statistics of four lipid traits, three neuropsychiatric disorders, and type 2 diabetes, and successfully replicate and validate the prioritized tissues and cell types. Together, EPIC's integrative analysis of cell-type-specific expressions and GWAS polygenic signals help to elucidate the underlying cell-type-specific disease etiology and prioritize important functional variants. EPIC is compiled as an open-source R package available at https://github.com/rujinwang/EPIC.

Material and Methods

Overview of methods

The goal of EPIC is to identify disease- or trait-relevant tissues or cell types. An overview of the framework is outlined in Figure 1. EPIC takes as input single-variant summary statistics from GWAS, which is used to aggregate SNP-level associations into genes, and gene expression datasets from either bulk tissue or single-cell RNA-seq. An external reference panel is adopted to account for the linkage disequilibrium (LD) between SNPs and genes. We first perform gene-level testing based on GWAS summary statistics from the single-variant analysis. The multivariate statistics for both common and rare variants can be recovered using covariance of the single-variant test statistics, which can be estimated from either the participating study or from a public database. We then develop a gene-based regression framework that can prioritize trait-relevant cell types from gene-level test statistics and cell-type-specific omics profiles while accounting for gene-gene correlations due to LD. The underlying hypothesis is that if a particular cell type influences a trait, then more of the GWAS polygenic signals would be concentrated in genes with greater cell-type-specific gene expression. For significantly enriched tissue or cell type, we further carry out a gene-specific influence analysis to identify genes that are highly influential in leading to the significance of the prioritized tissue or cell type.

Gene-level associations for common variants

Let $\beta = (\beta_1, \cdots, \beta_K)^T$ be the effect sizes of K common variants within a gene of interest. Let $\hat{\beta} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_K\right)^T$ be the estimators for β, with corresponding standard errors $\hat{\sigma} = (\hat{\sigma}_1, \cdots, \hat{\sigma}_K)^T$. Let $\hat{z} = (\hat{z}_1, \cdots, \hat{z}_K)^T$ be the z-scores, where $\hat{z}_j = \hat{\beta}_j / \hat{\sigma}_j$ is the standard-normal statistic for testing the null hypothesis of no association for SNP j. We approximate the correlation matrix of $\hat{\beta}$ (equivalent to the covariance matrix of $\hat{z}$) by the LD matrix $R = \{R_{jl}; j, l = 1, \dots, K\}$, where R_{jl} is the Pearson correlation between SNP j and SNP l. We further define $V = \mathrm{cov}\left(\hat{\beta}\right) = \mathrm{diag}(\hat{\sigma}) R \mathrm{diag}(\hat{\sigma})$ as the covariance matrix of $\hat{\beta}$. We have $\hat{\beta} \sim \mathrm{MVN}(0, V)$ under the null. To perform gene-level association testing for common variants, we construct a simple and powerful chi-square statistic for testing the null hypothesis of $\beta = 0$:

$$Q^c = \hat{\beta}^T V^{-1} \hat{\beta} = \hat{z}^T R^{-1} \hat{z} \sim \chi_K^2.$$

The correlation matrix R can be estimated from either the participating study or a publicly available reference panel. In this study, we utilize the 1000 Genomes Project European panel[31], which comprises genotypes of ~500 European individuals across ~23 million SNPs.

An effective chi-square test described above requires the covariance matrix to be well-conditioned. For most GWASs, the ratio of the number of SNPs and the number of subjects is greater than or close to one, making the sample covariance matrix ill-conditioned[32,33]. In these cases, smaller eigenvalues of the sample covariance matrix are underestimated[32], leading to inflated false positives in the gene-level association testing. To solve this issue, we choose to adopt the POET estimator[34], a principal orthogonal complement thresholding approach, to obtain a well-conditioned covariance matrix via sparse shrinkage under a high-dimensional setting. The estimator of $V = \{V_{jl}; j, l = 1, \dots, K\}$ is defined as $\hat{V}_H = \sum_{j=1}^{H} \hat{\lambda}_j \hat{v}_j \hat{v}_j^T + \hat{R}_H^*$, where $\hat{\lambda}_j$ is the jth eigenvalues of the covariance matrix with corresponding eigenvector $\hat{v}_j$, $\hat{R}_H^*$ is obtained from applying adaptive thresholding on $\hat{R}_H = \sum_{j=H+1}^{K} \hat{\lambda}_j \hat{v}_j \hat{v}_j^T$, and H is the number of spiked eigenvalues.

The degree of shrinkage is determined by a tuning parameter, and we choose one so that the positive definiteness of the estimated sparse covariance matrix is guaranteed. Notably, other sparse covariance matrix estimators[32,33,35,36] can also be used in a similar fashion.

Gene-level associations for rare variants

Recent advances in next-generation sequencing technology have made it possible to extend association testing to rare variants, which can explain additional disease risk or trait variability[37-39]. Previous work[40] has demonstrated that the gene-level testing of rare variants is powerful and able to achieve well-controlled type I error as long as the correlation matrix of single-variant test statistics can be accurately estimated. Here, we recover the burden test statistics from GWAS summary statistics for the gene-level association testing of rare variants. Suppose that a total of M rare variants residing in a gene are genotyped. Let $U = \{U_j;\ j = 1, \dots, M\}$ and $C = \{C_{jl}; j, l = 1, \dots, M\}$ be the score statistic and the corresponding covariance matrix for testing the null hypothesis of no association. Under H_0, the burden test statistic $T = \xi^T U / \sqrt{\xi^T C \xi}$ follows a standard normal distribution, where $\xi_{M\times 1} = (1, \cdots, 1)^T$. We approximate U_j and C_{jl} by

$$\widehat{U}_j = w_j \hat{\beta}_j / \hat{\sigma}_j = w_j \hat{z}_j$$

$$\hat{C}_{jl} = w_j R_{jl} w_l,$$

where R is the correlation or covariance matrix of $\hat{z}$ and $w_j = 1/\hat{\sigma}_j$ is an empirical approximation to $\sqrt{C_{jj}}$. Denote $w = (w_1, \dots, w_M)^T$. The burden test uses $Q^r = (w^T \hat{z})^2 / w^T R w$, which follows a chi-square distribution with one degree of freedom under the null $Q^r \sim \chi_1^2$.

Joint analysis for common and rare variants

Existing methods either remove rare variants from the analysis[16,18] or do not differentiate common and rare variants when only summary statistics are available[17]. Yet, existing GWASs have successfully uncovered both common and rare variants associated with complex traits and diseases[15,37-39], and rare variants should therefore not be ignored in the enrichment analysis. To incorporate rare variants into the common-variant gene association testing framework, we collapse genotypes of all rare variants within a gene to construct a pseudo-SNP. We then treat the aggregated pseudo-SNP as a common variant and concatenate the z-scores $\hat{z}^* = (\hat{z}_1, \cdots, \hat{z}_K, \hat{z}^r)^T$, where the first K elements are from the common variants and $\hat{z}^r = w^T\hat{z}/\sqrt{w^TRw}$ is from the burden test statistic for the combined rare variants. A joint chi-square test for common and rare variants is performed as below:

$$Q = \hat{z}^{*T} R^{*-1} \hat{z}^* \sim \chi^2_{K+1},$$

where R^* can be estimated using POET shrinkage with the pseudo-SNP included.

Gene-gene correlation

Proximal genes that share *cis*-SNPs inherit LD from SNPs and result in correlations among genes. Since the correlations between genes are caused by LD between SNPs, which quickly drops off as a function of distance, we adopt a sliding-window approach to only compute correlations for pairs of genes within a certain distance from each. It is worth noting that this also significantly reduces the computational burden. Specifically, let N be the number of genes from the same chromosome, and we adopt a sliding window of size d to estimate the sparse covariance matrix among genes $\{G_1, \ldots, G_d\}$, $\{G_2, \ldots G_{d+1}\} G_{N-d+1}, \ldots, G_N$□, respectively. By default, we set $d = 10$ so that gene-wise correlations can be recovered for a gene with its 18 neighboring genes (see Supplementary Figure S1 for the effect of sliding window size on EPIC's performance). Similar to MAGMA, correlations are only computed for pairs of genes within 5 megabases by default.

Recall that the gene-level association statistics are chi-square statistics in a quadratic form. Within a specific window, the gene-wise correlations are obtained via transformations of the SNP-wise LD information. Let $\hat{z}^{(s)}$ and $\hat{z}^{(t)}$ be the SNP-wise z-scores for genes s and t, respectively. Let $R^{(s)} = \left\{R^{(s)}_{jl}; j, l = 1, \ldots, K_s\right\}$, $R^{(t)} = \left\{R^{(t)}_{jl}; j, l = 1, \ldots, K_t\right\}$, and $R^{(s,t)} = \left\{R^{(s,t)}_{jl}; j = 1, \ldots, K_s, l = 1, \ldots, K_t\right\} = \mathrm{cor}\left(\hat{z}^{(s)}, \hat{z}^{(t)}\right)$ be the within- and between-gene correlation matrices obtained from the POET shrinkage estimation. We take advantage of the Cholesky decomposition to obtain the gene-gene correlation between $Q_s = \left(\hat{z}^{(s)}\right)^T \left(R^{(s)}\right)^{-1} \hat{z}^{(s)}$ and $Q_t = \left(\hat{z}^{(t)}\right)^T \left(R^{(t)}\right)^{-1} \hat{z}^{(t)}$:

$$\rho_{st} = \mathrm{cor}(Q_s, Q_t) = \frac{\sum_{j=1}^{K_s} \sum_{i=1}^{K_s+K_t} L_{ij}^2}{\sqrt{K_s K_t}},$$

where L_{ij}'s are entries of a lower triangular matrix L such that $\tilde{R}_{(K_s+K_t)\times(K_s+K_t)} = LL^T$ and

$$\tilde{R}_{(K_s+K_t)\times(K_s+K_t)} = \begin{pmatrix} I_{K_s} & {R^{(s)}}^{-1/2} R^{(s,t)} {R^{(t)}}^{-1/2} \\ {R^{(t)}}^{-1/2} R^{(t,s)} {R^{(s)}}^{-1/2} & I_{K_t} \end{pmatrix},$$

I_K is the identity matrix with dimension K. The full derivation is detailed in Supplementary Note S1. When rare variants are included in the framework, gene-gene correlations are calculated similarly by aggregating all rare variants that reside in a gene as a pseudo-SNP.

Prioritizing trait-relevant tissue(s) and cell type(s)

To detect tissue- or cell-type-specific enrichment for a specific trait of interest, we devise a regression framework based on generalized least squares to identify risk loci enrichment. The key underlying hypothesis is that if a particular cell type influences a trait, more GWAS polygenic signals would be concentrated in genes with greater cell-type-specific gene expression. Under this hypothesis, genes that are significantly associated with lipid traits are expected to be highly expressed in the liver since the liver is known to participate in cholesterol regulation. This relationship between the GWAS association signals and the gene expression specificity is modeled as below.

Let Q_g be the gene-level chi-square association test statistic for gene g. To account for the different number of SNPs within each gene, we adjust the degree of freedom of $K_g + 1$ to obtain $Y_g = Q_g/(K_g + 1)$, which is included as the outcome variable. For each cell type c, to test for its enrichment we fit a separate regression using its cell-type-specific gene expression E_{cg} (reads per kilobase million (RPKM) or transcripts per million (TPM)) as a dependent variable. To account for the baseline gene expression[24], we also include another covariate $A_g = \frac{1}{T}\sum_{c=1}^{T} E_{cg}$, which is the average gene expression across all T tissues/cell types. Taken together, we have

$$Y = \gamma_0 + E_c\gamma_c + A\gamma_A + \epsilon,$$

where $\epsilon \sim \mathrm{MVN}(0, \sigma^2 W)$, $W = DPD^T$, $D = \mathrm{Diag}\left(\sqrt{2/K_g}\right)$, and $P = \{\rho_{st}\}$ is the gene-gene correlation matrix. We adopt the generalized least squares approach to fit the model and perform a one-sided test against the alternative $\gamma_c > 0$, under which the gene-level association signals positively correlated with the cell-type-specific expression. For a significantly enriched tissue or cell type, we further carry out a statistical influence test to identify a set of tissue- or cell-type-specific influential genes, using the DFBETAS statistics[41]—large values of DFBETAS indicate observations (i.e., genes) that are influential in estimating γ_c. With a size-adjusted cutoff $2/\sqrt{N}$, where N is the number of genes used in the tissue- or cell-type-specific enrichment analysis, significantly influential genes allow for further pathway or gene set enrichment analyses.

GWAS summary statistics and transcriptomic data processing

We adopt GWAS summary statistics of eight traits, including four lipid traits[42] (low-density lipoprotein cholesterol (LDL), high-density lipoprotein cholesterol (HDL), total cholesterol (TC), and triglyceride levels (TG)), three neuropsychiatric disorders[39,43,44] (schizophrenia (SCZ), bipolar disorder (BIP), and schizophrenia and bipolar disorder (SCZBIP)), and type 2 diabetes[38] (T2D). The relevant tissues involved in these traits are well known/studied – liver for the lipid traits, brain for the neuropsychiatric disorders, and pancreas for the T2D – and we use this as ground truths to demonstrate EPIC and to benchmark against other methods. See Supplementary Table S1 for more information on the GWASs.

For each trait, we obtain SNP-level summary statistics and apply stringent quality control procedures to the data. We restrict our analyses to autosomes, filter out SNPs not in the 1000 Genomes Project Phase 3 reference panel, and remove SNPs with mismatched reference SNP ID numbers. We exclude SNPs from the major histocompatibility complex (MHC) region due to complex LD architecture[16,19,22]. In addition to SNP filtering, we align alleles of each SNP against those of the reference panel to harmonize the effect alleles of all processed GWAS summary statistics. A gene window is defined with 10kb upstream and 1.5kb downstream of each gene[14], and SNPs residing in the windows are assigned to the corresponding genes.

In the analysis that follows, we uniformly report results using a minor allele frequency (MAF) cutoff of 1% to define common and rare variants (see Supplementary Figure S2 for enrichment results with different MAF cutoffs). To reduce the computational cost and to alleviate the multicollinearity problem, we perform LD pruning using PLINK[45] with a threshold of $r^2 \leq 0.8$ to obtain a set of pruned-in common variants, followed by a second-round of LD pruning if the number of common SNPs per gene exceeds 200. See Supplementary Figure S3 for results with varying LD-pruning thresholds. For rare variants, we only carry out a gene-level rare variant association testing if the minor allele count (MAC), defined as the total number of minor alleles across subjects and SNPs within the gene, exceeds 20. We report the number of SNPs (common variants and rare variants), the number of genes, and the number of SNPs per gene for each GWAS trait in Supplementary Table S2.

We adopt a unified framework to process all transcriptomic data. For scRNA-seq data, we follow the Seurat[46] pipeline to perform gene- and cell-wise quality controls and focus on the top 8000 highly variable genes. Cell-type-specific RPKMs are calculated by combining read or UMI counts from all cells of a specific cell type, followed by log2 transformation with an added pseudo-count. For tissue-specific bulk RNA-seq data from the Genotype-Tissue Expression project (GTEx), we first calculate a tissue specificity score for each gene[19,47], and only focus on genes that are highly specific in at least one tissue. See Supplementary Note S2 for more details. We then perform log2 transformation on the tissue-specific TPM measurements with an added pseudo-count.

Benchmarking against RolyPoly, LDSC-SEG, and MAGMA

We benchmarked EPIC against three existing approaches: RolyPoly[16], LDSC-SEG[18], and MAGMA[17]. For all methods, we used RPKMs for each cell type and TPMs for each GTEx tissue in the benchmarking analysis. We made gene annotations the same for RolyPoly, MAGMA, and EPIC by defining the gene window as 10kb upstream and 1.5kb downstream of each gene. For LDSC-SEG, as recommended by the authors[18], the window size is set to be 100kb up and downstream of each gene's transcribed region. Since all methods adopt a hypothesis testing framework to identify trait-relevant tissue(s), for each trait-tissue pair, we reported and compared the corresponding p-values from the different methods.

RolyPoly takes as input GWAS summary statistics, gene expression data, gene annotations, and LD matrix from the 1000 Genomes Project Phase 3. As recommended by the developer for RolyPoly[16], we scaled the gene expression for each gene across tissues/cell types and took the absolute values of the scaled expression values. We performed 100 block bootstrapping iterations to test whether a tissue- or cell-type-specific gene expression annotation was significantly enriched in a joint model across all tissues or cell types. We also benchmarked LDSC-SEG, which computes *t*-statistics to quantify differential expression for each gene across tissues or cell types. We annotated genome-wide SNPs using the top 10% genes with the highest positive *t*-statistics and applied stratified LDSC to test the heritability enrichment of the annotations that were attributed to specifically expressed genes for each tissue. For MAGMA, we first obtained gene-level association statistics using MAGMA v1.08. We then carried out the gene-property analysis proposed in Watanabe et al.[24], with technical confounders being controlled by default, to test the positive relationship between tissue- or cell-type specificity of gene expression and genetic associations.

Results

Inferring trait-relevant tissues using bulk RNA-seq from GTEx

We started our analysis with tissue-specific transcriptomic profiles from the GTEx v8[6], which consists of bulk-tissue gene expression measurements of 17,382 samples from 54 tissues across 980 postmortem donors (Supplementary Table S1). Tissues with fewer than 100 samples were removed from the analysis. After sample-specific quality controls, we obtained gene expression profiles of 45 tissues, averaged across samples. For subsequent analyses, we focused on a set of 8,708 genes with tissue specificity scores greater than 5. We applied EPIC to the GTEx data with GWAS summary statistics for eight diseases and traits, including four lipid traits, three neuropsychiatric disorders, and T2D.

We first performed the gene-level chi-square association test with the shrinkage estimators and sliding-window approach. The quantile-quantile (Q-Q) plots of gene-level p-values are shown in Supplementary Figure S4, with a comparison against MAGMA. We observed elevated power in the Q-Q plots for four lipid traits. In Supplementary Table S3, we summarized a list of genes that have been shown to modulate lipid levels[42] and compared the gene-level association testing results from EPIC and MAGMA. Significant gene-level associations were detected between all lipid traits and variants in *APOB*, *APOE*, and *CETP*. Meanwhile, *PCSK9*, *ABCG5,* and *ABCG8* exhibited significant associations with LDL and TC. For neuropsychiatric disorders, we examined genes that are relevant to the etiology of schizophrenia[39], including genes that are targets of therapeutic drugs (*DRD2* and *GRM3*), genes that participate in neuronal calcium signaling (*CACNA1I*), and genes that are involved in synaptic function (*CNTN4* and *SNAP91*) and other neuronal pathways (*FXR1*, *CHRNA3*, *CHRNB4*, and *HCN1*). EPIC's chi-square test approach demonstrates higher power than MAGMA. We also compared the number of significant genes for eight traits – after Bonferroni correction, EPIC detected more genes than MAGMA (Supplementary Figure S5). Additionally, we report gene-level association tests for a set of housekeeping genes[48] and demonstrate that, while powerful, EPIC also controls for type I error (Supplementary Figure S6).

We next applied EPIC to identify the trait-relevant tissues by performing tissue-specific regression for each trait, with results shown in Figure 2, Figure 3A, and Figure 4A. All four lipid traits are significantly enriched in the liver, which plays a key role in lipid metabolism. Specifically, LDL, TC, and TG showed strong enrichment in the liver (Figure 2A, Figure 2C, and Figure 2D), suggesting that these three traits are embedded in a similar genetic architecture and share the same relevant tissue. The small intestine was marginally significant for TC – it has been shown that the small intestine plays an important role in cholesterol regulation and metabolism[49-51]. On the other hand, HDL exhibited a slightly different enrichment pattern (Figure 2B): liver and two adipose tissues are identified as being significantly enriched by both EPIC and MAGMA. Both LDSC-SEG and RolyPoly suffer from low power, although the liver was one of the top-ranked tissues for the lipid traits.

Neuropsychiatric disorders exhibited strong brain-specific enrichments, as expected. The frontal cortex of the brain was detected as being the most strongly enriched for SCZ, BIP, and SCZBIP (Figure 4A). The pituitary also demonstrated strong enrichment signals with SCZ and SCZBIP, while the spinal cord was found to be an irrelevant tissue with these three neuropsychiatric disorders. In comparison, LDSC-SEG identified part of the brain tissues as trait-relevant, while RolyPoly failed to return enrichment in any of the brain tissues (Figure 4A).

As a final proof of concept, we sought to infer T2D-relevant tissue(s) using the tissue-specific gene expression data GTEx. The pancreas and the liver were prioritized as the T2D-relevant tissues by EPIC, while MAGMA yielded significant results in the pancreas as well as the stomach (Figure 3A). RolyPoly identified the pancreas as the second most relevant tissue; LDSC-SEG reported liver as the only significantly enriched tissue (Figure 3A). For validation, we adopted a similar strategy as proposed by Shang et al.[19] – we carried out a PubMed search, resorting to previous literatures studying the trait of interest in relation to a particular tissue or cell type. Specifically, we counted the number of previous publications using the key word pairs of trait and tissue/cell type and calculated the Spearman's rank correlations between the number of publications and EPIC's tissue-/cell-type-specific p-values (Figure 5). Across all traits, we found strong positive correlations between EPIC's enrichment results and PubMed search results (Figure 5A).

Cell-type enrichment for T2D by scRNA-seq data of pancreatic islets

We next analyzed pancreatic islet scRNA-seq data to identify trait-relevant cell types for T2D. To assess reproducibility, EPIC was separately applied to two scRNA-seq datasets consisting of multiple endocrine cell types (Supplementary Table S1 and Supplementary Figure S7). The scRNA-seq data were generated using two different protocols: the Smart-seq2 protocol on six healthy donors from Segerstolpe et al.[9] and the InDrop protocol on three healthy individuals from Baron et al.[7]. Following the pre-processing step as described in Materials and Methods, we retained a total of 5,488 genes to prioritize pancreatic cell types for T2D. In both datasets, beta cells were identified as the trait-relevant cell types by EPIC (Figure 3B). This finding was supported by known biology, in that beta cells participate in insulin secretion and are gradually lost in T2D[52-54]. We also found that gamma cells were marginally associated with T2D in the Segerstolpe dataset. Pancreatic polypeptide, which is produced by gamma cells, is known to play a critical role in endocrine pancreatic secretion regulation[55-57]. However, neither MAGMA nor LDSC-SEG detected significant enrichment in beta cells, even though the enrichment was top-ranked. RolyPoly, on the other hand, did not report any enrichment of the beta cells compared to the other types of cells.

To identify specific genes that drive the significant enrichment in beta cells, we carried out the gene-specific influence test as outlined in Materials and Methods and identified 142 highly influential genes (Figure 3C). We then performed KEGG pathway analysis and Gene Ontology (GO) biological process enrichment analysis using the DAVID bioinformatics resources[58,59]. Beta-cell-specific influential genes are enriched in GO terms including glucose homeostasis and regulation of insulin secretion, as well as KEGG pathways including insulin secretion, maturity-onset diabetes of the young, and type II diabetes mellitus (Figure 3C and Supplementary Table S4). Additionally, the cell-type ranks obtained from EPIC's beta-cell-specific p-values was highly consistent with those from the PubMed search results (Figure 5B). We demonstrate the effectiveness of EPIC in identifying tissue-relevant cell types using scRNA-seq datasets generated by different protocols.

Cell-type enrichment for neuropsychiatric disorders by scRNA-seq data of brain

To further test EPIC in a more complex tissue, we sought to prioritize trait-relevant cell types in the brain. While the brain tissues are significantly enriched using the GTEx bulk-tissue RNA-seq data (Figure 4A), the relevant cell types in the brain for neuropsychiatric disorders are not as well defined and studied. We obtained droplet-based scRNA-seq data[8], generated on frozen adult human postmortem tissues from the GTEx project (Supplementary Table S1), to infer the relevant cell types. After pre-processing and stringent quality controls, the scRNA-seq data contains gene expression profiles of 17,698 genes across 14,137 single cells collected from the human hippocampus and prefrontal cortex tissues. The cells belong to ten cell types (Figure 4B), and we focused on the top 8,000 highly variable genes for subsequent analyses.

We evaluated EPIC's cell-type-specific enrichment results and found that all three neuropsychiatric disorders were significantly enriched in GABAergic interneurons (GABA), excitatory glutamatergic neurons from the prefrontal cortex (exPFC), and excitatory pyramidal neurons in the hippocampal CA region (exCA). Excitatory granule neurons from the hippocampal dentate gyrus region (exDG) were identified as relevant cell types for SCZ and SCZBIP (Figure 4C). EPIC successfully replicated the previously reported association of neuropsychiatric disorders with interneurons and excitatory pyramidal neurons[14,15].

We employed three strategies to validate the trait-relevant cell types for the neuropsychiatric disorders. First, we again found positive Spearman correlations with PubMed search results and EPIC's enrichment results for SCZ and SCZBIP (Figure 5C). Second, we adopted additional independent GWAS summary statistics for SCZ (SCZ2)[60] (Supplementary Table S1) and observed highly concordant enrichment results between SCZ and SCZ2 (Figure 4C). Third, we tested whether genes that are upregulated/downregulated for SCZ were enriched in the identified cell types to additionally implicate cell types involved in SCZ. Specifically, we performed differential expression (DE) analysis from an independent case-control study of SCZ using bulk RNA-seq[61], retaining 287 significant DE genes that also overlap the scRNA-seq data (Supplementary Figure S8). We reasoned that, if SCZ-relevant risk loci were enriched in a particular cell type, genes that are differentially expressed between SCZ cases and controls would demonstrate greater cell-type specificity in this cell type. We calculated cell-type specificities using the set of DE genes and observed GABA, exCA, exDG, and exPFC were the top four cell types with the lowest gene-specificity ranks (Figure 4D). Using three different strategies by querying external databases and adopting additional and orthogonal datasets, we validated the trait-cell-type relevance results.

Discussion

Over the last one and half decades, GWASs have successfully identified and replicated genetic variants associated with various complex traits. Meanwhile, bulk-tissue and single-cell transcriptomic sequencing allow tissue- and cell-type-specific gene expression characterization and have seen rapid technological development with ever-increasing sequencing capacities and throughputs. Here, we propose EPIC to address the problem of how GWAS summary statistics should be integrated with bulk-tissue or single-cell transcriptomic data to prioritize trait-relevant tissue or cell types and to elucidate disease etiology. To our best knowledge, EPIC is the first method that prioritizes tissues and/or cell types for both common and rare variants with a rigorous statistical framework to account for both within- and between-gene correlations. We demonstrate EPIC's effectiveness and outperformance compared to existing methods with extensive benchmark and validation studies.

For scRNA-seq data, all existing methods, including EPIC, resort to pre-clustered/annotated cell types and average across cells to obtain cell-type-specific expression profiles. However, scRNA-seq goes beyond the mean measurements[62,63], and how to make the best use of gene expression dispersion, nonzero fraction, and other aspects of its distribution needs further method development[64]. Additionally, while many efforts have been devoted to identifying enrichment of discretized cell types, how to carry out enrichment analysis for transient cell states needs further investigation. Last but not least, when multiple scRNA-seq datasets are available across different experiments, protocols, or species, borrowing information from additional sources can potentially boost the performance and increase the robustness of the enrichment analysis[65]. While it is nontrivial to directly perform gene expression data integration, a cross-dataset conditional analysis workflow was proposed by Watanabe et al.[24] to evaluate the association of cell types based on multiple independent scRNA-seq datasets. However, the linear conditional analysis may not be sufficient to capture the nonlinear batch effects[46,66].

It is also worth noting that CoCoNet, MAGMA, and EPIC first carry out a gene-level association test, so that the summary statistics and expressions are unified to be gene-specific. They adopt different methods to integrate SNP-wise summary statistics, and SNPs need to be annotated to genes based on a window surrounding each gene. While RolyPoly and LDSC-SEG model on the SNP level directly, each SNP still needs to be assigned to a gene so that the gene expression can be used as a SNP annotation. There is not a consensus on how to most accurately assign SNPs to genes, and more importantly, one would only be able to do so for SNPs that reside in gene bodies or promoter regions. Meanwhile, a large number of GWAS hits are in the non-coding regions, and their functions are yet to be fully understood. EPIC's framework can be easily extended to infer enrichment of non-coding variants when combined with the single-cell assay for transposase-accessible chromatin using sequencing (ATAC-seq) data[67,68]. Additionally, cell-type-specific expression quantitative trait loci from the non-coding regions[69] can also be integrated with the second-step gene-property analysis to boost power and to infer enrichment of non-coding variants.

Data Availability

GWAS summary statistics are downloaded from public repositories listed in Supplementary Table S1. Genotypes from the 1000 Genomes Project reference panel are available at https://ctg.cncr.nl/software/magma. Bulk RNA-seq and scRNA-seq data are downloaded from GTEx v8 at http://www.gtexportal.org/. ScRNA-seq read counts from two pancreatic islet studies are publicly available with accession GSE81433[7] and E-MTAB-5061[9]. We obtain a list of human housekeeping genes from the Housekeeping and Reference Transcript Atlas[48] at https://housekeeping.unicamp.br/.

Code Availability

EPIC is compiled as an open-source R package available at https://github.com/rujinwang/EPIC.

Acknowledgments

This work was supported by the National Institutes of Health (NIH) grant P01 CA142538 (to D.L. and Y.J.), R35 GM138342 (to Y.J.), and R01 HG009974 (to D.L.). The authors thank Drs. Yun Li, Michael Love, Karen Mohlke, and Jason Stein for helpful discussions and comments, and Drs. Alkes Price, Diego Calderon, and Kyoko Watanabe for providing support and insight on existing methods.

Authors' Contributions

Y.J. and D.L. initiated and envisioned the study. R.W., Y.J., and D.L. formulated the model; R.W. developed and implemented the algorithm. R.W., D.L., and Y.J. performed data analysis. R.W. and Y.J. wrote the manuscript, which was edited by D.L..

Competing Interests

The authors declare no competing interests.

Figure Legends

Figure 1. Overview of EPIC framework. EPIC starts from GWAS summary statistics and an external reference panel to account for LD structure. To ensure that the correlation matrix is well-conditioned, EPIC adopts the POET estimators to obtain a sparse shrinkage correlation matrix. EPIC performs LD pruning, computes the gene-level chi-square statistics for common variants, and calculates burden test statistics for rare variants. EPIC then integrates gene-level association statistics with transcriptomic profiles and prioritizes trait-relevant tissues or cell types using a regression-based framework while accounting for the gene-gene correlation structure.

Figure 2. Tissue enrichment for four lipid traits using GTEx bulk RNA-seq data. (A) LDL; (B) HDL; (C) TC; and (D) TG. The dashed line is the Bonferroni-corrected *p*-value threshold.

Figure 3. Tissue and cell-type enrichment of T2D risk loci. (A) T2D-relevant tissue identification using GTEx tissue-specific RNA-seq data. (B) T2D-relevant cell type identification using scRNA-seq data of human pancreatic islets. The dashed line is the Bonferroni-corrected *p*-value threshold. (C) Gene-specific influence analysis for the significantly enriched beta cells using scRNA-seq data of human pancreatic islets from Baron et al.. DFBETA measures the difference in the estimated coefficients in the gene-property analysis with and without each gene. Red lines are the size-adjusted cutoffs $\pm 2/\sqrt{N} \approx \pm 0.03$, where N is the number of genes.

Figure 4. Tissue and cell-type enrichment for three neuropsychiatric disorders. (A) Beeswarm plot of –log10(*p*-value) from the tissue enrichment analysis using GTEx bulk RNA-seq data. The dashed line is Bonferroni corrected *p*-value threshold (0.05/45). (B) Heatmap of –log10(*p*-value) from the cell-type enrichment analysis using GTEx scRNA-seq brain data. Bonferroni-significant results are marked with red asterisks (p<0.05/10). GABA: GABAergic interneurons; exPFC: excitatory glutamatergic neurons in the prefrontal cortex; exDG: excitatory granule neurons from the hippocampal dentate gyrus region; exCA: excitatory pyramidal neurons in the hippocampal Cornu Ammonis region;

OPC: oligodendrocyte precursor cells; ODC: oligodendrocytes; NSC: neuronal stem cells; ASC: astrocytes; MG: microglia cells; END: endothelial cells. (C) Boxplots of gene specificity ranks across ten brain cell types for differentially expressed genes from SCZ case-control studies.

Figure 5. Correlations of tissue or cell type ranks from enrichment analysis and PubMed Search. Spearman correlations are calculated between the PubMed search and EPIC's results. Trait-relevant tissues/cell types with statistical significance after Bonferroni correction are highlighted in red, where the top-ranking tissues/cell types are labeled.

References

1. Lang, U.E., Puls, I., Muller, D.J., Strutz-Seebohm, N., and Gallinat, J. (2007). Molecular mechanisms of schizophrenia. Cell Physiol Biochem 20, 687-702.
2. Ongen, H., Brown, A.A., Delaneau, O., Panousis, N.I., Nica, A.C., Consortium, G.T., and Dermitzakis, E.T. (2017). Estimating the causal tissues for complex traits and diseases. Nat Genet 49, 1676-1683.
3. Raj, T., Rothamel, K., Mostafavi, S., Ye, C., Lee, M.N., Replogle, J.M., Feng, T., Lee, M., Asinovski, N., Frohlich, I., et al. (2014). Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. Science 344, 519-523.
4. Uhlhaas, P.J., and Singer, W. (2010). Abnormal neural oscillations and synchrony in schizophrenia. Nat Rev Neurosci 11, 100-113.
5. Xiao, X., Chang, H., and Li, M. (2017). Molecular mechanisms underlying noncoding risk variations in psychiatric genetic studies. Mol Psychiatry 22, 497-511.
6. Consortium, G.T. (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science 369, 1318-1330.
7. Baron, M., Veres, A., Wolock, S.L., Faust, A.L., Gaujoux, R., Vetere, A., Ryu, J.H., Wagner, B.K., Shen-Orr, S.S., Klein, A.M., et al. (2016). A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure. Cell Syst 3, 346-360 e344.
8. Habib, N., Avraham-Davidi, I., Basu, A., Burks, T., Shekhar, K., Hofree, M., Choudhury, S.R., Aguet, F., Gelfand, E., Ardlie, K., et al. (2017). Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods 14, 955-958.
9. Segerstolpe, A., Palasantza, A., Eliasson, P., Andersson, E.M., Andreasson, A.C., Sun, X., Picelli, S., Sabirsh, A., Clausen, M., Bjursell, M.K., et al. (2016). Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes. Cell Metab 24, 593-607.
10. Gormley, P., Anttila, V., Winsvold, B.S., Palta, P., Esko, T., Pers, T.H., Farh, K.H., Cuenca-Leon, E., Muona, M., Furlotte, N.A., et al. (2016). Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. Nat Genet 48, 856-866.
11. Hu, X., Kim, H., Stahl, E., Plenge, R., Daly, M., and Raychaudhuri, S. (2011). Integrating autoimmune risk loci with gene-expression data identifies specific pathogenic immune cell subsets. Am J Hum Genet 89, 496-506.
12. Pers, T.H., Karjalainen, J.M., Chan, Y., Westra, H.J., Wood, A.R., Yang, J., Lui, J.C., Vedantam, S., Gustafsson, S., Esko, T., et al. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. Nat Commun 6, 5890.
13. Slowikowski, K., Hu, X., and Raychaudhuri, S. (2014). SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. Bioinformatics 30, 2496-2497.
14. Skene, N.G., Bryois, J., Bakken, T.E., Breen, G., Crowley, J.J., Gaspar, H.A., Giusti-Rodriguez, P., Hodge, R.D., Miller, J.A., Munoz-Manchado, A.B., et al. (2018). Genetic identification of brain cell types underlying schizophrenia. Nat Genet 50, 825-833.
15. Bryois, J., Skene, N.G., Hansen, T.F., Kogelman, L.J.A., Watson, H.J., Liu, Z., Eating Disorders Working Group of the Psychiatric Genomics, C., International Headache Genetics, C., andMe Research, T., Brueggeman, L., et al. (2020). Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. Nat Genet 52, 482-493.
16. Calderon, D., Bhaskar, A., Knowles, D.A., Golan, D., Raj, T., Fu, A.Q., and Pritchard, J.K. (2017). Inferring Relevant Cell Types for Complex Traits by Using Single-Cell Gene Expression. Am J Hum Genet 101, 686-699.
17. de Leeuw, C.A., Mooij, J.M., Heskes, T., and Posthuma, D. (2015). MAGMA: generalized gene-set analysis of GWAS data. PLoS Comput Biol 11, e1004219.

18. Finucane, H.K., Reshef, Y.A., Anttila, V., Slowikowski, K., Gusev, A., Byrnes, A., Gazal, S., Loh, P.R., Lareau, C., Shoresh, N., et al. (2018). Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. Nat Genet 50, 621-629.
19. Shang, L., Smith, J.A., and Zhou, X. (2020). Leveraging gene co-expression patterns to infer trait-relevant tissues in genome-wide association studies. PLoS Genet 16, e1008734.
20. Zhu, H., Shang, L., and Zhou, X. (2020). A Review of Statistical Methods for Identifying Trait-Relevant Tissues and Cell Types. Front Genet 11, 587887.
21. Bulik-Sullivan, B.K., Loh, P.R., Finucane, H.K., Ripke, S., Yang, J., Schizophrenia Working Group of the Psychiatric Genomics, C., Patterson, N., Daly, M.J., Price, A.L., and Neale, B.M. (2015). LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47, 291-295.
22. Finucane, H.K., Bulik-Sullivan, B., Gusev, A., Trynka, G., Reshef, Y., Loh, P.R., Anttila, V., Xu, H., Zang, C., Farh, K., et al. (2015). Partitioning heritability by functional annotation using genome-wide association summary statistics. Nat Genet 47, 1228-1235.
23. Jagadeesh, K.A., Dey, K.K., Montoro, D.T., Gazal, S., Engreitz, J.M., Xavier, R.J., Price, A.L., and Regev, A. (2021). Identifying disease-critical cell types and cellular processes across the human body by integration of single-cell profiles and human genetics. 2021.2003.2019.436212.
24. Watanabe, K., Umicevic Mirkov, M., de Leeuw, C.A., van den Heuvel, M.P., and Posthuma, D. (2019). Genetic mapping of cell type specificity for complex traits. Nat Commun 10, 3222.
25. Kalra, G., Milon, B., Casella, A.M., Herb, B.R., Humphries, E., Song, Y., Rose, K.P., Hertzano, R., and Ament, S.A. (2020). Biological insights from multi-omic analysis of 31 genomic risk loci for adult hearing difficulty. PLoS Genet 16, e1009025.
26. Timshel, P.N., Thompson, J.J., and Pers, T.H. (2020). Genetic mapping of etiologic brain cell types for obesity. Elife 9.
27. Tran, M.N., Maynard, K.R., Spangler, A., Collado-Torres, L., Sadashivaiah, V., Tippani, M., Barry, B.K., Hancock, D.B., Hicks, S.C., Kleinman, J.E., et al. (2020). Single-nucleus transcriptome analysis reveals cell type-specific molecular signatures across reward circuitry in the human brain. bioRxiv, 2020.2010.2007.329839.
28. Yurko, R., Roeder, K., Devlin, B., and G'Sell, M. (2021). H-MAGMA, inheriting a shaky statistical foundation, yields excess false positives. Ann Hum Genet 85, 97-100.
29. Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89, 82-93.
30. Lin, D.Y., and Tang, Z.Z. (2011). A general framework for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89, 354-367.
31. Genomes Project, C., Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature 491, 56-65.
32. Ledoit, O., and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. Journal of Multivariate Analysis 88, 365-411.
33. Cai, T., and Liu, W. (2011). Adaptive Thresholding for Sparse Covariance Matrix Estimation. Journal of the American Statistical Association 106, 672-684.
34. Fan, J., Liao, Y., and Mincheva, M. (2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements. J R Stat Soc Series B Stat Methodol 75.
35. Bickel, P.J., and Levina, E. (2008). Covariance Regularization by Thresholding. Ann Stat 36, 2577-2604.
36. Ledoit, O., and Wolf, M. (2015). Spectrum estimation: A unified framework for covariance matrix estimation and PCA in large dimensions. Journal of Multivariate Analysis 139, 360-384.

37. Lange, L.A., Hu, Y., Zhang, H., Xue, C., Schmidt, E.M., Tang, Z.Z., Bizon, C., Lange, E.M., Smith, J.D., Turner, E.H., et al. (2014). Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. Am J Hum Genet 94, 233-245.

38. Mahajan, A., Taliun, D., Thurner, M., Robertson, N.R., Torres, J.M., Rayner, N.W., Payne, A.J., Steinthorsdottir, V., Scott, R.A., Grarup, N., et al. (2018). Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. Nat Genet 50, 1505-1513.

39. Schizophrenia Working Group of the Psychiatric Genomics, C. (2014). Biological insights from 108 schizophrenia-associated genetic loci. Nature 511, 421-427.

40. Hu, Y.J., Berndt, S.I., Gustafsson, S., Ganna, A., Genetic Investigation of, A.T.C., Hirschhorn, J., North, K.E., Ingelsson, E., and Lin, D.Y. (2013). Meta-analysis of gene-level associations for rare variants based on single-variant statistics. Am J Hum Genet 93, 236-248.

41. Belsley, D.A., Kuh, E., and Welsch, R.E. (1980). Regression diagnostics : identifying influential data and sources of collinearity.(New York: Wiley).

42. Willer, C.J., Schmidt, E.M., Sengupta, S., Peloso, G.M., Gustafsson, S., Kanoni, S., Ganna, A., Chen, J., Buchkovich, M.L., Mora, S., et al. (2013). Discovery and refinement of loci associated with lipid levels. Nat Genet 45, 1274-1283.

43. Bipolar, D., Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address, d.r.v.e., Bipolar, D., and Schizophrenia Working Group of the Psychiatric Genomics, C. (2018). Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. Cell 173, 1705-1715 e1716.

44. Stahl, E.A., Breen, G., Forstner, A.J., McQuillin, A., Ripke, S., Trubetskoy, V., Mattheisen, M., Wang, Y., Coleman, J.R.I., Gaspar, H.A., et al. (2019). Genome-wide association study identifies 30 loci associated with bipolar disorder. Nat Genet 51, 793-803.

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

46. Stuart, T., Butler, A., Hoffman, P., Hafemeister, C., Papalexi, E., Mauck, W.M., 3rd, Hao, Y., Stoeckius, M., Smibert, P., and Satija, R. (2019). Comprehensive Integration of Single-Cell Data. Cell 177, 1888-1902 e1821.

47. Sonawane, A.R., Platig, J., Fagny, M., Chen, C.Y., Paulson, J.N., Lopes-Ramos, C.M., DeMeo, D.L., Quackenbush, J., Glass, K., and Kuijjer, M.L. (2017). Understanding Tissue-Specific Gene Regulation. Cell Rep 21, 1077-1088.

48. Hounkpe, B.W., Chenou, F., de Lima, F., and De Paula, E.V. (2021). HRT Atlas v1.0 database: redefining human and mouse housekeeping genes and candidate reference transcripts by mining massive RNA-seq datasets. Nucleic Acids Res 49, D947-D955.

49. Field, F.J., Kam, N.T., and Mathur, S.N. (1990). Regulation of cholesterol metabolism in the intestine. Gastroenterology 99, 539-551.

50. Ko, C.W., Qu, J., Black, D.D., and Tso, P. (2020). Regulation of intestinal lipid metabolism: current concepts and relevance to disease. Nat Rev Gastroenterol Hepatol 17, 169-183.

51. Kruit, J.K., Groen, A.K., van Berkel, T.J., and Kuipers, F. (2006). Emerging roles of the intestine in control of cholesterol metabolism. World J Gastroenterol 12, 6429-6439.

52. Cerf, M.E. (2013). Beta cell dysfunction and insulin resistance. Front Endocrinol (Lausanne) 4, 37.

53. Donath, M.Y., Ehses, J.A., Maedler, K., Schumann, D.M., Ellingsgaard, H., Eppler, E., and Reinecke, M. (2005). Mechanisms of beta-cell death in type 2 diabetes. Diabetes 54 Suppl 2, S108-113.

54. Maedler, K., and Donath, M.Y. (2004). Beta-cells in type 2 diabetes: a loss of function and mass. Horm Res 62 Suppl 3, 67-73.

55. Chandra, R., and Liddle, R.A. (2009). Neural and hormonal regulation of pancreatic secretion. Curr Opin Gastroenterol 25, 441-446.

56. Chandra, R., and Liddle, R.A. (2014). Recent advances in the regulation of pancreatic secretion. Curr Opin Gastroenterol 30, 490-494.

57. Washabau, R.J. (2013). Chapter 1 - Integration of Gastrointestinal Function. In Canine and Feline Gastroenterology, R.J. Washabau and M.J. Day, eds. (Saint Louis, W.B. Saunders), pp 1-31.

58. Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4, 44-57.

59. Huang da, W., Sherman, B.T., and Lempicki, R.A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res 37, 1-13.

60. Pardinas, A.F., Holmans, P., Pocklington, A.J., Escott-Price, V., Ripke, S., Carrera, N., Legge, S.E., Bishop, S., Cameron, D., Hamshere, M.L., et al. (2018). Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. Nat Genet 50, 381-389.

61. Fromer, M., Roussos, P., Sieberts, S.K., Johnson, J.S., Kavanagh, D.H., Perumal, T.M., Ruderfer, D.M., Oh, E.C., Topol, A., Shah, H.R., et al. (2016). Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci 19, 1442-1453.

62. Jiang, Y., Zhang, N.R., and Li, M. (2017). SCALE: modeling allele-specific gene expression by single-cell RNA sequencing. Genome Biol 18, 74.

63. Korthauer, K.D., Chu, L.F., Newton, M.A., Li, Y., Thomson, J., Stewart, R., and Kendziorski, C. (2016). A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. Genome Biol 17, 222.

64. Wang, J., Huang, M., Torre, E., Dueck, H., Shaffer, S., Murray, J., Raj, A., Li, M., and Zhang, N.R. (2018). Gene expression distribution deconvolution in single-cell RNA sequencing. Proc Natl Acad Sci U S A 115, E6437-E6446.

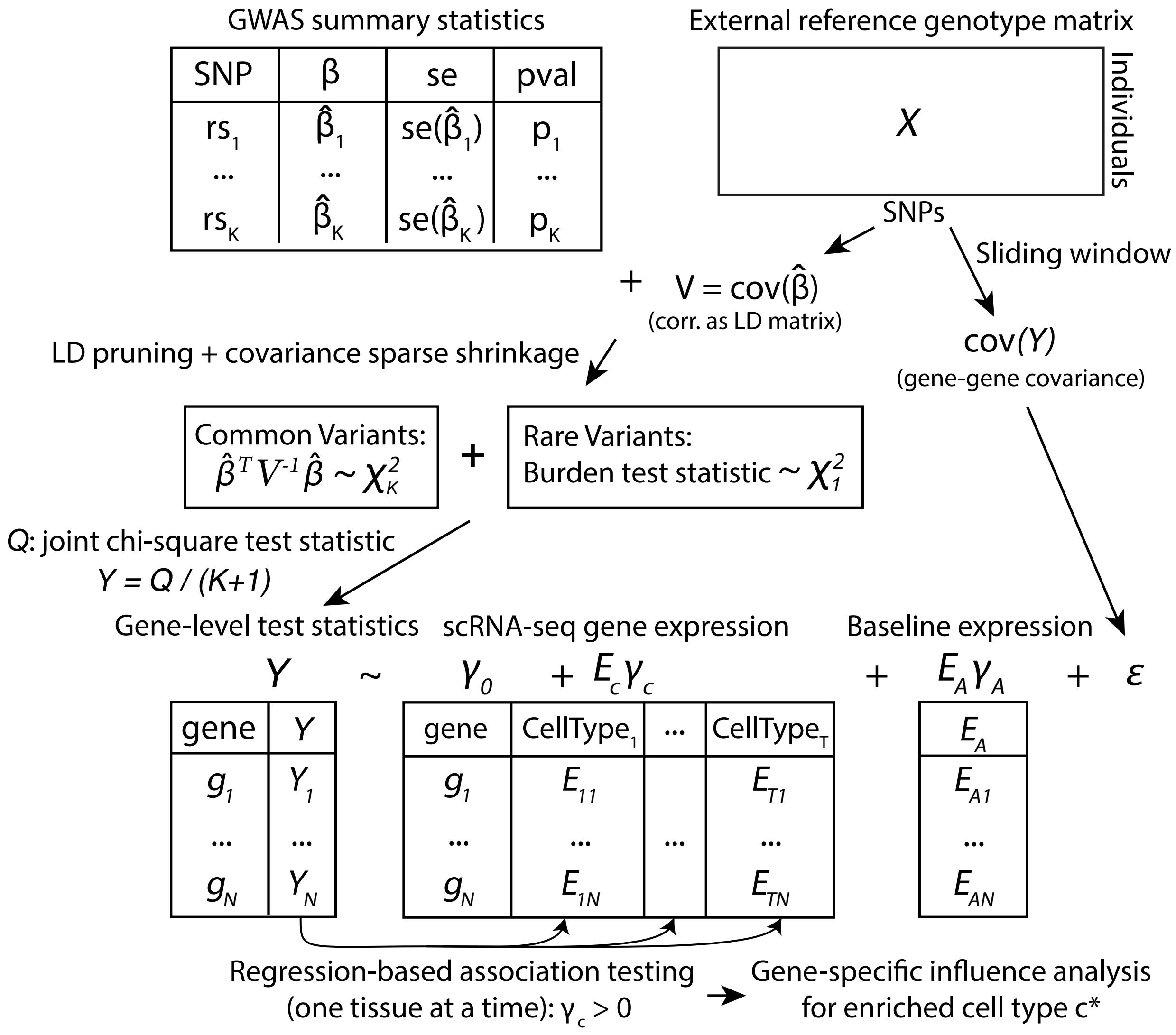
65. Dong, M., Thennavan, A., Urrutia, E., Li, Y., Perou, C.M., Zou, F., and Jiang, Y. (2021). SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. Brief Bioinform 22, 416-427.

66. Haghverdi, L., Lun, A.T.L., Morgan, M.D., and Marioni, J.C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol 36, 421-427.

67. Urrutia, E., Chen, L., Zhou, H., and Jiang, Y. (2019). Destin: toolkit for single-cell analysis of chromatin accessibility. Bioinformatics 35, 3818-3820.

68. Granja, J.M., Corces, M.R., Pierce, S.E., Bagdatli, S.T., Choudhry, H., Chang, H.Y., and Greenleaf, W.J. (2021). ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. Nat Genet 53, 403-411.

69. van der Wijst, M.G.P., Brugge, H., de Vries, D.H., Deelen, P., Swertz, M.A., LifeLines Cohort, S., Consortium, B., and Franke, L. (2018). Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. Nat Genet 50, 493-497.


GWAS summary statistics
External reference genotype matrix
X
Individuals
SNPs
Sliding window
+ $V = \text{cov}(\hat{\beta})$
(corr. as LD matrix)
$\text{cov}(Y)$
(gene-gene covariance)
LD pruning + covariance sparse shrinkage
Common Variants: $\hat{\beta}^T V^{-1} \hat{\beta} \sim \chi^2_K$
+
Rare Variants: Burden test statistic $\sim \chi^2_1$
Q: joint chi-square test statistic
$Y = Q / (K+1)$
Gene-level test statistics
scRNA-seq gene expression
Baseline expression
$Y \sim \gamma_0 + E_c\gamma_c + E_A\gamma_A + \varepsilon$
Regression-based association testing (one tissue at a time): $\gamma_c > 0$
Gene-specific influence analysis for enriched cell type c*

SNP	β	se	pval
rs_1	$\hat{\beta}_1$	$se(\hat{\beta}_1)$	p_1
...	...	...	...
rs_K	$\hat{\beta}_K$	$se(\hat{\beta}_K)$	p_K

gene	Y
g_1	Y_1
...	...
g_N	Y_N

gene	CellType_1	...	CellType_T
g_1	E_{11}		E_{T1}
...	...	...	...
g_N	E_{1N}		E_{TN}

E_A
E_{A1}
...
E_{AN}

Figure 1

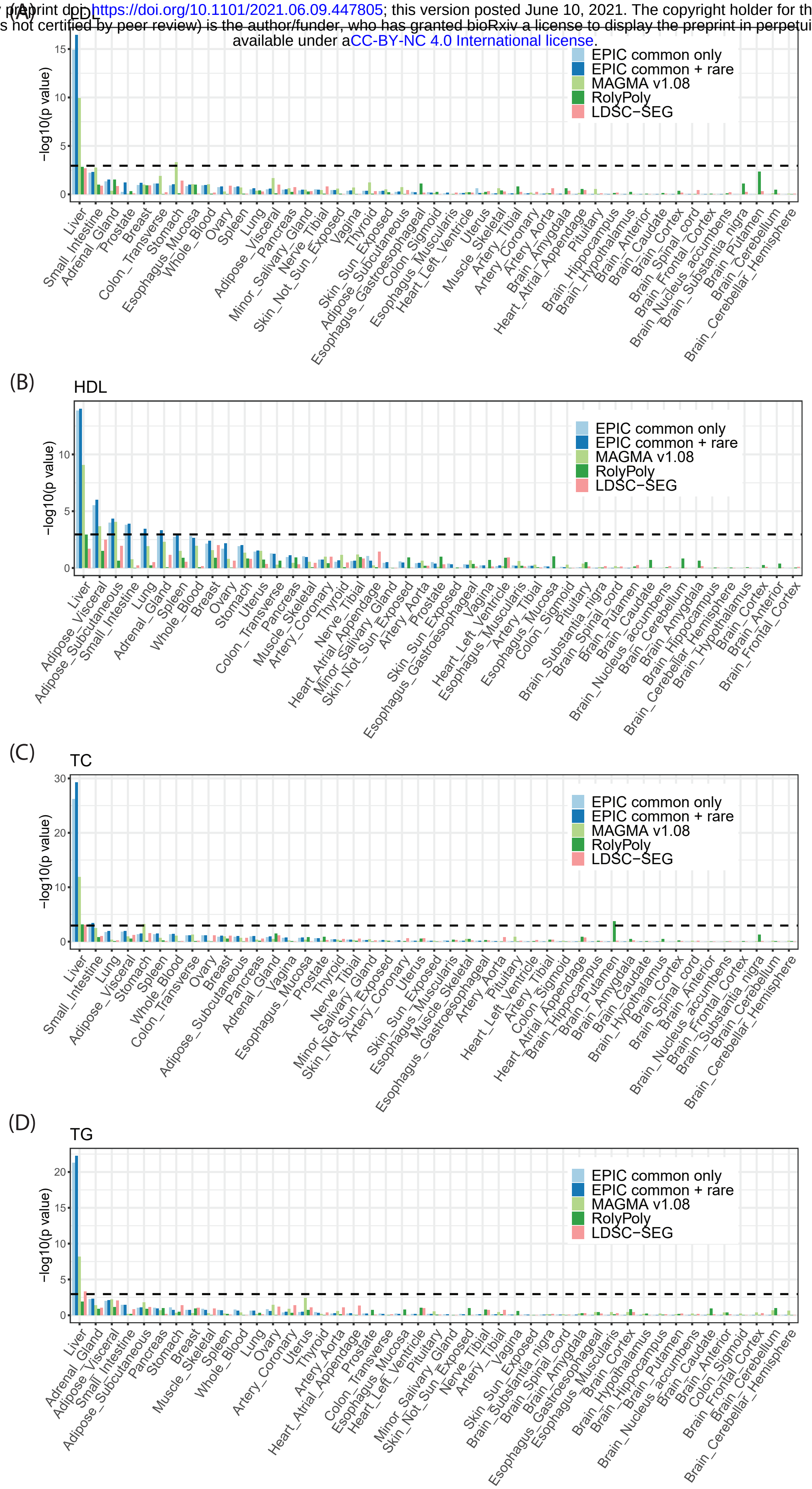
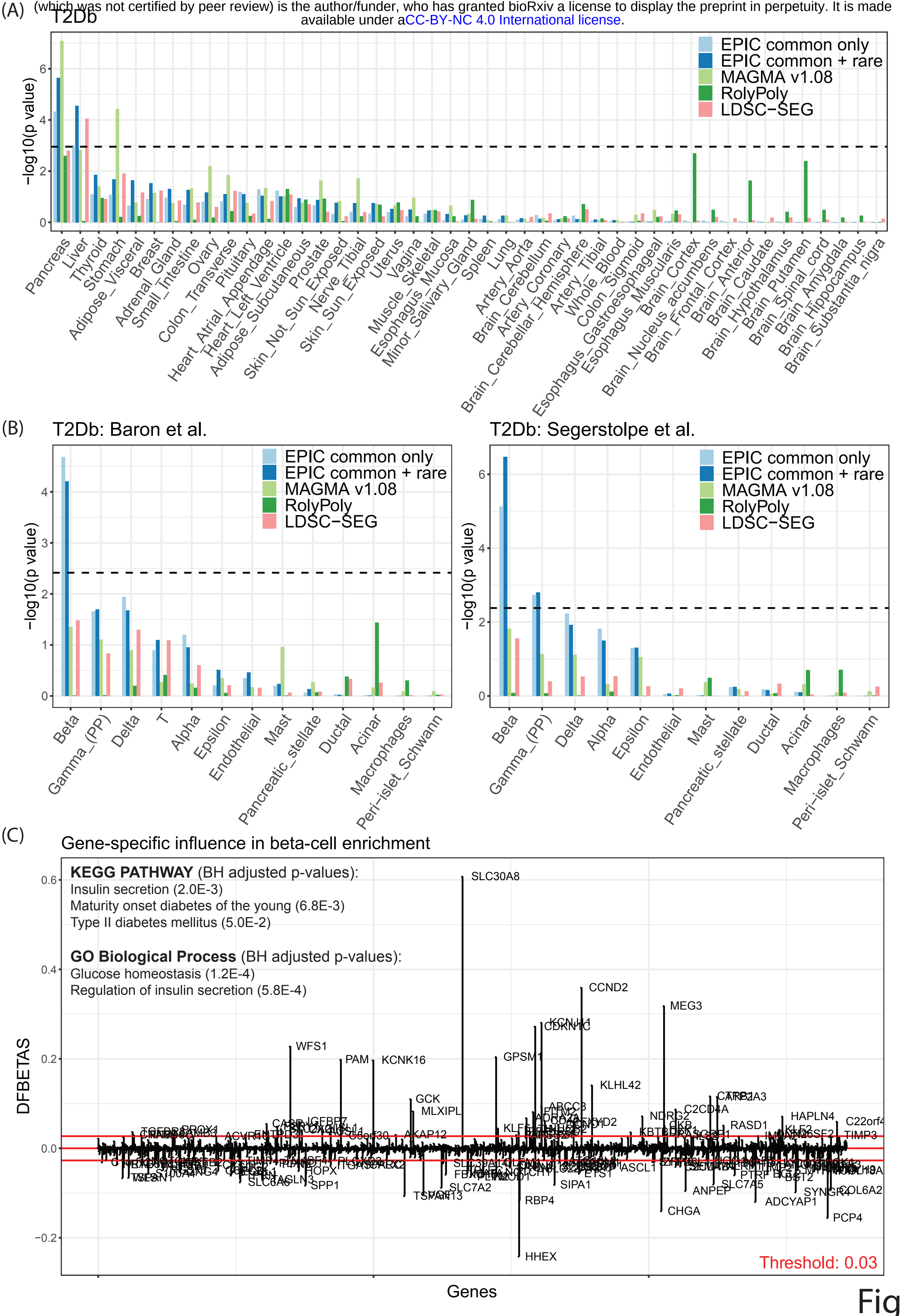

(A)
LDL
−log10(p value)
0
5
10
15
EPIC common only
EPIC common + rare
MAGMA v1.08
RolyPoly
LDSC−SEG
Liver
Small_Intestine
Adrenal_Gland
Prostate
Breast
Colon_Transverse
Stomach
Esophagus_Mucosa
Whole_Blood
Ovary
Spleen
Lung
Adipose_Visceral
Pancreas
Minor_Salivary_Gland
Nerve_Tibial
Skin_Not_Sun_Exposed
Vagina
Thyroid
Skin_Sun_Exposed
Adipose_Subcutaneous
Esophagus_Gastroesophageal
Colon_Sigmoid
Esophagus_Muscularis
Heart_Left_Ventricle
Uterus
Muscle_Skeletal
Artery_Tibial
Artery_Coronary
Artery_Aorta
Brain_Amygdala
Heart_Atrial_Appendage
Pituitary
Brain_Hippocampus
Brain_Hypothalamus
Brain_Anterior
Brain_Caudate
Brain_Cortex
Brain_Spinal_cord
Brain_Frontal_Cortex
Brain_Nucleus_accumbens
Brain_Substantia_nigra
Brain_Putamen
Brain_Cerebellum
Brain_Cerebellar_Hemisphere
(B)
HDL
−log10(p value)
0
5
10
EPIC common only
EPIC common + rare
MAGMA v1.08
RolyPoly
LDSC−SEG
Liver
Adipose_Visceral
Adipose_Subcutaneous
Small_Intestine
Lung
Adrenal_Gland
Spleen
Whole_Blood
Breast
Ovary
Stomach
Uterus
Colon_Transverse
Pancreas
Muscle_Skeletal
Artery_Coronary
Thyroid
Nerve_Tibial
Heart_Atrial_Appendage
Minor_Salivary_Gland
Skin_Not_Sun_Exposed
Artery_Aorta
Prostate
Skin_Sun_Exposed
Esophagus_Gastroesophageal
Vagina
Heart_Left_Ventricle
Esophagus_Muscularis
Artery_Tibial
Esophagus_Mucosa
Colon_Sigmoid
Pituitary
Brain_Substantia_nigra
Brain_Spinal_cord
Brain_Putamen
Brain_Caudate
Brain_Nucleus_accumbens
Brain_Cerebellum
Brain_Amygdala
Brain_Hippocampus
Brain_Cerebellar_Hemisphere
Brain_Hypothalamus
Brain_Cortex
Brain_Anterior
Brain_Frontal_Cortex
(C)
TC
−log10(p value)
0
10
20
30
EPIC common only
EPIC common + rare
MAGMA v1.08
RolyPoly
LDSC−SEG
Liver
Small_Intestine
Lung
Adipose_Visceral
Stomach
Spleen
Whole_Blood
Colon_Transverse
Ovary
Breast
Adipose_Subcutaneous
Pancreas
Adrenal_Gland
Vagina
Esophagus_Mucosa
Prostate
Thyroid
Nerve_Tibial
Minor_Salivary_Gland
Skin_Not_Sun_Exposed
Artery_Coronary
Uterus
Skin_Sun_Exposed
Esophagus_Muscularis
Muscle_Skeletal
Esophagus_Gastroesophageal
Artery_Aorta
Heart_Left_Ventricle
Pituitary
Artery_Tibial
Colon_Sigmoid
Heart_Atrial_Appendage
Brain_Hippocampus
Brain_Putamen
Brain_Amygdala
Brain_Caudate
Brain_Hypothalamus
Brain_Cortex
Brain_Spinal_cord
Brain_Anterior
Brain_Nucleus_accumbens
Brain_Frontal_Cortex
Brain_Substantia_nigra
Brain_Cerebellum
Brain_Cerebellar_Hemisphere
(D)
TG
−log10(p value)
0
5
10
15
20
EPIC common only
EPIC common + rare
MAGMA v1.08
RolyPoly
LDSC−SEG
Liver
Adrenal_Gland
Adipose_Visceral
Small_Intestine
Adipose_Subcutaneous
Pancreas
Stomach
Breast
Muscle_Skeletal
Spleen
Whole_Blood
Lung
Ovary
Artery_Coronary
Uterus
Thyroid
Artery_Aorta
Heart_Atrial_Appendage
Prostate
Colon_Transverse
Esophagus_Mucosa
Heart_Left_Ventricle
Pituitary
Minor_Salivary_Gland
Skin_Not_Sun_Exposed
Nerve_Tibial
Artery_Tibial
Vagina
Skin_Sun_Exposed
Brain_Substantia_nigra
Brain_Spinal_cord
Brain_Amygdala
Esophagus_Gastroesophageal
Esophagus_Muscularis
Brain_Cortex
Brain_Hypothalamus
Brain_Hippocampus
Brain_Putamen
Brain_Nucleus_accumbens
Brain_Caudate
Brain_Anterior
Colon_Sigmoid
Brain_Frontal_Cortex
Brain_Cerebellum
Brain_Cerebellar_Hemisphere

Figure 2



KEGG PATHWAY (BH adjusted p-values):
Insulin secretion (2.0E-3)
Maturity onset diabetes of the young (6.8E-3)
Type II diabetes mellitus (5.0E-2)

GO Biological Process (BH adjusted p-values):
Glucose homeostasis (1.2E-4)
Regulation of insulin secretion (5.8E-4)

Figure 3

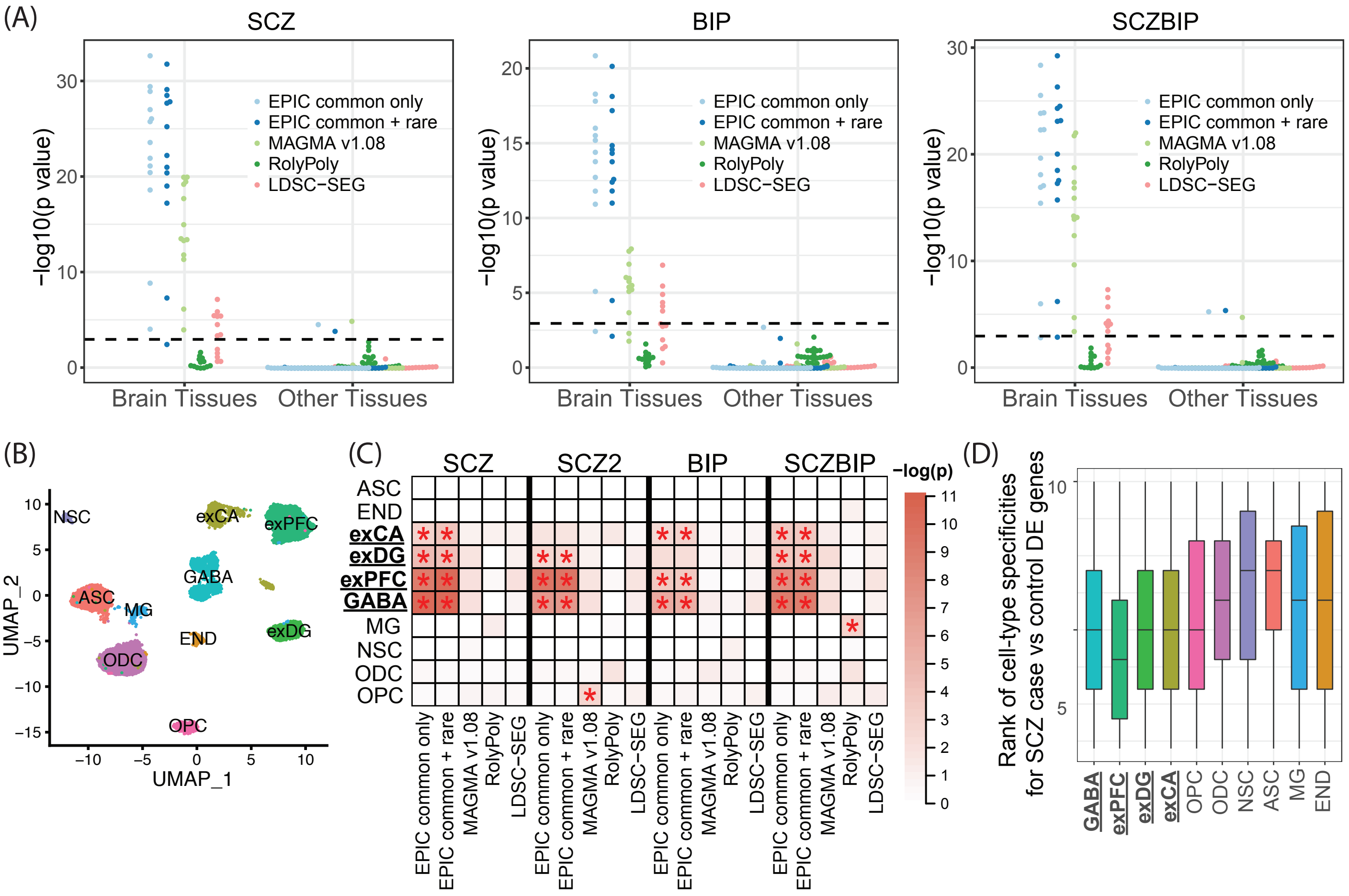

(A)
SCZ
BIP
SCZBIP
−log10(p value)
0
10
20
30
5
15
Brain Tissues
Other Tissues
EPIC common only
EPIC common + rare
MAGMA v1.08
RolyPoly
LDSC−SEG
(B)
NSC
exCA
exPFC
GABA
ASC
MG
END
exDG
ODC
OPC
UMAP_2
UMAP_1
10
5
0
−5
−10
−15
(C)
SCZ
SCZ2
BIP
SCZBIP
−log(p)
ASC
END
exCA
exDG
exPFC
GABA
MG
NSC
ODC
OPC
EPIC common only
EPIC common + rare
MAGMA v1.08
RolyPoly
LDSC−SEG
*
11
10
9
8
7
6
5
4
3
2
1
0
(D)
Rank of cell-type specificities for SCZ case vs control DE genes
10
5
GABA
exPFC
exDG
exCA
OPC
ODC
NSC
ASC
MG
END

Figure 4

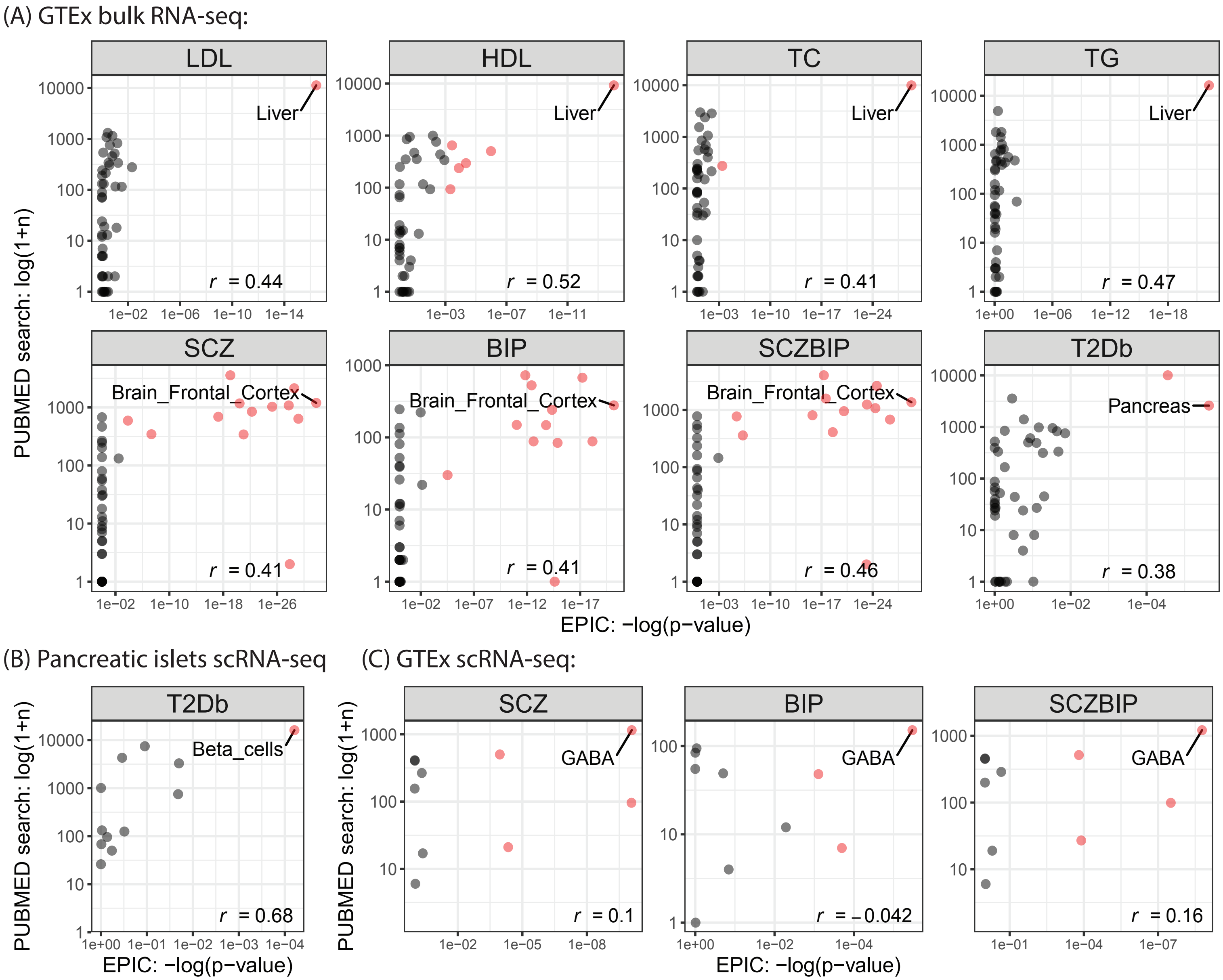

(A) GTEx bulk RNA-seq:
LDL
Liver
r = 0.44
HDL
Liver
r = 0.52
TC
Liver
r = 0.41
TG
Liver
r = 0.47
SCZ
Brain_Frontal_Cortex
r = 0.41
BIP
Brain_Frontal_Cortex
r = 0.41
SCZBIP
Brain_Frontal_Cortex
r = 0.46
T2Db
Pancreas
r = 0.38
PUBMED search: log(1+n)
EPIC: −log(p−value)
10000
1000
100
10
1
1e−02
1e−06
1e−10
1e−14
1e−03
1e−07
1e−11
1e−10
1e−17
1e−24
1e+00
1e−12
1e−18
1e−18
1e−26
1e−12
1e−04
(B) Pancreatic islets scRNA-seq
T2Db
Beta_cells
r = 0.68
PUBMED search: log(1+n)
EPIC: −log(p−value)
1e+00
1e−01
1e−02
1e−03
1e−04
(C) GTEx scRNA-seq:
SCZ
GABA
r = 0.1
BIP
GABA
r = −0.042
SCZBIP
GABA
r = 0.16
PUBMED search: log(1+n)
EPIC: −log(p−value)
1e−02
1e−05
1e−08
1e−01
1e−04
1e−07

Figure 5