Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types

Hilary K. Finucane[1,2,*], Yakir A. Reshef[3], Verneri Anttila[4,5,6], Kamil Slowikowski[7], Alexander Gusev[2], Andrea Byrnes[4,5,6], Steven Gazal[2], Po-Ru Loh[2], Giulio Genovese[5,6], Arpiar Saunders[8], Evan Macosko[8], Samuela Pollack[2], The Brainstorm Consortium, John R.B. Perry[9], Soumya Raychaudhuri[6,10-13], Steven McCarroll[5,6,8], Benjamin M. Neale[4,5,6], Alkes L. Price[2,*]

[1]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
[2]Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA.
[3]Department of Computer Science, Harvard University, Cambridge, Massachusetts, USA.
[4]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts, USA.
[5]Stanley Center for Psychiatric Research, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
[6]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
[7]Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA.
[8]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.
[9]Medical Research Council (MRC) Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK.
[10]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[11]Division of Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, Massachusetts, USA.
[12]Partners Center for Personalized Genetic Medicine, Boston, Massachusetts, USA.
[13]Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK.

*Correspondence should be addressed to HKF (hilaryf@mit.edu) or ALP (aprice@hsph.harvard.edu).

ABSTRACT

Genetics can provide a systematic approach to discovering the tissues and cell types relevant for a complex disease or trait. Identifying these tissues and cell types is critical for following up on non-coding allelic function, developing ex-vivo models, and identifying therapeutic targets. Here, we analyze gene expression data from several sources, including the GTEx and PsychENCODE consortia, together with genome-wide association study (GWAS) summary statistics for 48 diseases and traits with an average sample size of 86,850, to identify disease-relevant tissues and cell types. We develop and apply an approach that uses stratified LD score regression to test whether disease heritability is enriched in regions surrounding genes with the highest specific expression in a given tissue. We detect tissue-specific enrichments at FDR < 5% for 30 diseases and traits across a broad range of tissues that recapitulate known biology. In our analysis of traits with observed central nervous system enrichment, we detect an enrichment of neurons over other brain cell types for several brain-related traits, enrichment of inhibitory neurons over excitatory neurons for bipolar disorder, and enrichments in the cortex for schizophrenia and in the striatum for migraine. In our analysis of traits with observed immunological enrichment, we identify enrichments of alpha beta T cells for asthma and eczema, B cells for primary biliary cirrhosis, and myeloid cells for lupus and Alzheimer's disease. Our results demonstrate that our polygenic approach is a powerful way to leverage gene expression data for interpreting GWAS signal.

INTRODUCTION

There are many diseases whose causal tissues or cell types are uncertain or unknown. Identifying these tissues and cell types is critical for developing systems to explore gene regulatory mechanisms that contribute to disease. In recent years, researchers have been gaining an increasingly clear picture of which parts of the genome are active in a range of tissues and cell types: for example, which parts of the genome are accessible, which enhancers are active, and which genes are expressed[1–3]. Combining this type of information with GWAS data offers the potential to identify causal tissues and cell types for disease.

Many different types of data characterizing tissue- and cell-type-specific activity have been analyzed together with GWAS data to identify disease-relevant tissues and cell types: histone marks[4–7], DNase I hypersensitivity (DHS)[8–11], eQTLs[3,12], and gene expression data[13–16]. Of these data types, gene expression data (without genotypes or eQTLs) has the advantage of being available in the widest range of tissues and cell types. Therefore, methods for integrating gene expression data with GWAS data have the potential not only to identify system-level differences among traits—e.g., brain enrichment vs. immune enrichment—but also to obtain high resolution within a system—e.g., differentiating among brain regions or among immune cell types.

Indeed, previous work has shown that gene expression can be a useful source of information for identifying disease-relevant tissues and cell types from GWAS data. An initial application of the SNPsea algorithm[13,14] analyzed a data set with gene expression in 249 immune cell types from mouse, together with genome-wide significant SNPs from GWAS of several immunological diseases, and reported disease-specific patterns of enrichment[13]. The DEPICT software[15] includes a method for joint analysis of GWAS summary statistics with a large gene expression data set[17], and has been used to identify enriched tissues for height[18] and BMI[19]. In a recent study of migraine[16], an analysis of genome-wide significant loci with expression data from the GTEx project identified cardiovascular and digestive/smooth muscle enrichments. These studies show that gene expression data are informative for disease-relevant tissues and cell types, and have led to biological insights about the diseases and traits studied. However, the methods applied in these studies restrict their analyses to subsets of SNPs that pass a significance threshold. To our knowledge, no previous study has modeled genome-wide polygenic signals to identify disease-relevant tissues and cell types from GWAS and gene expression data.

Here, we apply stratified LD score regression[6], a method for partitioning heritability from GWAS summary statistics, to sets of specifically expressed genes to identify disease-relevant tissues and cell types across 48 diseases and traits with an average GWAS sample size of 86,850. We first analyze two gene expression data sets[3,15,17] containing a wide range of tissues to infer system-level enrichments, recapitulating known biology. We also analyze chromatin data from the Roadmap Epigenomics project[2] across the same set of diseases and traits, and conclude that gene expression and chromatin provide complementary information. We then analyze gene expression data sets that allow us to achieve higher resolution within a system[3,20–22], identifying enriched brain regions, brain cell types, and immune cell types for several brain- and immune-related diseases and traits. Our results

underscore that a heritability-based framework applied to gene expression data allows us to achieve high-resolution enrichments, even for very polygenic traits.

RESULTS

Overview of methods

We analyzed the five gene expression data sets listed in **Table 1**, mapping mouse genes to orthologous human genes when necessary. To assess the enrichment of a focal tissue for a given trait, we follow the procedure described in **Figure 1**. We begin with a matrix of normalized gene expression values across genes, with samples from multiple tissues including the focal tissue. For each gene, we compute a t-statistic for specific expression in the focal tissue (Online Methods). We rank all genes by their t-statistic, and define the 10% of genes with the highest t-statistic to be the gene set corresponding to the focal tissue; we call this the set of specifically expressed genes, but we note that this includes not only genes that are strictly specifically expressed (i.e. only expressed in the focal tissue), but also genes that are weakly specifically expressed (i.e. higher average expression in the focal tissue). For a few of the datasets analyzed, we modified our approach to constructing the set of specifically expressed genes to better take advantage of the data available (Online Methods). We add 100kb windows on either side of the transcribed region of each gene in the set of specifically expressed genes to construct a genome annotation corresponding to the focal tissue. (The choice of the parameters 10% and 100kb is discussed in Online Methods.) Finally, we apply stratified LD score regression[6] to GWAS summary statistics to evaluate the contribution of the focal genome annotation to trait heritability (Online Methods). We jointly model the annotation corresponding to the focal tissue, a genome annotation corresponding to all genes, and the 52 annotations in the "baseline model"[6] (including genic regions, enhancer regions, and conserved regions; see **Table S1**). A positive regression coefficient for the focal annotation in this regression represents a positive contribution of this annotation to trait heritability, conditional on the other annotations. We report regression coefficients, normalized by mean per-SNP heritability, together with a P-value to test whether the regression coefficient is significantly positive. Stratified LD score regression requires GWAS summary statistics for the trait of interest, together with an LD reference panel (e.g. 1000 Genomes[23]), and has been shown to produce robust results with properly controlled type I error[6]. We have released open source software implementing our approach, and have also released all genome annotations derived from the publicly available gene expression data that we analyzed (see **URLs**). We call our approach LD score regression applied to specifically expressed genes (LDSC-SEG).

Analysis of 48 complex traits across multiple tissues

We first analyzed two gene expression data sets. The first data set, from the GTEx consortium v6p[3], consists of RNA-seq data for 53 tissues, with an average of 161 samples per tissue (**Table S2,** Online Methods). The second data set, which we call the Franke lab data set, is an aggregation of publicly available microarray gene expression data sets comprising 37,427 samples in human, mouse, and rat[15,17]. After removing redundant data,

this data set contained 152 tissues, including much better representation of immune tissues and cell types than the GTEx data set (**Table S3,** Online Methods). The gene expression values in the Franke lab data set already quantify relative expression for a tissue/cell-type rather than absolute expression for a single sample, and so we used these values in place of our t-statistics. For visualization purposes, we classified the 205 tissues and cell types in these data sets into nine categories; the classification is described in **Table S2** and **Table S3**. The main goal of this multiple-tissue analysis was to identify system-level enrichments.

We analyzed GWAS summary statistics for 48 diseases and traits with an average sample size of 86,850 (**Table S4**), applying LDSC-SEG for each of the 205 specifically expressed gene annotations in turn. The 48 traits included 12 traits from the UK Biobank[24], 17 traits with publicly available GWAS summary statistics[25–36], and 19 traits from the Brainstorm Consortium[37–46]. We excluded the HLA region from all analyses, due to its unusual genetic architecture and pattern of LD. For 30 of the 48 traits, at least one tissue was significant at FDR<5% (**Figure 2**, **Figure S1** and **Table S5)**. Averaging across the most significant tissue for each of these 30 traits, the specifically expressed gene annotation spanned 17% of the genome and explained 38% of SNP-heritability (**Table S5**). Several of our results recapitulate known biology: immunological traits exhibit immune cell-type enrichments, psychiatric traits exhibit strong brain enrichment, LDL and triglycerides exhibit liver-specific enrichments, BMI-adjusted waist-hip ratio exhibits adipose enrichment, and height exhibits enrichments in a variety of tissues in a pattern similar to previous analyses of this trait[18]. In addition, several of our results validate very recent findings from other genetic analyses: in particular, smoking status, years of education, BMI, and age at menarche show robust brain enrichments that recapitulate results from our previous analysis of genetic data together with chromatin data[6]. We also observe a cardiovascular enrichment for intracerebral hemorrhage, consistent with genetic evidence that this trait shares risk alleles with blood pressure levels[47], and a brain enrichment for epilepsy, consistent with parallel unpublished work[48].

In a data set with many tissues/cell types, related tissues will have highly overlapping gene sets. Because of this, and because we fit each tissue without adjusting for the other tissues analyzed, related tissues often appear enriched as a group. In this analysis, we are focused on identifying system-level enrichments, and so these correlated results do not limit interpretability. The following section similarly focuses on identifying system-level enrichments, while in later sections we focus on differentiating among related tissues/cell types within a system.

Comparison of expression-based approach to chromatin-based approach

We compared our approach to analyses of the same 48 diseases and traits using stratified LD score regression[6] in conjunction with chromatin data from the Roadmap Epigenomics project[2] (see URLs) instead of gene expression data. We constructed 397 cell-type-/tissue-specific annotations from narrow peaks in six chromatin marks—DNase hypersensitivity, H3K27ac, H3K4me3, H3K4me1, H3K9ac, and H3K36me3—each in a subset of 88 primary cell types/tissues. This analysis differed from our previous analysis of chromatin data[6] in

that we used more recently available data on a larger set of chromatin marks, we consistently used narrow peaks as processed by Roadmap for all marks, and we controlled not only for the union of annotations for each mark, but also for the average of annotations for each mark (Online Methods).

We analyzed GWAS summary statistics for the 48 traits, applying stratified LD score regression to each of the 397 tissue-specific chromatin-based annotations in turn. For 43 of the 48 traits, at least one tissue was significant at FDR<5% (**Figure S2** and **Table S6**). Averaging across the most significant annotation for each of these 43 traits, the tissue-specific chromatin annotation spanned 2.8% of the genome and explained 41% of the SNP-heritability (**Table S6**). Our results using chromatin data were generally concordant with the results of our gene expression analysis (**Figure 3a**). However, in many instances the analysis of chromatin data detected more enrichments and/or enrichments at higher significance levels than the analysis of gene expression data. There are two potential explanations for this. First, the set of tissues and cell types for which data is available is different for the two analyses; while in general gene expression is available in a wider variety of tissues and cell types (particularly for within-system analyses; see below), in some instances the most significantly enriched tissue in the chromatin analysis was not available in the GTEx or Franke lab data sets. For example, fetal lung was highly significantly enriched for lung capacity (FEV1/FVC) in our analysis of chromatin data, but there was no data on fetal lung in the GTEx or Franke lab data sets. Second, the enrichments were generally much larger for the chromatin-based annotations than for the gene expression-based annotations that we analyzed. However, the gene expression–based annotations were larger (i.e. spanned more of the genome) than the chromatin-based annotations and were comprised of larger regions, reducing the amount of LD between SNPs in the annotation and SNPs not in the annotation; this explains why LDSC-SEG was well-powered to identify much smaller enrichments.

We observed notable differences between the enrichments identified by the two approaches for migraine (**Figure 3b**). There is a long-standing scientific debate as to whether migraine has a primarily neurological or vascular basis[49], and a previous analysis of the migraine GWAS data (not restricted to any subtype) together with the GTEx gene expression data reported both cardiovascular and digestive/smooth muscle enrichments[16]. Our analysis of gene expression data did not identify any significant enrichments for this migraine data set, and identified a cardiovascular enrichment but no significant digestive/smooth muscle enrichment for migraine without aura. On the other hand, our analysis with chromatin data identified a significant neurological enrichment as well as quantitatively smaller and less significant cardiovascular and digestive/smooth muscle enrichments for the migraine data set, but identified only a borderline significant enrichment in fibroblasts for migraine without aura (**Figure 3b**). We hypothesize that this difference reflects a difference in power and in the cell types available in the two sources of data. For example, the top annotations for migraine in the chromatin analysis were neurospheres and fetal brain, neither of which was present in the gene expression data analyzed. Our results are consistent with the hypothesis that migraine without aura does indeed have a vascular component, and that another subtype of migraine may have a neurological basis which is sufficiently cell-type specific that the relevant cell types are not

represented in either the GTEx or Franke lab data sets. Our results demonstrate that for a multiple-tissue analysis, chromatin and gene expression data are complementary sources of data, and that it is of interest to test both gene expression annotations and chromatin annotations for enrichment, since there are diseases such as migraine and migraine without aura for which only one of the two types of data yields a significant enrichment.

A major advantage of gene expression data is that it is available at finer tissue/cell-type resolution within several systems. In the within-system analyses that follow, we analyzed gene expression data from tissues/cell types for which we did not have comparable chromatin data to investigate more detailed patterns of tissue/cell-type specificity. Thus, these analyses could not have been conducted using chromatin data.

Analysis of 13 brain-related traits using fine-scale brain expression data

We identified 13 traits with CNS enrichment at FDR<5% in our gene expression and/or chromatin analyses: schizophrenia, bipolar disorder, Tourette syndrome, epilepsy, generalized epilepsy, ADHD, migraine, depressive symptoms, BMI, smoking status, years of education, neuroticism, and systolic blood pressure. The nervous system has been implicated, either with genetic evidence or non-genetic evidence, for each of these traits[6,26,37,46,48–52], including systolic blood pressure, which is regulated in part via the autonomic nervous system[51].

We first investigated whether some brain regions are enriched over other brain regions for these traits. While the multiple-tissue analysis included annotations for many different brain regions, the gene sets for the different brain regions were often highly overlapping so that for many traits, many brain regions were identified as enriched. For example, nearly every brain region in either the GTEx or Franke lab data was found to be enriched at FDR<5% (**Figure 2**) in schizophrenia. To differentiate among brain regions, we restricted ourselves to gene expression data only from samples from the brain in the GTEx data. We computed t-statistics within the brain-only data set; e.g. we computed t-statistics for cortex vs. other brain regions instead of cortex vs. other tissues in GTEx, and we used these new t-statistics to construct and test gene sets as in the multiple-tissue analysis. Individual-level data was not available for the Franke lab data set, and thus we could not compute within-brain t-statistics for this data set.

An alternative approach would be to undertake a joint analysis of the original 13 annotations from the multiple-tissue analysis. However, joint analysis of 13 highly correlated annotations is likely to be underpowered, while re-computing t-statistics within the brain allows us to construct new annotations with lower correlations (**Figure S3**), increasing our power. Moreover, differential expression within the brain may allow us to isolate signals from cell types or processes that are unique to a single brain region, separately from the cell types or processes that are unique to the brain but shared among brain regions. Thus, we use differential expression within the brain, rather than joint analysis of the original annotations, to differentiate among brain regions.

The results of our analysis comparing brain regions are displayed in **Figure 4a**, **Figure S4a** and **Table S7a**. We identified significant enrichments in the cortex relative to other brain regions at FDR<5% for bipolar disorder, schizophrenia, depressive symptoms, and BMI, and in the striatum for migraine. These enrichments are consistent with our understanding of the biology of these traits[53–56], but to our knowledge have not previously been reported in any integrative analysis using genetic data. We also identified enrichments in cerebellum for bipolar disorder, years of education, smoking status, and BMI. However, we caution that differential gene expression in samples from different brain regions can reflect the cell type composition of these brain regions as well as their function. In particular, the cerebellum is known to have a very high concentration of neurons[57], and thus cerebellar enrichments could indicate either that the cerebellum is a region that is important in disease etiology, or that neurons are an important cell type.

To address the question of the relative importance of brain cell types, as opposed to brain regions, we analyzed the same set of traits using a publicly available data set of specifically expressed genes identified from different brain cell types purified from mouse forebrain[20]. The authors of this data set made lists of specifically expressed genes for each of the three brain cell types available, and these lists were all approximately the same size as the sets of specifically expressed genes in our previous analyses. We created annotations from these lists in the same way that we created annotations from the lists of top 10% expressed genes. The results of this analysis are displayed in **Figure 4b**, **Figure S4b** and **Table S7b**. We identified neuronal enrichments at FDR<5% for seven traits: bipolar disorder, schizophrenia, years of education, smoking status, BMI, neuroticism, and systolic blood pressure. The other cell types did not exhibit significant enrichment for any of the 13 brain-related traits. The enrichment of neurons for all four of the traits with enrichment in cerebellum in the brain-region analysis supports the hypothesis that analyses of brain regions may be confounded by cell-type composition. The enrichment for systolic blood pressure is consistent with the role of autonomic regulation of this trait[51].

To more precisely characterize the neuronal enrichments, we analyzed the seven traits with neuronal enrichment at FDR<5% using t-statistics computed by the PsychENCODE consortium[21] on differential expression in glutamatergic (excitatory) vs. GABAergic (inhibitory) neurons. The results are displayed in **Figure 4c**, **Figure S4c** and **Table S7c**; we used Bonferroni correction in this analysis, as we were testing only 7x2=14 hypotheses. For bipolar disorder, genes that are specifically expressed in GABAergic neurons exhibited heritability enrichment, while genes specific to glutamatergic neurons did not. This result supports the theory that pathology in GABAergic neurons can contribute causally to risk for bipolar disorder[58,59].

Analysis of 22 immune-related traits using immune cell expression data

We identified 22 traits with immune enrichment at FDR<5% in our gene expression and/or chromatin analyses. This includes many immunological disorders: celiac disease, Crohn's disease, inflammatory bowel disease, lupus, primary biliary cirrhosis, rheumatoid arthritis, type 1 diabetes, ulcerative colitis, asthma, eczema, and multiple sclerosis. It also includes Alzheimer's and Parkinson's diseases, which are neurodegenerative diseases with an

immune component previously identified from genetics[60,61], as well as several brain-related traits---ADHD, anorexia nervosa, bipolar disorder, schizophrenia, Tourette syndrome, and neuroticism---and HDL, LDL, and BMI. Several of the brain-related traits have been previously suggested to have an immune component[46,62,63], HDL and LDL have been linked to immune activation[64–66], and obesity, in addition to contributing to inflammation[67], can also be induced in mice through alterations of the immune system[68]. We investigated cell-type-specific enrichments for these traits in 292 immune cell types using gene expression data from the ImmGen project[22], which contains microarray data on these cell types from mice. This data set contains data for many immune cell types that are not available in the multiple-tissue analysis, and because we compute t-statistics within the data set---i.e., each immune cell vs. all other immune cells---the gene sets are less overlapping than those of immune cell types in the multiple-tissue analysis.

We identified enrichments at FDR<5% for 13 traits. Results are displayed in **Figure 5**, **Figure S5** and **Table S8**, and reveal highly trait-specific patterns of enrichment. For primary biliary cirrhosis, we identified an enrichment in B cells, consistent with literature on the importance of B cells for this trait[69,70]. Lupus and Alzheimer's disease both exhibit enrichment in myeloid cells. The Alzheimer's disease result is consistent with existing literature on the importance of the innate immune system in Alzheimer's disease etiology[71]. Asthma and eczema both exhibited enrichment in alpha beta T cells. Several subclasses of alpha beta T cells have been shown to be important in asthma[72]; to our knowledge, this result has not previously been reported in analyses of genetic data. Rheumatoid arthritis, Crohn's disease, inflammatory bowel disease, and multiple sclerosis all exhibited enrichments in a variety of cell types, consistent with complex etiologies for these diseases that involve many different immune cell types[73–75]. Schizophrenia and bipolar disorder both exhibited an enrichment in alpha beta T cells. Patients with bipolar disorder have been shown to have a reduction in types of alpha beta T cells, but have equal levels of B cells, NK cells, and monocytes compared to controls[76]. T cell levels have been shown to vary between schizophrenia cases and controls, but existing literature is not consistent in its description of the direction of effect[77]. Note that our analysis excludes the HLA region; a previous analysis of the HLA region for schizophrenia implicated the complement system through its role in synaptic pruning, a signal that is distinct from the signal we observe here[78]. Finally, we identified an enrichment in gamma delta T cells for BMI. While obesity is known to cause inflammation[67], and gamma delta T cells are known to be involved in obesity-related inflammation[79], gamma delta T cells have not to our knowledge been previously suggested to have a role in BMI etiology.

DISCUSSION

We have shown that applying stratified LD score regression to sets of specifically expressed genes identifies disease-relevant tissues and cell types. Our approach, LDSC-SEG, allows us to take advantage of the large amount of gene expression data available---including fine-grained data for which we do not currently have a comparable chromatin counterpart---to ask questions ranging in resolution from whether a trait is brain-related to whether excitatory or inhibitory neurons are more important for disease etiology. We identified many significant enrichments that confirm or extend our current understanding of biology,

including an enrichment of striatum for migraine, enrichment of GABAergic neurons for bipolar disorder, and an enrichment of myeloid cells for Alzheimer's disease. These results improve our understanding of these diseases, and highlight the power of GWAS as a source of biological insight.

There are several key differences between LDSC-SEG, which relies on gene expression data without genotypes or eQTLs, and approaches that require eQTL data[3,12]. First, our approach can be applied to expression data sets such as the Franke lab data set, the Cahoy data set, the PsychENCODE data set, and the ImmGen data set that do not have genotypes or eQTLs available (**Table 1**). Second, to our knowledge, no method based on eQTLs has been shown to consistently identify system-level enrichments such as brain enrichments for psychiatric traits and immune enrichment for immunological traits, as we do here[3,12]. Third, methods based on eQTLs require gene expression sample sizes that are large enough to detect eQTLs. In an analysis of data from the GTEx project, we determined that we could identify strong enrichments such as brain enrichment for schizophrenia with just one brain sample, though subtler enrichments had decreasing levels of significance as the gene expression data were down-sampled (**Figure S6**, Online Methods). Results from our analysis of ImmGen data, which has 2.8 samples per cell type on average, confirm that LDSC-SEG can identify significant enrichments even when the gene expression data has a small number of samples per tissue/cell type, in contrast to eQTL-based methods.

Our polygenic approach also differs from other gene expression-based approaches such as SNPsea[13,14] and DEPICT[15], which restrict their analyses to subsets of SNPs that pass a significance threshold. For comparison purposes, we repeated the multiple-tissue analysis using SNPsea and DEPICT. We also repeated the multiple-tissue analysis by analyzing our annotations using MAGMA, a recently developed gene set enrichment method[80] instead of stratified LD score regression[6]. Results are displayed in **Figures S7-S11** (see Online Methods). Many broad patterns were consistent across all approaches: immune enrichment for many immunological diseases, liver enrichment for lipid traits, adipose enrichment for BMI-adjusted waist-hip ratio, and enrichment in several tissues for height and heel T-score. However, there were also several discrepancies. First, SNPsea and DEPICT, the two approaches based on top SNPs, did not identify many of the CNS enrichments for brain-related traits identified by LDSC-SEG and by MAGMA. Second, DEPICT and MAGMA identified more enrichments than LDSC-SEG overall, including some enrichments with unclear relationships to known biology. We hypothesized that LDSC-SEG did not identify some of these enrichments because we jointly model our gene expression-based annotations with the many potential genomic confounders that are included in the baseline model (e.g. exons). We conducted simulations that confirmed that LDSC-SEG is the only approach that is well-powered to identify true enrichments for polygenic traits while avoiding genomic confounding (**Figure S12**; see Online Methods).

Our work has several limitations. First, our approach is fundamentally limited by the availability of gene expression data; for example, if the tissue/cell type that is most relevant for a disease occurs in a stage of development that has not been assayed, then we cannot identify enrichments in that tissue/cell type. Second, when analyzing gene expression data from different tissues, cell type composition can confound the analysis, as we demonstrated

in our comparison of brain regions. Third, tissues/cell types with similar gene expression profiles to a causal tissue/cell type will be identified as relevant to disease, just as SNPs in LD with a causal SNP will be identified as associated to disease in a GWAS; thus, significant tissues/cell types should be cautiously interpreted as the "best proxy" for the truly causal tissue/cell type, which may be unobserved. Finally, because our approach uses stratified LD score regression, it cannot be applied to custom array data, and it requires a sequenced reference panel that matches the population studied in the GWAS[6].

Our power to identify disease-relevant tissues and cell types will improve as GWAS sample sizes continue to grow and gene expression data is generated in new tissues and cell types. This will help advance our understanding of disease biology and lay the groundwork for future experiments exploring specific variants and mechanisms.

ACKNOWLEDGEMENTS

We are thankful to Tune Pers, Sam Riesenfeld, Rebecca Herbst, Adrian Veres, and Eran Hodis for helpful conversations. This research has been conducted using the UK Biobank Resource (Application Number: 16549). This research was funded by NIH grants R01 MH107649, R01 MH109978 and U01 CA194393. HKF is supported by the Fannie and John Hertz Foundation.

URLs

- LDSC software, including LDSC-SEG: https://github.com/bulik/ldsc.
- Gene sets and LD scores from this paper: https://data.broadinstitute.org/alkes group/LDSCORE/.
- GTEx: http://www.gtexportal.org.
- Franke lab data: https://data.broadinstitute.org/mpg/depict/depict_download/ tissue_expression.
- Cahoy et al. data: http://jneurosci.org/content/suppl/2008/01/03/ 28.1.264.DC1, see Tables S4-S6.
- PsychENCODE: https://www.synapse.org//#!Synapse:syn4921369/wiki/235539.
- ImmGen, https://www.immgen.org/.
- Roadmap Epigenomics: http://www.roadmapepigenomics. org.
- GERA data set (database of Genotypes and Phenotypes (dbGaP), phs000674.v1.p1): http://www-ncbi-nlm-nih-gov.libproxy.mit.edu/projects/gap/cgi-bin/study.cgi? study_id=phs000674.v1.p1.
- PLINK: https://www.cog-genomics.org/plink2.
- makegenes.sh: https://github.com/freeseek/gwaspipeline

References

1. The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489,** 57–74 (2012).

2. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518,** 317–330 (2015).
3. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science* **348,** 648–660 (2015).
4. Trynka, G. *et al.* Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45,** 124–130 (2013).
5. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518,** 337–343 (2014).
6. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47,** 1228–1235 (2015).
7. Li, Y. & Kellis, M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw627
8. Maurano, M. T. *et al.* Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. *Science* **337,** 1190–1195 (2012).
9. Pickrell, J. K. Joint Analysis of Functional Genomic Data and Genome-wide Association Studies of 18 Human Traits. *Am. J. Hum. Genet.* **94,** 559–573 (2014).
10. Kichaev, G. *et al.* Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genet.* **10,** e1004722 (2014).
11. Gusev, A. *et al.* Partitioning heritability of regulatory and cell-type-specific variants across 11 common diseases. *Am. J. Hum. Genet.* **95,** 535–552 (2014).
12. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *bioRxiv* (2016).

13. 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).

14. 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).

15. Pers, T. H. *et al.* Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6,** 5890 (2015).

16. Gormley, P. *et al.* Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nat. Genet.* **48,** 856–866 (2016).

17. Fehrmann, R. S. N. *et al.* Gene expression analysis identifies global gene dosage sensitivity in cancer. *Nat. Genet.* **47,** 115–125 (2015).

18. Wood, A. R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat. Genet.* **46,** 1173–1186 (2014).

19. Locke, A. E. *et al.* Genetic studies of body mass index yield new insights for obesity biology. *Nature* **518,** 197–206 (2015).

20. Cahoy, J. D. *et al.* A Transcriptome Database for Astrocytes, Neurons, and Oligodendrocytes: A New Resource for Understanding Brain Development and Function. *J. Neurosci.* **28,** 264–278 (2008).

21. Akbarian, S. *et al.* The PsychENCODE project. *Nat. Neurosci.* **18,** 1707–1712 (2015).

22. Heng, T. S. P., Painter, M. W. & Immunological Genome Project Consortium. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9,** 1091–1094 (2008).

23. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).

24. Sudlow, C. *et al.* UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med* **12,** e1001779 (2015).

25. Okbay, A. *et al.* Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533,** 539–542 (2016).

26. Okbay, A. *et al.* Genetic variants associated with subjective well-being, depressive symptoms, and neuroticism identified through genome-wide analyses. *Nat. Genet.* **48,** 624–633 (2016).

27. Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466,** 707–713 (2010).

28. Schunkert, H. *et al.* Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43,** 333–338 (2011).

29. Morris, A. P. *et al.* Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat. Genet.* **44,** 981–990 (2012).

30. Manning, A. K. *et al.* A genome-wide approach accounting for body mass index identifies genetic variants influencing fasting glycemic traits and insulin resistance. *Nat. Genet.* **44,** 659–669 (2012).

31. Okada, Y. *et al.* Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506,** 376–381 (2013).

32. Jostins, L. *et al.* Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491,** 119–124 (2012).

33. Bradfield, J. P. *et al.* A Genome-Wide Meta-Analysis of Six Type 1 Diabetes Cohorts Identifies Multiple Associated Loci. *PLOS Genet* **7,** e1002293 (2011).

34. Dubois, P. C. A. *et al.* Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42,** 295–302 (2010).

35. Bentham, J. *et al.* Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat. Genet.* **47,** 1457–1464 (2015).

36. Cordell, H. J. *et al.* International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nat. Commun.* **6,** 8019 (2015).

37. Anttila, V. *et al.* Analysis of shared heritability in common disorders of the brain. *bioRxiv* 048991 (2016).

38. Lambert, J.-C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45,** 1452–1458 (2013).

39. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet.* **45,** 984–994 (2013).

40. International League Against Epilepsy Consortium on Complex Epilepsies. Genetic determinants of common epilepsies: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **13,** 893–903 (2014).

41. Woo, D. *et al.* Meta-analysis of genome-wide association studies identifies 1q22 as a susceptibility locus for intracerebral hemorrhage. *Am. J. Hum. Genet.* **94,** 511–521 (2014).

42. Traylor, M. *et al.* Genetic risk factors for ischaemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol.* **11,** 951–962 (2012).

43. Gormley, P. *et al.* Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nat. Genet.* **48,** 856–866 (2016).

44. Patsopoulos, N. A. *et al.* Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann. Neurol.* **70,** 897–912 (2011).

45. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46,** 989–993 (2014).

46. Ripke, S. *et al.* Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511,** 421–427 (2014).

47. Falcone, G. J. *et al.* Burden of risk alleles for hypertension increases risk of intracerebral hemorrhage. *Stroke J. Cereb. Circ.* **43,** 2877–2883 (2012).

48. Backenroth, D. *et al.* Tissue-specific functional effect prediction of genetic variation and applications to complex trait genetics. *bioRxiv* (2016).

49. Tfelt-Hansen, P. C. & Koehler, P. J. One hundred years of migraine research: major clinical and scientific observations from 1910 to 2010. *Headache* **51,** 752–778 (2011).

50. Timothy E. Wilens, Joseph Biederman & Spencer, and T. J. Attention Deficit/Hyperactivity Disorder Across the Lifespan. *Annu. Rev. Med.* **53,** 113–131 (2002).

51. Lilly, L. S. *Pathophysiology of Heart Disease: A Collaborative Project of Medical Students and Faculty*. (Lippincott Williams & Wilkins, 2011).

52. Davis, L. K. *et al.* Partitioning the Heritability of Tourette Syndrome and Obsessive Compulsive Disorder Reveals Differences in Genetic Architecture. *PLOS Genet* **9,** e1003864 (2013).

53. Hanford, L. C., Nazarov, A., Hall, G. B. & Sassi, R. B. Cortical thickness in bipolar disorder: a systematic review. *Bipolar Disord.* **18,** 4–18 (2016).

54. Callicott, J. H. *et al.* Physiological Dysfunction of the Dorsolateral Prefrontal Cortex in Schizophrenia Revisited. *Cereb. Cortex* **10,** 1078–1092 (2000).

55. Medic, N. *et al.* Increased body mass index is associated with specific regional alterations in brain structure. *Int. J. Obes.* **40,** 1177–1182 (2016).

56. Maleki, N. *et al.* Migraine attacks the Basal Ganglia. *Mol. Pain* **7,** 71 (2011).

57. Herculano-Houzel, S. & Lent, R. Isotropic Fractionator: A Simple, Rapid Method for the Quantification of Total Cell and Neuron Numbers in the Brain. *J. Neurosci.* **25,** 2518–2521 (2005).

58. Sakai, T. *et al.* Changes in density of calcium-binding-protein-immunoreactive GABAergic neurons in prefrontal cortex in schizophrenia and bipolar disorder. *Neuropathology* **28,** 143–150 (2008).

59. Benes, F. M. & Berretta, S. GABAergic Interneurons: Implications for Understanding Schizophrenia and Bipolar Disorder. *Neuropsychopharmacology* **25,** 1–27 (2001).

60. Gjoneska, E. *et al.* Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518,** 365–369 (2015).

61. Gagliano, S. A. *et al.* Genomics implicates adaptive and innate immunity in Alzheimer's and Parkinson's. *bioRxiv* (2016). doi:10.1101/059519

62. Rege, S. & Hodgkinson, S. J. Immune dysregulation and autoimmunity in bipolar disorder: Synthesis of the evidence and its clinical application. *Aust. N. Z. J. Psychiatry* **47,** 1136–1151 (2013).

63. Elamin, I., Edwards, M. J. & Martino, D. Immune dysfunction in Tourette syndrome. *Behav. Neurol.* **27,** 23–32 (2013).

64. Jin, W., Millar, J. S., Broedl, U., Glick, J. M. & Rader, D. J. Inhibition of endothelial lipase causes increased HDL cholesterol levels in vivo. *J. Clin. Invest.* **111,** 357–362 (2003).

65. Broedl, U. C. *et al.* Endothelial lipase promotes the catabolism of ApoB-containing lipoproteins. *Circ. Res.* **94,** 1554–1561 (2004).

66. Feingold, K. R. & Grunfeld, C. The role of HDL in innate immunity. *J. Lipid Res.* **52,** 1–3 (2011).

67. Hotamisligil, G. S. Inflammation and metabolic disorders. *Nature* **444,** 860–867 (2006).

68. Zlotnikov-Klionsky, Y. *et al.* Perforin-Positive Dendritic Cells Exhibit an Immuno-regulatory Role in Metabolic Syndrome and Autoimmunity. *Immunity* **43,** 776–787 (2015).

69. Dhirapong, A. *et al.* B cell depletion therapy exacerbates murine primary biliary cirrhosis. *Hepatol. Baltim. Md* **53,** 527–535 (2011).

70. Zhang, J. *et al.* Ongoing activation of autoantigen-specific B cells in primary biliary cirrhosis. *Hepatol. Baltim. Md* **60,** 1708–1716 (2014).

71. Heneka, M. T., Golenbock, D. T. & Latz, E. Innate immunity in Alzheimer's disease. *Nat. Immunol.* **16,** 229–236 (2015).

72. Lloyd, C. M. & Hessel, E. M. Functions of T cells in asthma: more than just TH2 cells. *Nat. Rev. Immunol.* **10,** (2010).

73. Müller-Ladner, U., Pap, T., Gay, R. E., Neidhart, M. & Gay, S. Mechanisms of disease: the molecular and cellular basis of joint destruction in rheumatoid arthritis. *Nat. Clin. Pract. Rheumatol.* **1,** 102–110 (2005).

74. Xavier, R. J. & Podolsky, D. K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448,** 427–434 (2007).

75. Sospedra, M. & Martin, R. Immunology of Multiple Sclerosis. *Annu. Rev. Immunol.* **23,** 683–747 (2005).

76. Barbosa, I. G., Machado-Vieira, R., Soares, J. C. & Teixeira, A. L. The immunology of bipolar disorder. *Neuroimmunomodulation* **21,** 117–122 (2014).

77. Steiner, J. *et al.* Acute schizophrenia is accompanied by reduced T cell and increased B cell immunity. *Eur. Arch. Psychiatry Clin. Neurosci.* **260,** 509–518 (2010).

78. Sekar, A. *et al.* Schizophrenia risk from complex variation of complement component 4. *Nature* **530,** 177–183 (2016).

79. Mehta, P., Nuotio-Antar, A. M. & Smith, C. W. γδ T cells promote inflammation and insulin resistance during high fat diet-induced obesity in mice. *J. Leukoc. Biol.* **97,** 121–134 (2015).

80. 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).

Name	Organism	Tissues/cell types	Technology
GTEx[3]	Human	53 tissues/cell types	RNA-seq
Franke lab[15,17]	Human/mouse/rat	152 tissues/cell types	Array
Cahoy[20]	Mouse	3 brain cell types	Array
PsychENCODE[21]	Human	2 neuronal cell types	RNA-seq
ImmGen[22]	Mouse	292 immune cell types	Array

Table 1: List of gene expression data sets used in this study. We analyzed five gene expression data sets: two (GTEx and Franke lab) containing a wide range of tissues and three (Cahoy, PsychENCODE, ImmGen) with more detailed information about a particular tissue.

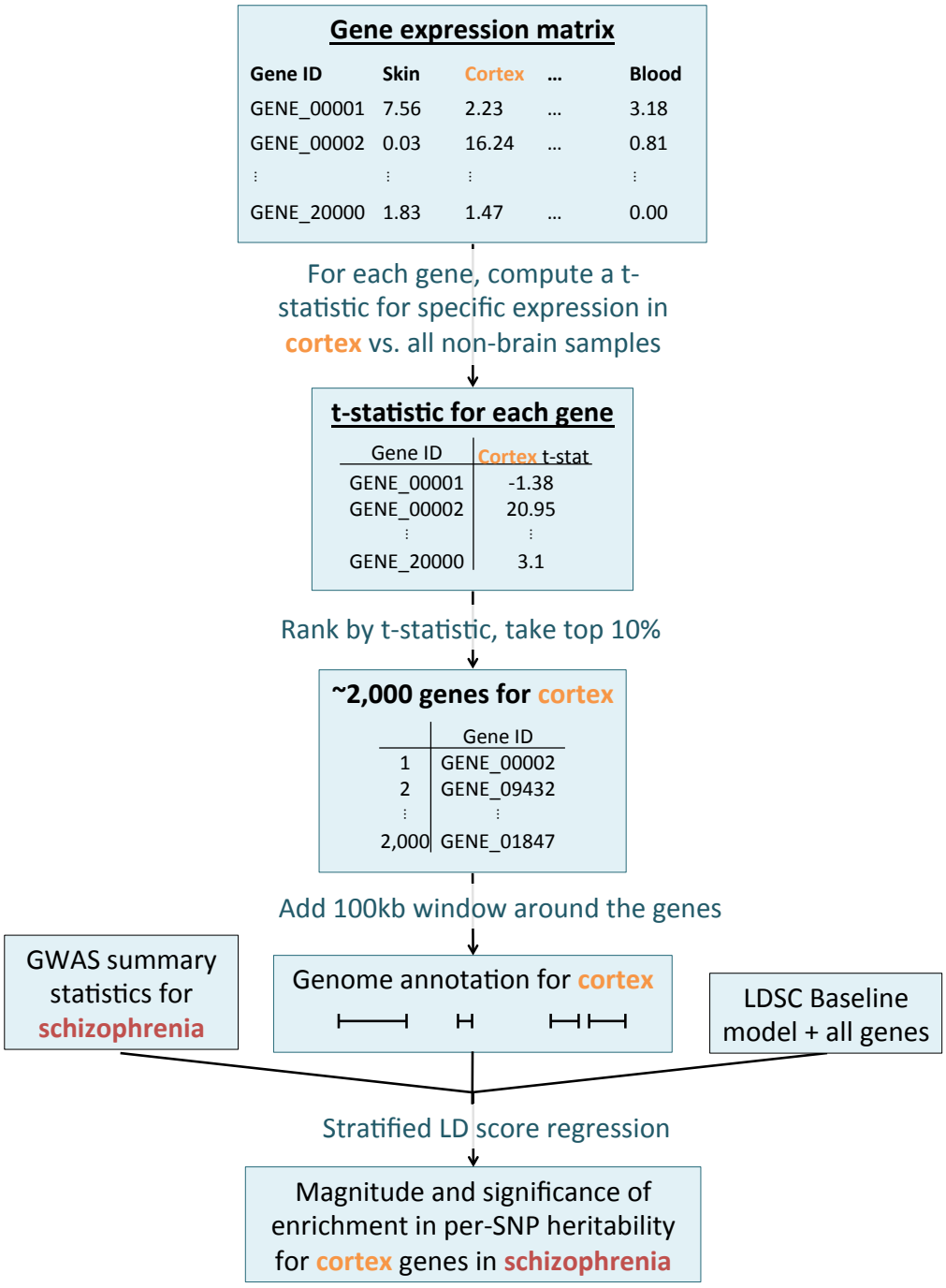

Figure 1: Overview of the approach. For each tissue in our gene expression data set, we compute t-statistics for differential expression for each gene. We then rank genes by t-statistic, take the top 10% of genes, and add a 100kb window to get a genome annotation. We use stratified LD score regression[6] to test whether this annotation is significantly enriched for per-SNP heritability, conditional on the baseline model[6] and the set of all genes.

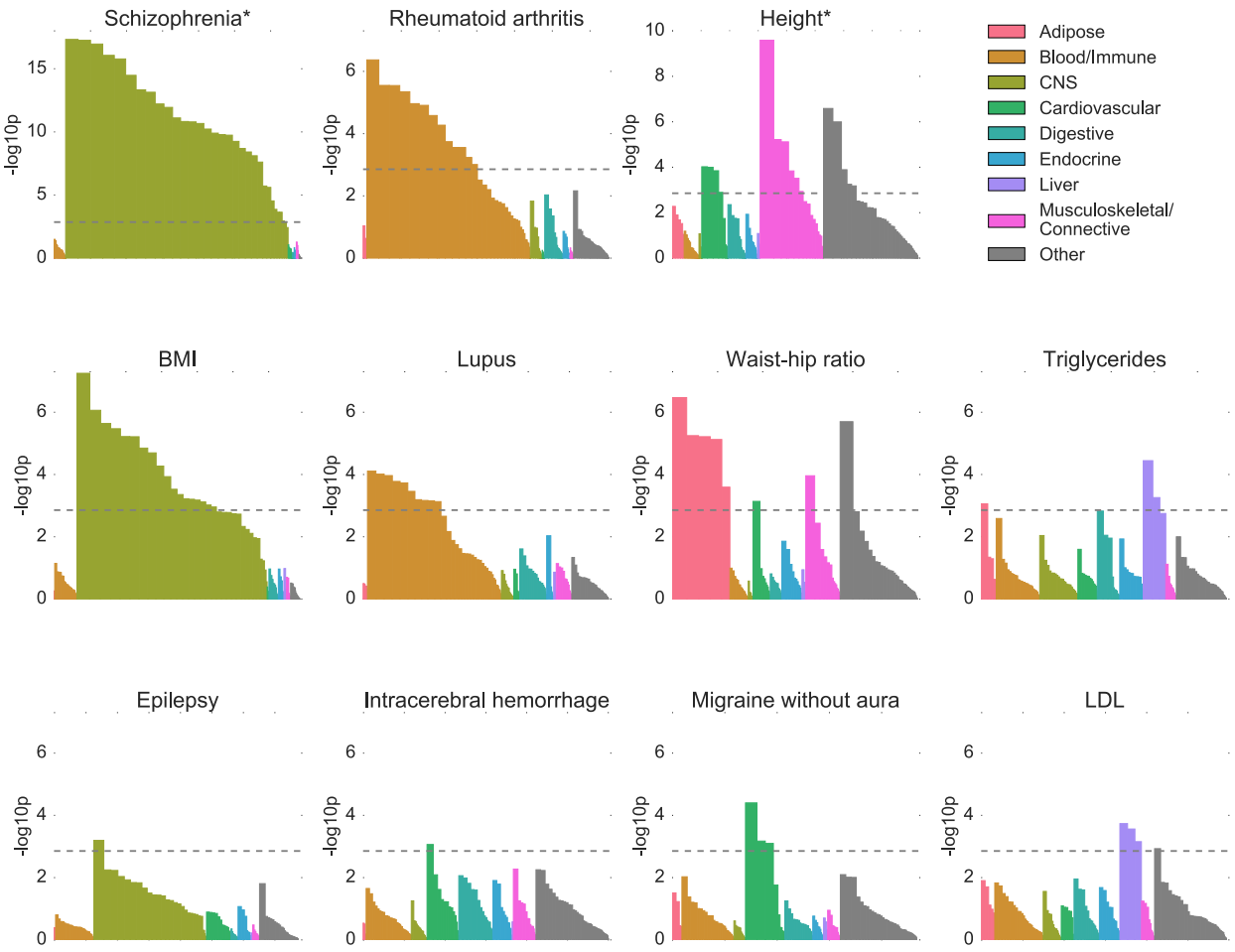

Figure 2: Results of multiple-tissue analysis for selected traits. Results for the remaining traits are displayed in **Figure S1**. Each bar represents a tissue/cell type from either the GTEx data set or the Franke lab data set. The width of each bar is proportional to its height, for easier visualization. *: y-axis has been rescaled to fit the data. The dashed line represents the FDR<5% cutoff, $-\log_{10}(P)=2.86$. Numerical results are reported in **Table S5**.

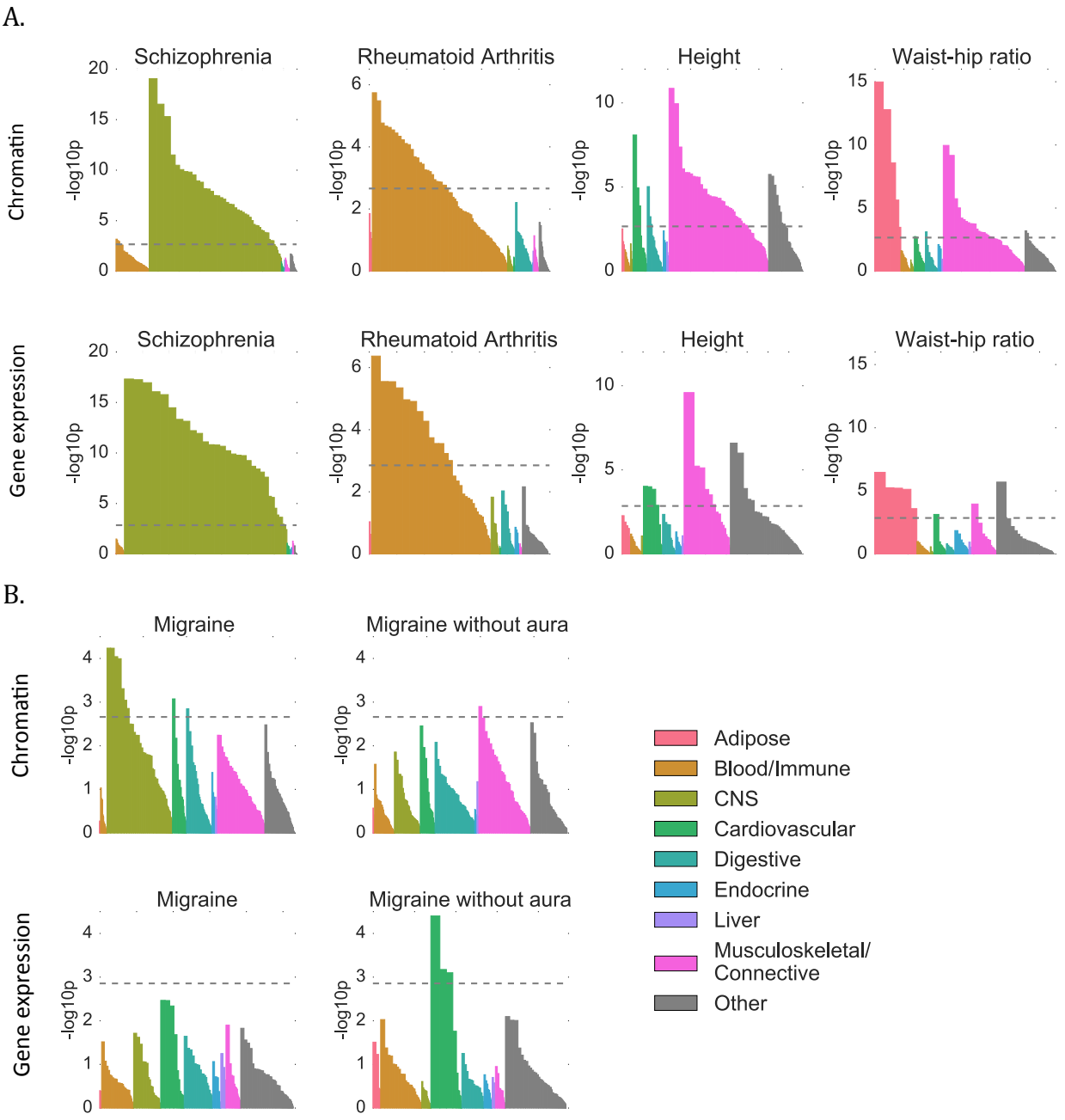

Figure 3: Comparison of chromatin and gene expression results. (A) Results using chromatin data from Roadmap (top) and results from the multiple-tissue gene expression analysis (bottom), for selected traits. Results using chromatin data from Roadmap for the remaining traits are displayed in **Figure S2**, with numerical results in **Table S6**. For the chromatin results, each bar represents a track of narrow peaks for H3K4me3, H3K4me1, H3K9ac, H3K27ac, H3K36me3, or DHS in a single cell type/tissue. The width of each bar is proportional to its height, for easier visualization. The dashed line represents the FDR<5% cutoff, $-\log_{10}(P)=2.85$ (chromatin) or $-\log_{10}(P)=2.99$ (gene expression). (B) Results using chromatin and gene expression data for migraine and migraine without aura.

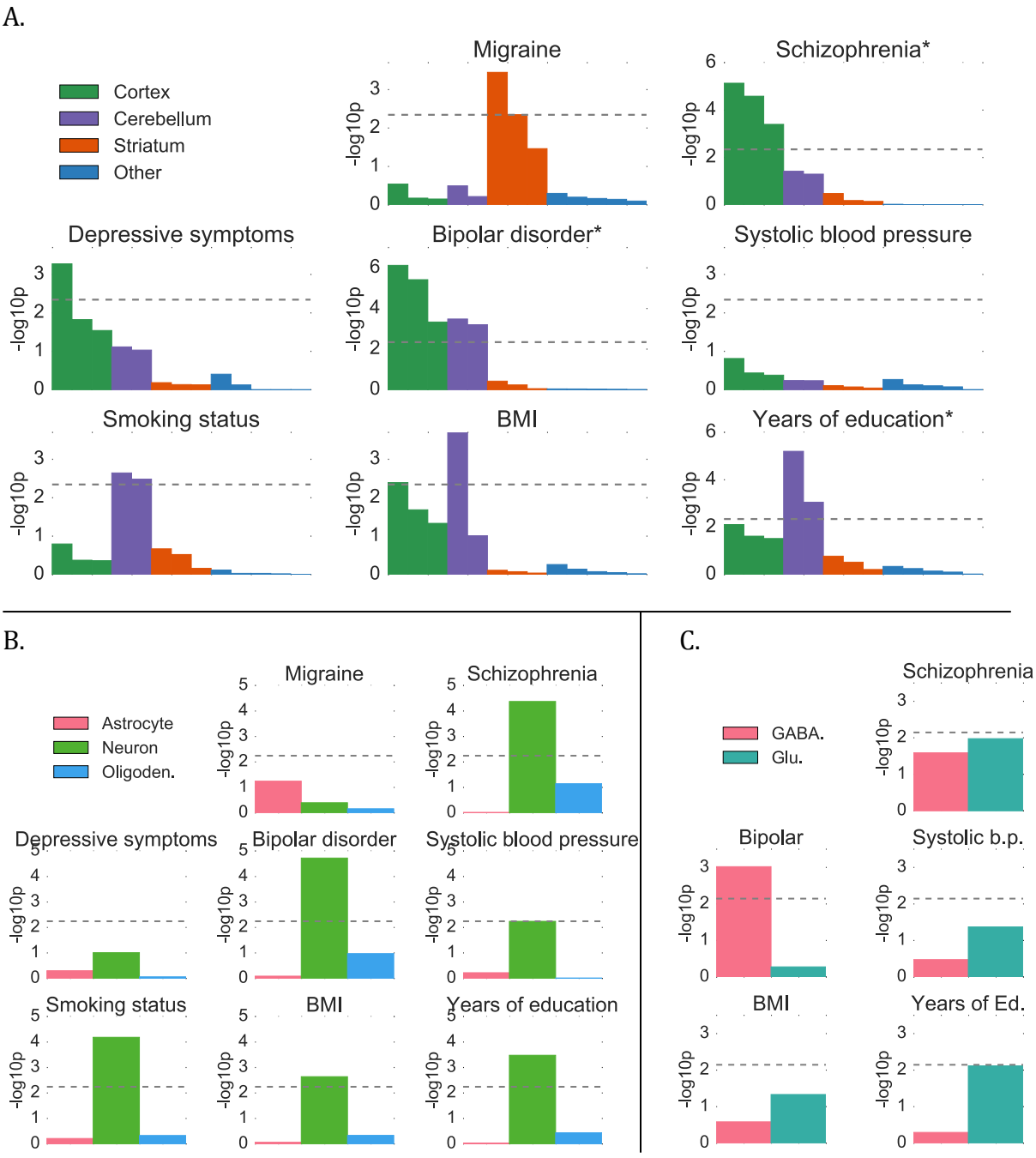

Figure 4: Results of brain analysis for selected traits. Results for remaining traits are displayed in **Figure S4**, with numerical results reported in **Table S7**. (A) Results from within-brain analysis of 13 brain regions in GTEx, classified into four groups, for eight of 13 brain-related traits. *: y-axis has been rescaled to fit the data. The dashed line represents the FDR<5% cutoff, $-\log_{10}(P)=2.35$. (B) Results from the data of Cahoy et al. on three brain cell types for eight of 13 brain-related traits. The dashed line represents the FDR<5% cutoff, $-\log_{10}(P)=2.24$. (C) Results from PyschENCODE data on two neuronal subtypes for five of seven neuron-related traits. The dashed line represents the cutoff for Bonferroni significance, $-\log_{10}(P)=2.72$.

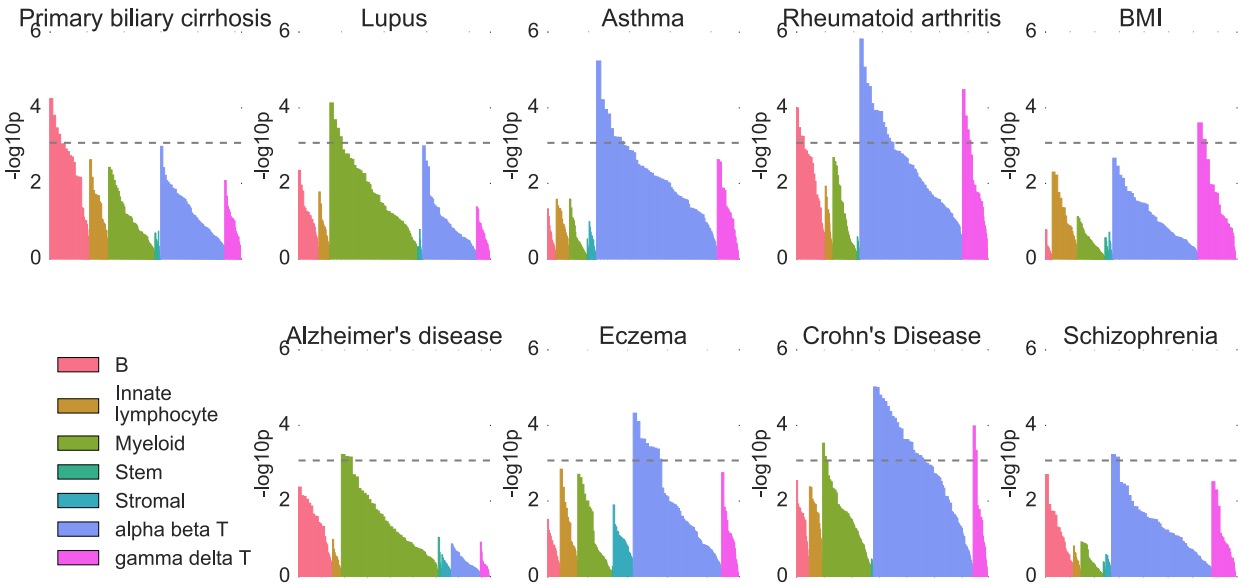

Figure 5: Results of ImmGen analysis for selected traits. Results for the remaining traits are displayed in **Figure S5**. The width of each bar is proportional to its height, for easier visualization. The dashed line represents the FDR<5% cutoff, $-\log_{10}(P)=3.08$. Numerical results are reported in **Table S8**.

ONLINE METHODS

Computing t-statistics. When computing the t-statistic of each gene for a focal tissue, we excluded all samples from the same tissue category (see "Tissue categories and covariates" below). For example, when computing the t-statistic of specific expression for each gene in cortex using GTEx data, we compared expression in cortex samples to expression in all other samples, excluding other brain regions. We chose to exclude other brain regions because we wanted to include genes that are more highly expressed in brain tissues than in non-brain tissues, even if they are not specific to cortex within the brain. This procedure results in a higher correlation among the t-statistics for the different brain regions; in a separate analysis, we compute within-brain t-statistics to disentangle this signal.

Thus, for a focal tissue (e.g., cortex) in a larger tissue category (e.g., brain), we computed the t-statistic for gene g as follows. We first constructed a design matrix X where each row corresponds to a sample either in cortex or outside of the brain. The first column of X has a 1 for every cortex sample and a -1 for every non-brain sample. The remaining columns are an intercept and covariates (see "Tissue categories and covariates" below). The outcome Y in our model is expression. We fit this model via ordinary least squares, and compute a t-statistic for the first explanatory variable in the standard way:

$$t = \frac{(X^T X)^{-1} X^T Y[0]}{\sqrt{MSE \cdot (X^T X)^{-1}[0,0]}}$$

where MSE is the mean squared error of the fitted model; i.e.,

$$MSE = \frac{1}{N}(Y - X(X^T X)^{-1} X^T Y)\text{^}T(Y - X(X^T X)^{-1} X^T Y)$$

where N is the number of rows in X. This gives us a t-statistic for each gene for the focal tissue. We then select the top 10% of genes, add a 100kb window around their transcribed regions, and apply stratified LD score regression to the resulting genome annotations as described below.

Modifications of our approach. For some analyses, we modified our approach to constructing sets of specifically expressed genes to better take advantage of the data available.

- *Franke lab data set*. The values in the publicly available matrix are not a quantification of expression intensity, but rather a quantification of differential expression relative to other tissues in this data set. Thus, it was not appropriate to compute t-statistics in this data set. We used the original values in place of our t-statistics, then proceeded as described in **Figure 1**.
- *Cahoy data set*. The data set of Cahoy et al. had available sets of specifically expressed genes for the three cell types that each had between 1,700 and 2,100 genes. We took these to be the gene sets for the three cell types, then proceeded as in the standard approach, adding a 100kb window and applying stratified LD score regression.
- *PsychENCODE data set*. The PsychENCODE data set had available t-statistics for GABAergic neurons vs. Glutamatergic neurons. We used these t-statistics, rather than computing our own.

For the other data sets we analyzed (GTEx, GTEx brain regions, ImmGen), we used the approach described in **Figure 1**. We view it as an advantage of our method that it can be flexibly adapted to many different types of data.

Tissue categories and covariates.

- For the multiple-tissue GTEx analysis, we used the "SMTS" variable ("Tissue Type, area from which the tissue sample was taken") to define the tissue categories (**Table S2**). We used age and sex as covariates.
- For the analysis of GTEx brain regions, we set each tissue to be its own category, and we used age and sex as covariates.
- For the ImmGen analysis, we defined tissue categories using the classification on the main page of immgen.org of cell types into categories: B cells, gamma delta T cells, alpha beta T cells, innate lymphocytes, myeloid cells, stromal cells, and stem cells (**Table S8**). The classification at immgen.org also has a "T cell activation" category that we collapsed into the alpha beta T cell category because it had data on alpha beta T cells at different stages of activation. We did not have any covariates.
- For the Franke lab data set, Cahoy data set, and PsychENCODE data set, we did not compute t-statistics and so we did not have tissue categories or covariates (see "Modifications of our approach" above).

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.

Application of stratified LD score regression. Stratified LD score regression[6] is a method for partitioning heritability. Given (potentially overlapping) genomic annotations $C_1, \ldots, C_K$, one of which is the category of all SNPs, we model the causal effect of SNP j on phenotype Y as drawn from a distribution with mean 0 and variance

$$Var(\beta_i) = \sum_k \tau_k \mathbf{1}\{i \in C_k\}. \qquad (1)$$

(If the genomic annotations are real-valued rather than subsets of SNPs, we can replace $\mathbf{1}\{i \in C_k\}$ with any other function of the SNP indices[81].) We then model the phenotype Y as depending linearly on genotype: $Y = X \cdot \beta + \epsilon$, where X is a vector of SNP values for an individual, and each SNP has been standardized to mean 0 and variance 1 in the population. Because each SNP is standardized, and because β_i has mean zero, we can call $Var(\beta_i)$ the per-SNP heritability of SNP i. (Note that here, because we model β as random, our definition of heritability is different from definitions of heritability in which β is fixed, and so we are estimating a fundamentally different quantity than some other methods[82].)

Under this model, the expected marginal chi-square association statistic for SNP i reflects the causal contributions not only of SNP i but of SNPs in LD with SNP i. Specifically,

$$E[\chi_i^2] = 1 + Na + N\sum_k \tau_k \ell(i,k),$$

where N is the GWAS sample size, a is a constant that reflects population structure and other sources of confounding,[83] and $\ell(i,k)$ is the LD score of SNP i to category C_k, defined as $\ell(i,k) = \sum_j r^2(i,j)\mathbf{1}\{j \in C_k\}$, where $r^2(i,j)$ is the squared correlation between SNPs i and j in the population. To estimate the τ_k, we first estimate $\ell(i,k)$ from a reference panel, and we then perform weighted regression χ_i^2 on $N \cdot \ell(i,k)$, using a jackknife over blocks of SNPs to estimate standard errors.

The regression coefficient τ_k quantifies the importance of annotation C_k, correcting for all other annotations in the model; τ_k will equal zero if C_k is not enriched, will be negative if belonging to C_k decreases per-SNP heritability accounting for all other annotations included, and will be positive if belonging to C_k increases per-SNP heritability, accounting for all other factors. Thus, as in our previous cell-type-specific anlaysis[6], we compute P-values that test whether τ_k is positive. When reporting quantitative results, we normalize the coefficient τ_k by our estimate of the mean per-SNP heritability $\sum_i Var(\beta_i)/M$ to make it comparable across phenotypes. The normalized coefficient can be interpreted as the proportion by which the per-SNP heritability of an average SNP would increase if τ_k were added to it. In addition, it is possible to estimate the total heritability, defined as $\sum_i Var(\beta_i)$, as well as the heritability in category C_k, defined as $\sum_{i \in C_k} Var(\beta_i)$, by plugging estimates of τ_k into Equation (1), and to compare the proportion of heritability, $\sum_{i \in C_k} Var(\beta_i)/\sum_i Var(\beta_i)$, to the proportion of SNPs, $|C_k|/M$, where M is the total number of SNPs[6].

We analyzed autosomes only and excluded the HLA from all analyses. In each analysis, we jointly fit the following annotations:

1. The annotation created for our focal tissue by adding 100kb windows around the top 10% of genes ranked by t-statistic.
2. An identical annotation created for all genes included in the gene expression data set being analyzed.
3. The baseline model with 52 functional categories, described previously[6] and listed in **Table S1**.

Gene expression data: quality control and normalization.

- *GTEx data set.* We downloaded the RNA-seq read counts from GTEx v6p (see URLs), removed genes for which fewer than 4 samples had at least one read count per million, removed samples for which fewer than 100 genes had at least one read count per million, and applied TPM normalization[84]. We used the "SMTSD" variable ("Tissue Type, more specific detail of tissue type") to define our tissues (**Table S2**).
- *Franke lab data set*. We downloaded the publicly available gene expression data from the DEPICT website (see URLs). We determined that several pairs of tissues had values that were correlated at r^2>0.99, including several that had r^2=1. We pruned our data so

that no two tissues had r^2>0.99. Most of the closely correlated pairs were also biologically closely related so that the interpretation did not depend on which tissue we chose to keep (e.g., plasma and plasma cells, joint and joint capsule). For pairs of tissues where one tissue was more specific than the second, we kept the more specific pair (e.g., nose vs. nasal mucosa, quadriceps muscle vs. skeletal muscle). There were two clusters of highly correlated tissues for which we decided to remove the entire cluster, not keeping any of the tissues, because these clusters had very strong but biologically implausible correlations. The first such cluster was made up of eyelids, conjunctiva, anterior eye segment, tarsal bones, foot bones, and bones of the lower extremity. The second such cluster was made up of connective tissue, bone and bones, skeleton, and bone marrow. After pruning, this data set contained 152 tissues, listed in **Table S3**.

- *Cahoy et al. data set*. We downloaded sets of specifically expressed genes for each of the three cell types (see URLs). To obtain a list of all genes, we also downloaded a list of all genes that passed quality control in their analysis (Table S3b of Cahoy et al.). We mapped from mouse to human genes using orthologs from ENSEMBL (see URLs).
- *PsychENCODE data set*. We used the t-statistics released by the PsychENCODE consortium for differential expression in GABAergic vs. Glutamatergic neurons[21]. These t-statistics were computed using limma[85].
- *ImmGen data set.* We downloaded publicly available gene expression data from the ImmGen Consortium (see URLs). We used both Phase 1 (GSE15907) and Phase 2 (GSE37448) data. The data on GEO were on an exponential scale, so we log transformed the data and mapped to human genes using ENSEMBL orthologs. We tested each of the 297 cell types.

We modified the makegenes.sh script[86] (see URLs) for some of our data processing.

Chromatin analysis. We downloaded narrow peaks from the Roadmap Epigenomics consortium for DNase hypersensitivity and five activating histone marks: H3K27ac, H3K4me3, H3K4me1, H3K9ac, and H3K36me3 (see URLs). Each of these six features was present in a subset of the 88 primary cell types/tissues, for a total of 397 cell-type-/tissue-specific annotations. For each of these annotations, we tested for enrichment by adding the annotation to the baseline model (see **Table S1**), together with the union of cell-type-specific annotations within each mark and the average of cell-type-specific annotations within each mark. A positive regression coefficient for a tissue-/cell-type-specific annotation represents a positive contribution of the annotation to per-SNP heritability, conditional on the other annotations. We again computed a P-value to test whether the regression coefficient was positive.

Our analysis of chromatin in this work differs from our previous analysis of chromatin data[6] in three ways. First, we use a larger range of marks and tissues/cell types: every track available from the Roadmap Epigenomics website (see URLs) for any of six activating marks, H3K27ac, H3K4me1, H3K4me3, H3K9ac, H3K36me3, and DHS, in any of the 88 primary tissues and cell types available, for a total of 397 annotations. Second, we used narrow peaks from Roadmap for all of the marks. Previously, we analyzed H3K27ac data from one source[5] and H3K4me1, H3K4me3, and H3K9ac data from another source[4,11]; now that there is a single standard source with uniformly processed data for all marks of

interest, we have switched to using this data. Finally, we controlled more strictly for confounders by including the average across cell types of the cell-type-specific annotations for a given mark as an annotation in the model, so that annotations that tend to fall in areas that are more active overall are not falsely interpreted as cell-type-specific signal.

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 S6A**). 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 S6B**). 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[13,14] 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 S7**. 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[15] 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[15,17], 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 S8** and **S9**. 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[80], a gene set enrichment method that allows inclusion of a window around each gene and leverages all SNPs in the gene set (**Figure S10**). MAGMA and LDSC-SEG identified many of the same enrichments, but MAGMA identified several enrichments that LDSC-SEG did not. In simulations described below, we determined that MAGMA can report significant results in the absence of true enrichment due to uncorrected genomic confounding.

For comparison purposes, we report LDSC-SEG results for the multiple tissue analysis as a heatmap in **Figure S11**, in addition to the bar charts 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[87–89] 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[90] (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 S12**.

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). This is consistent with the fact MAGMA does not control for exon content.

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 only LDSC-SEG and SNPsea control type I error, and that of these two methods, LDSC-SEG is better powered for polygenic traits.

References

81. Gazal, S. *et al.* Linkage disequilibrium dependent architecture of human complex traits reveals action of negative selection. *bioRxiv* 082024 (2016). doi:10.1101/082024

82. Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am. J. Hum. Genet.* **99,** 139–153 (2016).

83. Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47,** 291–295 (2015).

84. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci. Theor. Den Biowissenschaften* **131,** 281–285 (2012).

85. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15,** R29 (2014).

86. Genovese, G. *et al.* Increased burden of ultra-rare protein-altering variants among 4,877 individuals with schizophrenia. *Nat. Neurosci.* **19,** 1433–1441 (2016).

87. 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).

88. 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).

89. 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).

90. Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4,** 7 (2015).