SUPPLEMENTARY NOTE

Choice of parameters. Our approach includes two parameters: the proportion of genes selected, which we set to 10%, and the window size around each gene, which we set to 100kb. To choose these two parameters, we ran the approach with six different parameter settings ({2%, 5%, 10% of genes} x {20kb, 100kb windows}) on two diseases—schizophrenia and rheumatoid arthritis—and two corresponding GTEx tissues—brain (all brain regions) and blood (LCLs and whole blood)—which are widely known to be disease-relevant tissues. We determined that of the parameter settings we tested, 10% of genes and 100kb produced the most significant P-values for identifying brain enrichment for schizophrenia and blood enrichment for rheumatoid arthritis, so we used these parameters for the remaining analyses. Our results were robust to these choices (Figure S2).

Number of gene expression samples needed. Because the GTEx consortium data set included tens of samples for many of the tissues, we were able to assess how sensitive our results were to the sample size of the gene expression data set used to construct the gene sets. To do this, we repeatedly sub-sampled our data set to a variety of sample sizes, each time re-creating gene sets using the smaller sub-sampled data set. We chose two results to re-analyze in this way. First, we re-analyzed cortex enrichment for schizophrenia, in which cortex was compared to all non-brain samples and was highly significant (**Figure 2**). This result was very robust: the enrichment was highly significant in all of our downsampled data sets, even with only a single cortex sample (**Figure S11a**). We then assessed enrichment for schizophrenia in the within-brain analysis, in which cortex was compared to all other brain regions and was moderately significant (**Figure 4a**). In this analysis, sample size was more important, and while there was high variance in z-score among random samples at a given sample size, there was a clear trend that increasing the sample size increases the significance of the enrichment on average (**Figure S11b**). In conclusion, these analyses provide evidence that sample size can be important when the enrichment being identified is near the border of significance, but that our method is well-powered to detect strong signals even with a single sample in the tissue of interest.

Comparison to existing methods: real phenotypes. To our knowledge, SNPsea[1,2] is the only existing method that takes as input GWAS summary statistics, together with a matrix of gene expression values, and identifies enriched tissues and cell types. SNPsea leverages only genome-wide significant SNPs, rather than all SNPs, a notable difference from our approach. We ran SNPsea on the summary statistics and gene expression data analyzed in our multiple-tissue analysis; results are displayed in **Figure S12 and Table S11**. We found that SNPsea identified biological plausible enrichments at high levels of significance for traits such as LDL for which a large proportion of SNP-heritability lies in genome-wide significant loci, but that it was not well-powered for more polygenic traits; for example, it found zero tissues with FDR < 5% for bipolar disorder, while our approach found many brain regions to be enriched at P-values as low as 2e-12 (**Figure S1**). The lack of power of SNPsea on more polygenic traits is unsurprising, as SNPsea leverages only genome-wide significant loci.

The DEPICT software[3] includes a method for identifying disease-relevant tissues and cell types from GWAS summary statistics and gene expression data. However, this method takes as input only the GWAS summary statistics and not gene expression data; the method is designed to be run only with the Franke lab data set[3,4], which is built into the software. Thus, DEPICT could not be used to obtain the results in our brain-specific and immune-specific analyses, for which we analyzed data sets that allowed us to differentiate among tissues and cell types within each of these systems. However, DEPICT does perform a multiple-tissue analysis analogous to the Franke lab data set component of our multiple-tissue analysis, and so we ran DEPICT on the set of summary statistics that we analyzed. Like SNPsea, DEPICT is run on a subset of SNPs, but unlike SNPsea, DEPICT documentation recommends that it be run twice, once on SNPs that pass genome-wide significance at 5e-8, and once on SNPs that pass a less stringent threshold of 1e-5; we followed this recommendation, and our results are displayed in **Figures S13** and **S14** and **Tables S12** and **S13**. We determined that DEPICT failed to identify some enrichments identified by our analysis of the Franke lab data set, such as brain enrichment for several brain-related traits (epilepsy, Tourette syndrome, neuroticism, and smoking status), but that it identified a large number of enrichments for other traits and tissues that our approach did not find. In simulations described below, we found that DEPICT sometimes reported significant results in the absence of true enrichment.

Our approach, described in Figure 1, has two main steps: constructing a genome annotation from gene expression data, and testing this annotation for enrichment with GWAS summary statistics using stratified LD score regression. We tested whether the success of our approach depended on using stratified LD score regression in the second step by instead analyzing the specifically expressed gene annotations from the first step using MAGMA[5], a gene set enrichment method that allows inclusion of a window around each gene and leverages all SNPs in the gene set (**Figure S15, Table S14**). MAGMA and LDSC-SEG identified many of the same enrichments, but MAGMA identified several enrichments that LDSC-SEG did not. We hypothesized that this may occur because in this analysis, we did not use the option in MAGMA to incorporate gene-level covariates. In simulations described below, we determined that MAGMA can report significant results in the absence of true enrichment due to uncorrected genomic confounding if no covariates are included to ameliorate potential confounding. We leave an exploration of how best to use covariates in MAGMA to account for potential confounding while preserving power for future work.

For comparison purposes, we report LDSC-SEG results for the multiple tissue analysis as a heatmap in **Figure S2a**, in addition to the scatter plots in **Figure 2** and **Figure S1**.

Comparison to existing methods: simulated phenotypes. We performed simulations using genotypes from Genetic Epidemiology Research on Aging (GERA) data set[6–8] with 47,360 individuals and 6,507,309 SNPs with imputation $R^2 > 0.5$. We simulated five genetic architectures, where “null” refers to a heritable trait with no tissue-specific enrichment and “causal” refers to a heritable trait with cortex enrichment:

1. (Polygenic null) All SNPs causal, causal SNP effects are drawn independently from a normal distribution with mean zero and constant variance across the genome, with a total heritability of 0.9.
2. (Sparse null) Same as (1), but each SNP has probability 0.001 of being causal.
3. (Exon-enriched null) A SNP is causal if and only if it is in an exon, causal SNP effects are drawn independently from a normal distribution with mean zero and constant variance for all exonic SNPs, with a total heritability of 0.9.
4. (Polygenic causal) We use the annotation corresponding to cortex genes from the multiple-tissue analysis to simulate a true effect. All SNPs are causal, causal SNP effects are drawn independently from a normal distribution with a constant variance within the cortex annotation and constant variance outside of the cortex annotation so that 50% of the total heritability is assigned to the cortex annotation, 50% of the total heritability is distributed uniformly across the genome, and the total heritability is 0.2. We chose a smaller value of heritability in the causal simulations because we wanted to test power to identify true enrichment rather than control of type I error.
5. (Sparse causal) Same as (4), but each SNP has a probability of 0.001 to be causal.

For each genetic architecture, we simulated phenotypes and summary statistics using PLINK[9] (see URLs) with 100 replicates for each genetic architecture. We then ran the multiple-tissue analysis as described above for every method on each of the simulated data sets, and for each method and each simulated genetic architecture we performed FDR correction within the set of 100 simulated phenotypes. Results are displayed in **Figure S16** and **Table S15**.

Of the five methods tested (LDSC-SEG, SNPsea, DEPICT (1e-5), DEPICT (5e-8), and MAGMA), only LDSC-SEG and SNPsea correctly reported no significant enrichments passing FDR<5% for all 3 null simulations (scenarios 1-3). In particular, DEPICT with a threshold of 1e-5 reported significant enrichments at FDR<5% for all three null simulations (scenarios 1-3), while DEPICT with a threshold of 5e-8 reported significant enrichments at FDR < 5% for the sparse null simulation (scenario 2). MAGMA correctly reported no significant enrichment for the null simulations with no enrichment (scenarios 1-2) but reported a large number of significant enrichments at FDR<5% for the null simulation with enrichment in exons (scenario 3); we note that we ran MAGMA without taking advantage of the option to incorporate gene-level covariates which would likely ameliorate the false positives.

All five methods reported significant cortex enrichments at FDR<5% for the sparse causal simulation (scenario 5), but only MAGMA and LDSC-SEG reported significant cortex enrichments for the polygenic causal simulation (scenario 4). These simulations, together with the analysis of real phenotypes described above, indicate that when MAGMA is run without covariates, only LDSC-SEG and SNPsea control type I error, and that of these two methods, LDSC-SEG is better powered for polygenic traits.

Relationship of power and sample size at very large sample sizes. Power increases only modestly with sample size at very large sample sizes, as the finite size of the genome is a stricter constraint for highly heritable traits at these sample sizes: for example, LD score

regression coefficients of baseline model annotations had s.e. that were only 1.29x lower on average in analyses of the full UK Biobank data set (average *N*=438,682) vs. the interim UK Biobank data set (average *N*=140,026.)

References

1. Hu, X. *et al.* Integrating Autoimmune Risk Loci with Gene-Expression Data Identifies Specific Pathogenic Immune Cell Subsets. *Am. J. Hum. Genet.* **89,** 496–506 (2011).
2. Slowikowski, K., Hu, X. & Raychaudhuri, S. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. *Bioinformatics* **30,** 2496–2497 (2014).
3. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6,** 5890 (2015).
4. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47,** 115–125 (2015).
5. Leeuw, C. A. de, Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set Analysis of GWAS Data. *PLOS Comput Biol* **11,** e1004219 (2015).
6. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200,** 1285–1295 (2015).
7. Loh, P.-R. *et al.* Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. *Nat. Genet.* **47,** 1385–1392 (2015).
8. Galinsky, K. J. *et al.* Fast Principal-Component Analysis Reveals Convergent Evolution of ADH1B in Europe and East Asia. *Am. J. Hum. Genet.* **98,** 456–472 (2016).
9. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** 7 (2015).