nature genetics

ANALYSIS

https://doi.org/10.1038/s41588-018-0081-4

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

Hilary K. Finucane[1,2,3*], Yakir A. Reshef[4], Verneri Anttila[1,5], Kamil Slowikowski[1,6,7], Alexander Gusev[3], Andrea Byrnes[1,5], Steven Gazal[3], Po-Ru Loh[3], Caleb Lareau[1,8], Noam Shoresh[1], Giulio Genovese[1], Arpiar Saunders[9], Evan Macosko[9], Samuela Pollack[3], The Brainstorm Consortium[10], John R. B. Perry[11], Jason D. Buenrostro[1,12], Bradley E. Bernstein[1,13], Soumya Raychaudhuri[1,7,14,15,16], Steven McCarroll[1,9], Benjamin M. Neale[1,5], and Alkes L. Price[1,3*]

We introduce an approach to identify disease-relevant tissues and cell types by analyzing gene expression data together with genome-wide association study (GWAS) summary statistics. Our approach uses stratified linkage disequilibrium (LD) score regression to test whether disease heritability is enriched in regions surrounding genes with the highest specific expression in a given tissue. We applied our approach to gene expression data from several sources together with GWAS summary statistics for 48 diseases and traits (average $N = 169{,}331$) and found significant tissue-specific enrichments (false discovery rate (FDR) < 5%) for 34 traits. In our analysis of multiple tissues, we detected a broad range of enrichments that recapitulated known biology. In our brain-specific analysis, significant enrichments included an enrichment of inhibitory over excitatory neurons for bipolar disorder, and excitatory over inhibitory neurons for schizophrenia and body mass index. Our results demonstrate that our polygenic approach is a powerful way to leverage gene expression data for interpreting GWAS signals.

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 that characterize tissue- and cell-type-specific activity have been analyzed together with GWAS data to identify disease-relevant tissues and cell types—including histone marks[4–8], DNase I–hypersensitive sites (DHS)[9–12], expression quantitative trait loci (eQTLs)[3,13] and gene expression data[14–17]. Of these data types, gene expression data (without genotypes or eQTLs) have the advantage of being available in the widest range of tissues and cell types. Previous studies have shown that gene expression data are informative for disease-relevant tissues and cell types, and these have led to biological insights about the diseases and traits studied[14–17]. 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 systematically from GWAS and gene expression data.

Here we applied stratified LD score regression[7], 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 169,331. We first analyzed two gene expression datasets[3,17,18] that contained a wide range of tissues to infer system-level enrichments. We then analyzed chromatin data from the Roadmap Epigenomics and ENCODE projects[1,2] across the same set of diseases and traits to validate these results. Finally, we analyzed gene expression datasets that allowed us to achieve higher resolution within a system[3,19–21] and identified enriched brain regions, brain cell types and immune cell types for several brain- and immune-related diseases and traits; we validated several of our immune enrichments using independent chromatin data. 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.

[1]Broad Institute of MIT and Harvard, Cambridge, MA, USA. [2]Department of Mathematics, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [4]Department of Computer Science, Harvard University, Cambridge, MA, USA. [5]Analytic and Translational Genetics Unit, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [6]Bioinformatics and Integrative Genomics, Harvard University, Cambridge, MA, USA. [7]Division of Genetics, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [8]Department of Biostatistics, Harvard T. H. Chan School of Public Health, Boston, MA, USA. [9]Department of Genetics, Harvard Medical School, Boston, MA, USA. [10]A list of members and affiliations appears in the Supplementary Note.[11]Medical Research Council (MRC) Epidemiology Unit, University of Cambridge School of Clinical Medicine, Institute of Metabolic Science, Cambridge Biomedical Campus, Cambridge, UK. [12]Harvard Society of Fellows, Harvard University, Cambridge, MA, USA. [13]Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA. [14]Division of Rheumatology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA. [15]Partners Center for Personalized Genetic Medicine, Boston, MA, USA. [16]Faculty of Medical and Human Sciences, University of Manchester, Manchester, UK.
*e-mail: finucane@broadinstitute.org; aprice@hsph.harvard.edu

Table 1 | List of gene expression datasets used in this study

Name	Organism	Tissue or cell type	Technology
GTEx[3]	Human	53 tissues or cell types	RNA-seq
Franke lab[17,18]	Human, mouse and rat	152 tissues or cell types	Array
Cahoy[19]	Mouse	3 brain cell types	Array
PsychENCODE[20]	Human	2 neuronal cell types	RNA-seq
ImmGen[21]	Mouse	292 immune cell types	Array

We analyzed five gene expression datasets: two (GTEx and Franke lab) that contained a wide range of tissues and three (Cahoy, PsychENCODE and ImmGen) with more detailed information about a particular tissue.

Results

Overview of methods. We analyzed the five gene expression datasets 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 followed the procedure described in Fig. 1. We began with a matrix of normalized gene expression values across genes, with samples from multiple tissues including the focal tissue. For each gene, we computed a *t*-statistic for specific expression in the focal tissue (Methods). We ranked all of the genes by their *t*-statistic and defined the 10% of genes with the highest *t*-statistic to be the gene set corresponding to the focal tissue; we called this the set of specifically expressed genes, but we note that this includes genes that are not only strictly specifically expressed (i.e., only expressed in the focal tissue) but also those that are weakly specifically expressed (i.e., have 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 (Methods). We added 100-kb windows on either side of the transcribed region of each gene in the set of specifically expressed genes to construct a genome annotation that corresponded to the focal tissue (the choice of the parameters 10% and 100-kb window is discussed in the Supplementary Note; our results are robust to these choices; see below). Finally, we applied stratified LD score regression[7] to GWAS summary statistics to evaluate the contribution of the focal genome annotation to trait heritability (Methods). We jointly modeled the annotation that corresponded to the focal tissue, a genome annotation that corresponded to all of the genes, as well as the 52 annotations in the 'baseline model'[7] (including genic regions, enhancer regions and conserved regions; see Supplementary Table 1). 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 (for example, 1000 Genomes[22]), and has been shown to produce robust results with properly controlled type I error[7]. We have released open-source software implementing our approach and have also released all of the genome annotations that were derived from the publicly available gene expression data we analyzed (see URLs). We refer to our approach as LD score regression applied to specifically expressed genes (LDSC-SEG).

Analysis of 48 complex traits across multiple tissues. We first analyzed two gene expression datasets—one from the Genotype–Tissue Expression (GTEx) project and another that we call the 'Franke lab' dataset—and we classified the 205 tissues and cell types in these datasets into nine categories for visualization (Supplementary Tables 2 and 3, and Methods). We analyzed GWAS summary statistics for 48 diseases and traits from the UK Biobank[23] (Methods), the Brainstorm Consortium[16,24–32] and publicly available sources[33–43] (with an average sample size of 169,331; Supplementary Table 4) by applying LDSC-SEG for each of the 205 specifically expressed gene annotations in turn. We excluded the human leukocyte antigen (HLA) region from all analyses due to its unusual genetic architecture and pattern of LD.

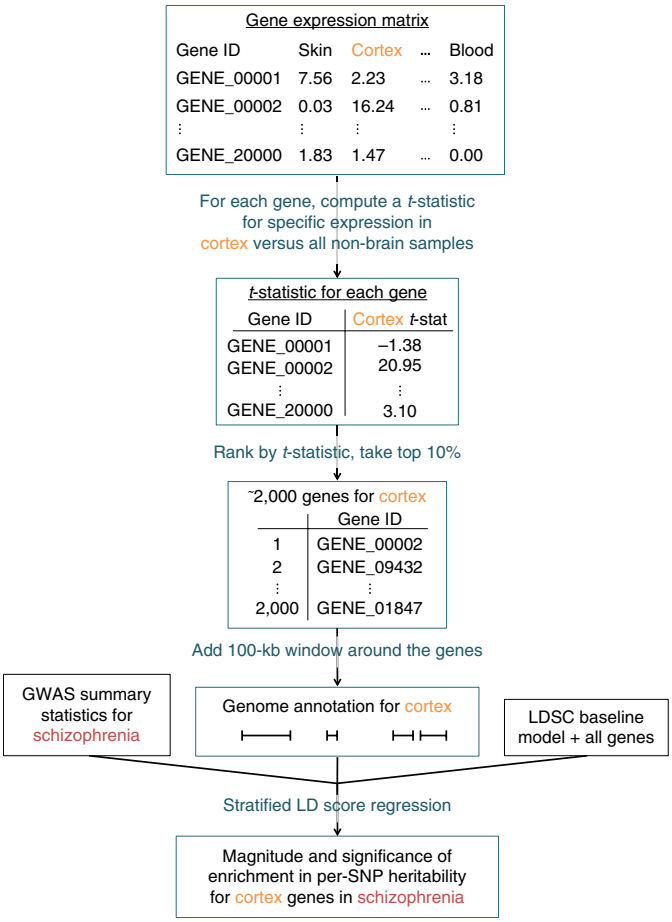

Fig. 1 | Overview of the approach. For each tissue in our gene expression dataset, we compute *t*-statistics for differential expression for each gene. We then rank genes by the *t*-statistic, take the top 10% of genes and add a 100-kb window to get a genome annotation. We use stratified LD score regression[7] to test whether this annotation is significantly enriched for per-SNP heritability, conditional on the baseline model[7] and the set of all genes.

For 34 of the 48 traits, at least one tissue was significant at FDR < 5% (Fig. 2, Supplementary Fig. 1 and Supplementary Tables 5 and 6). Several of our results recapitulated known biology: immunological traits exhibited immune cell–type enrichments, psychiatric traits exhibited strong brain-specific enrichments, low-density lipoprotein (LDL) and triglycerides exhibited liver-specific enrichments, body mass index (BMI)-adjusted waist–hip ratio exhibited adipose-specific enrichment, type 2 diabetes exhibited enrichment in the pancreas, and height exhibited enrichments in a variety of tissues in a pattern similar to those from previous analyses of this trait[44]. In addition, several of our results validated very recent findings from other genetic analyses; in particular, smoking status, years of education, BMI and age at menarche showed robust brain-specific enrichments that recapitulated results from our previous analysis of genetic data together with chromatin data[7]. Our results were robust to the choice of the percentage of genes used (10%)

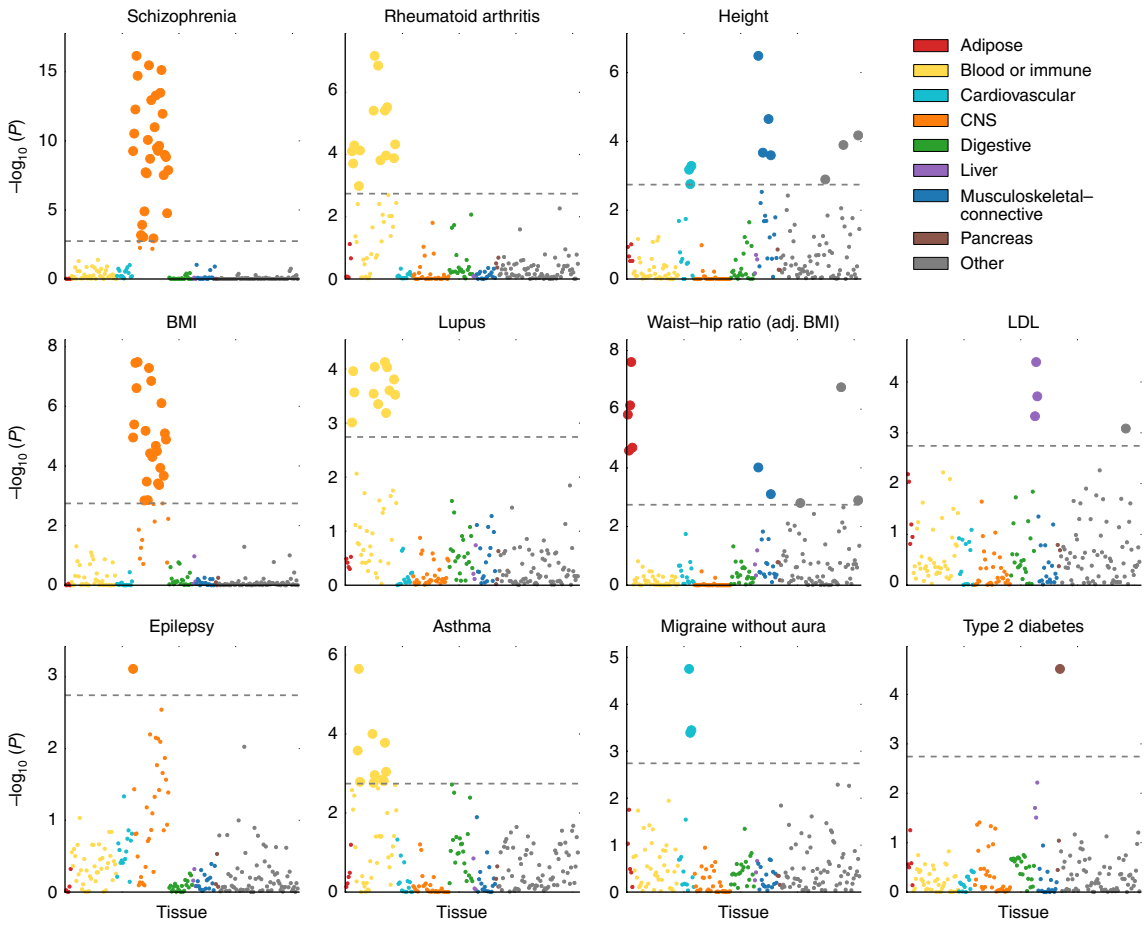

Fig. 2 | Results of the multiple-tissue analysis for selected traits. Each circle represents a tissue or cell type from either the GTEx dataset (total $N = 8{,}550$) or the Franke lab dataset (total $N = 37{,}427$). Results for the remaining traits are displayed in Supplementary Fig. 1. Large circles pass the cutoff of FDR < 5% at $-\log_{10}(P) = 2.75$. GWAS data are described in Supplementary Table 4, gene expression data are described in the Methods and in Supplementary Tables 2 and 3, and the statistical method is described in the "Overview of methods" section above and the Methods. Numerical results are reported in Supplementary Table 6.

and to the size of the window used (100 kb) (Supplementary Fig. 2). We assessed correlations in enrichment patterns for pairs of traits (Methods) and found large and significant (FDR < 5%) correlations among many brain-related phenotypes, among many immune-related phenotypes, and among a third set of phenotypes including height and blood pressure that tended to have enrichments in the musculoskeletal–connective, cardiovascular and other categories (Supplementary Fig. 3). The most significant annotation for each of these 34 traits spanned 11–23% (mean 16%) of the genome and explained 21–62% (mean 36%) of SNP heritability, with enrichments varying from 1.4× to 4.7× (mean 2.3×) (Supplementary Table 5).

Because related tissues have highly overlapping gene sets and we fit each tissue without adjusting for the other tissues, related tissues often appear enriched as a group. In this analysis and the analysis in the next section, both of which were focused on identifying system-level enrichments, these correlated results did not limit interpretability. In later sections, we focused on differentiating among related tissues or cell types within a system. We note also that the correlation structure among annotations can lead to a distribution of *P* values that is highly non-uniform (Methods).

Validation using independent chromatin data. We analyzed the same 48 diseases and traits using stratified LD score regression[7] in conjunction with chromatin data from the Roadmap Epigenomics and ENCODE projects[1,2] (see URLs) instead of from gene expression data, with three goals: (i) to validate the results from our analysis of gene expression data using a different type of data from an independent source, (ii) to identify new enrichments using chromatin data that we did not observe using gene expression data, and (iii) to compare enrichments from the two types of data. The ENCODE data we used were from a subproject called EN-TEx, which includes epigenetic data on a set of tissues that match a subset of the tissues from the GTEx project but are from different donors. In total, we analyzed 489 tissue-specific chromatin-based annotations from peaks for six epigenetic marks (Methods).

We considered two types of validation for the results of the multiple-tissue analysis of gene expression described above: validation at the system level and validation at the tissue or cell-type level. For validation at the system level, we classified the top tissue or cell type for each trait with a significant enrichment into one of nine systems (Methods), and we considered an enrichment to be validated if a tissue or cell type from the same system passed FDR < 5% for the same phenotype in the chromatin analysis. For validation at the tissue or cell-type level, we analyzed only the 27 tissues present in both the GTEx and EN-TEx datasets, and we considered an enrichment of a tissue in GTEx to be validated if any mark in the same tissue in EN-TEx passed FDR < 5% for the same phenotype. The top enrichment from our multi-tissue analysis of gene expression was validated at the system level for 33 of 34 phenotypes (Fig. 3a and

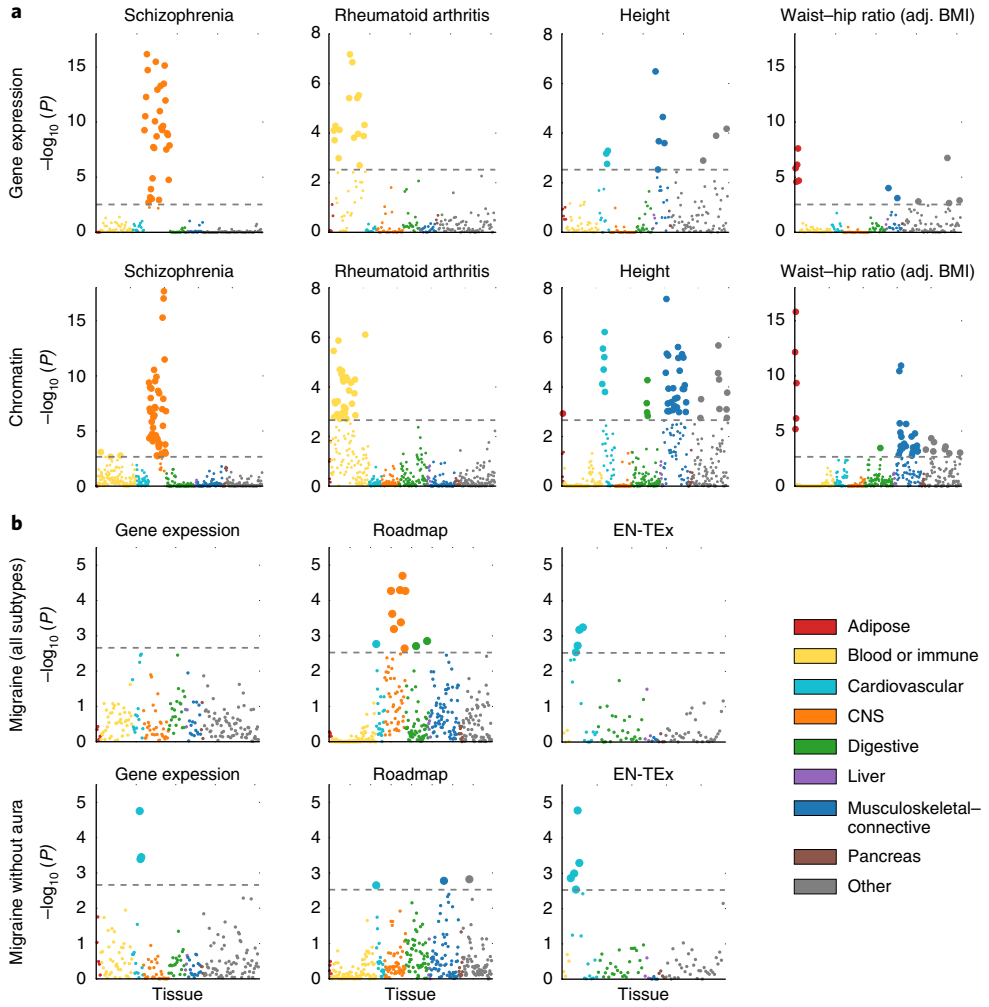

Fig. 3 | Validation of gene expression results with chromatin data. a, Examples of validation using chromatin data (bottom) of results from gene expression data (top), for selected traits. Results using chromatin data for all traits are displayed in Supplementary Fig. 5, with numerical results presented in Supplementary Table 7. For the chromatin results, each circle represents a track of peaks for trimethylated lysine 4 on histone H3 (H3K4me3), monomethylated lysine 4 on histone H3 (H3K4me1), acetylated lysine 9 on histone H3 (H3K9ac), acetylated lysine 27 on histone H3 (H3K27ac), trimethylated lysine 36 on histone H3 (H3K36me3) or DHS in a single tissue or cell type. **b**, Results—using gene expression (including GTEx), Roadmap and EN-TEx data—for migraine (all subtypes) and migraine without aura. For each plot, the large circles pass the cutoff of FDR < 5% at either $-\log_{10}(P) = 2.85$ (chromatin) or $-\log_{10}(P) = 2.75$ (gene expression). GWAS data are described in Supplementary Table 4, gene expression data and chromatin data are described in the Methods and in Supplementary Tables 2, 3 and 7, and the statistical method is described in the "Overview of methods" section above and the Methods.

Supplementary Table 5), and the top enrichment of a tissue or cell type shared between GTEx and EN-TEx was validated at the tissue or cell-type level for 13 of 20 phenotypes, which increased to 16 with a more lenient definition (Supplementary Table 5 and Methods). In many instances, the analysis of chromatin data detected a greater number of enrichments, larger enrichments and/or enrichments at higher significance levels than the analysis of gene expression data, although this was not always the case (Supplementary Figs. 4 and 5, Supplementary Table 7 and Methods). The enrichment correlations in this analysis showed a similar pattern to that of the gene expression analysis above (Supplementary Fig. 6).

There is a long-standing scientific debate as to whether migraine has a primarily neurological or vascular basis[45]. We analyzed GWAS summary statistics for migraine with aura, migraine without aura, and migraine (all subtypes)[16]. The migraine (all subtypes) dataset contained the datasets for migraine with aura and for migraine without aura, as well as data for a large number of additional subjects whose subtype was unknown. We found cardiovascular enrichments for migraine without aura with gene expression data, and for migraine without aura and migraine (all subtypes) with EN-TEx data, consistent with previous work[16] (Fig. 3b). Our analysis of Roadmap data, however, yielded qualitatively different results—the strongest enrichment for migraine (all subtypes) was a neurological enrichment. The top two annotations were neurospheres and fetal brain, neither of which was present in the gene expression data we analyzed nor in the EN-TEx dataset. The correlation in enrichments between migraine (all subtypes) and migraine without aura in the gene expression analysis was estimated to be 0.48 (s.e. 0.15), whereas it was estimated to be 0.60 (s.e. 0.13) in the chromatin data. 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 that is sufficiently cell-type specific that the relevant cell types are not represented in either the GTEx or Franke lab datasets. These results highlight the importance of having as many tissues and cell types as possible represented in a multiple-tissue analysis.

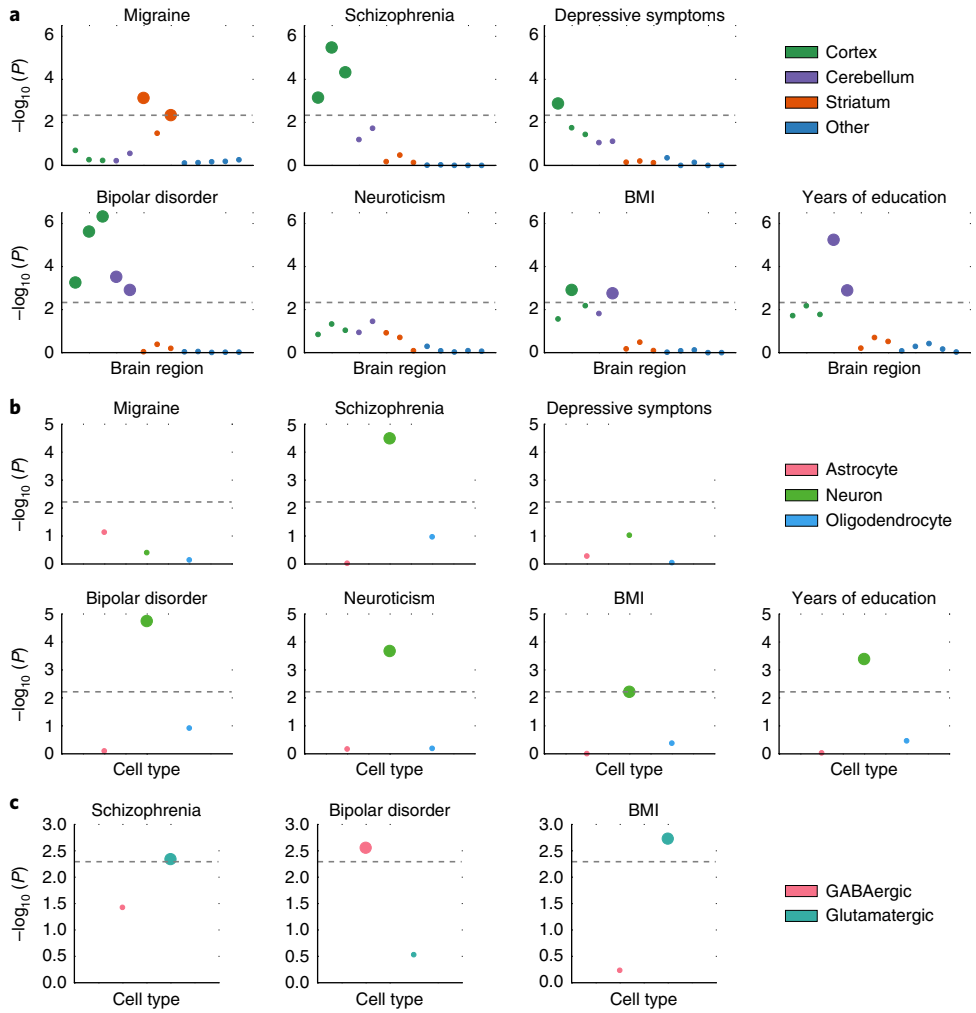

Fig. 4 | Results of the brain analysis for selected traits. Numerical results for all traits are reported in Supplementary Table 8. **a**, Results from the within-brain analysis of 13 brain regions in GTEx, classified into four groups, for 7 of 12 brain-related traits. Large circles passed the cutoff of FDR < 5% at $-\log_{10}(P) = 2.34$. **b**, Results from the data of Cahoy et al.[19] on three brain cell types for 7 of 12 brain-related traits. Large circles passed the cutoff of FDR < 5% at $-\log_{10}(P) = 2.22$. **c**, Results from PyschENCODE data on two neuronal subtypes (GABAergic and glutamatergic) for three of five neuron-related traits. Large circles passed the Bonferroni significance threshold in this analysis ($-\log_{10}(P) = 2.06$). GWAS data are described in Supplementary Table 4, gene expression data are described in the Methods and in Supplementary Table 8, and the statistical method is described in the "Overview of methods" section above and the Methods.

A major advantage of gene expression data is that it is available at finer tissue and cell-type resolution within several systems. In the within-system analyses that follow, we investigated these finer patterns of tissue and cell-type specificity.

Analysis of 12 brain-related traits using fine-scale brain expression data. We identified 12 traits with central nervous system (CNS) enrichment at FDR < 5% in our gene expression and/or chromatin analyses (Methods). We first investigated whether some brain regions were enriched relative to other brain regions for these traits using gene expression data from GTEx (Supplementary Fig. 7 and Methods). The results are displayed in Fig. 4a and Supplementary Table 8a. 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[46–49] but to our knowledge have not previously been reported in any integrative analysis using genetic data. We also identified enrichments in the cerebellum for bipolar disorder, years of education 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[50], and thus cerebellar enrichments could indicate either that the cerebellum is a region important in disease etiology or that neurons are an important cell type. Although many pairs of phenotypes had high estimated enrichment correlations in this analysis, migraine tended to have low enrichment correlations with other phenotypes (Supplementary Fig. 8); for example, the estimated enrichment correlation between migraine and schizophrenia was 0.06 (s.e. 0.30), whereas the estimated enrichment correlation between bipolar disorder and schizophrenia was 0.96 (s.e. 0.05).

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 dataset of specifically expressed genes that were identified from different brain cell types purified from mouse forebrain[19] (Methods). The results of this analysis are displayed in Fig. 4b and Supplementary Table 8b. We identified neuronal enrichments at FDR < 5% for five traits: bipolar disorder, schizophrenia, years of education, BMI and neuroticism. The other cell types did not

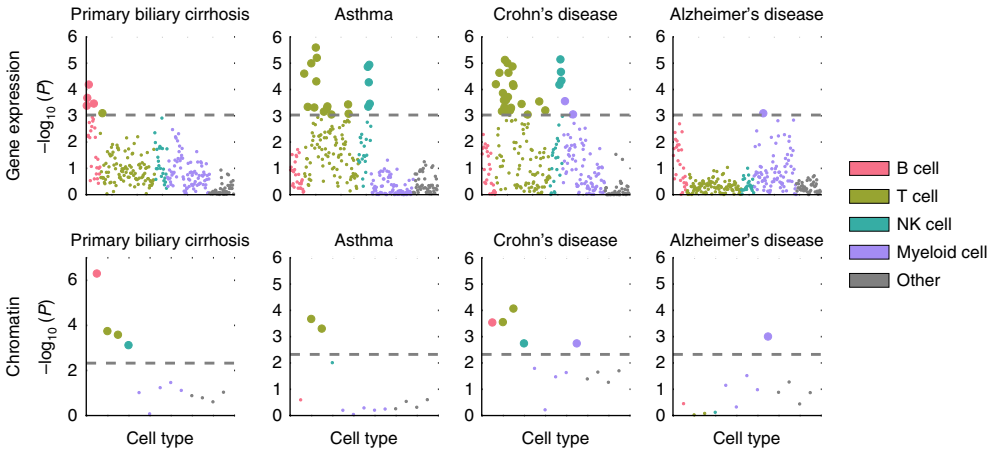

Fig. 5 | Results of the immune analysis for selected traits. Results of the analysis of ImmGen gene expression data (top) and hematopoiesis ATAC-seq data (bottom) for selected traits (results for the remaining traits are displayed in Supplementary Fig. 9). Large circles passed the cutoff of FDR < 5% at either $-\log_{10}(P) = 3.03$ (gene expression) or $-\log_{10}(P) = 2.32$ (chromatin). Numerical results are reported in Supplementary Table 10, GWAS data are described in Supplementary Table 4, gene expression and chromatin data are described in the Methods and in Supplementary Table 10, and the statistical method is described in the "Overview of methods" section above and the Methods.

exhibit significant enrichment for any of the 12 brain-related traits. The enrichment of neurons for all three 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.

To more precisely characterize the neuronal enrichments, we analyzed the five traits with neuronal enrichment at FDR < 5% using *t*-statistics computed by the PsychENCODE consortium[20] on differential expression in glutamatergic (excitatory) versus GABAergic (inhibitory) neurons (Methods). The results are displayed in Fig. 4c and Supplementary Table 8c; we used Bonferroni correction in this analysis, as we were testing only $5 \times 2 = 10$ hypotheses. For bipolar disorder, genes that were specifically expressed in GABAergic neurons exhibited heritability enrichment, whereas genes that were specific to glutamatergic neurons did not. This result supports the theory that pathology in GABAergic neurons can contribute causally to risk for bipolar disorder[51,52]. For BMI and schizophrenia, however, we found significant enrichment in glutamatergic neurons but not in GABAergic neurons.

We were unable to validate the results of these analyses using independent chromatin data. For the two analyses of brain cell types, this was because we were not aware of any available datasets with analogous chromatin data. For the analysis of brain regions, this was because the chromatin annotations that we analyzed were highly correlated across different brain regions, and thus some phenotypes showed enrichment in nearly every brain region; we did not consider these nonspecific enrichments to be a meaningful validation of our region-specific results using gene expression data.

Analysis of 25 immune-related traits using immune cell expression data. We identified 25 traits with immune enrichment at FDR < 5% in our gene expression and/or chromatin analyses (Methods). We investigated cell-type-specific enrichments for these traits using gene expression data from the Immunological Genome (ImmGen) project[21], which contains microarray data on 292 immune cell types from mice (Methods). This dataset contains data for many immune cell types that are not available in the multiple-tissue analysis, and because we compute *t*-statistics within the dataset—i.e., each immune cell versus 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 16 traits. Results show highly trait-specific patterns of enrichment (Fig. 5, Supplementary Fig. 9 and Supplementary Tables 9 and 10). For primary biliary cirrhosis, the largest and most significant enrichment was in B cells, which was consistent with literature on the importance of B cells for this trait[53,54]. Alzheimer's disease exhibits enrichment in myeloid cells, as seen previously from genetics[55,56]. Asthma and eczema both exhibited enrichment in T cells and NKT cells; several subclasses of T cells have been shown to be important in asthma[57], and a previous study using chromatin data found an enrichment in T cells for asthma but not in other immune cell types[6]. 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[58–60]. Schizophrenia and bipolar disorder both exhibited an enrichment in T cells. Patients with bipolar disorder have been shown to have a reduction in certain types of T cells, but have equal levels of B cells, NK cells and monocytes, as compared to controls[61]. T cell levels have been shown to vary between individuals with schizophrenia and controls; however, the existing literature is not consistent in its description of the direction of effect[62]. Note that our analysis excluded the HLA region; a previous analysis of the HLA region for individuals with schizophrenia implicated the complement system through its role in synaptic pruning, a signal that is distinct from the signal we observed here[63]. Finally, we identified an enrichment in stromal cells for both diastolic and systolic blood pressure. For each of these two traits, we identified enrichments in the musculoskeletal–connective category in the multiple-tissue analysis that were stronger than the immune enrichments in that analysis, and thus we hypothesize that the enrichment in stromal cells does not provide better resolution on the immune enrichment but instead reflects the more general importance of connective tissue. In enrichment correlation analyses, schizophrenia and bipolar disorder clustered with immunological diseases, whereas metabolic traits, neurological diseases and other psychiatric diseases did not (Supplementary Fig. 10).

To validate these results, we analyzed ATAC-seq (chromatin accessibility) data from 13 cell types that spanned the hematopoietic hierarchy in humans[64]. We validated 10 of the 14 top results (Supplementary Table 9 and Methods). The only immunological disease whose result was not validated was lupus; the top result for lupus in the ImmGen analysis was a myeloid cell type, whereas the largest and most significant enrichment in the hematopoiesis

dataset was a B cell enrichment, which was consistent with other genetic studies of this trait[14].

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, allowed us to take advantage of the large amount of gene expression data available—including fine-grained data for which we currently do not 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. Our results were able to improve understanding of the phenotypes studied here and to highlight the power of GWAS as a source of biological insight, and they may also be useful for choosing the relevant tissue or cell type for in vitro experiments to further elucidate the molecular mechanisms underlying significant loci across the genome that were identified in GWAS.

There are several key differences between LDSC-SEG, which relies on gene expression data without genotypes or eQTLs, and approaches that require eQTL data[3,13] (Supplementary Fig. 11, Supplementary Note and Methods). Our polygenic approach also differs from other gene expression–based approaches such as SNPsea[14,15] and DEPICT[17], which restrict their analyses to subsets of SNPs that pass a significance threshold (Supplementary Figs. 12–16, Supplementary Tables 11–15 and Supplementary Note).

We cannot conclusively say whether gene expression or chromatin data are preferable when both types of data are available for the same tissues and cell types (Supplementary Figs. 4 and 17, Supplementary Tables 10 and 16, and Methods). Instead, we conclude that the question of which type of data is preferable may depend on complex factors, such as which chromatin marks were analyzed, the sample size with which the specifically expressed genes are called and the overall quality of the dataset. When gene expression and chromatin data are available for the same set of tissues or cell types, it may be possible to combine these types of data to improve power—for example, by restricting an annotation to tissue-specific chromatin marks near specifically expressed genes or by combining the *P* values from separate analyses of the two types of data. We defer a thorough exploration of this set of possibilities to future work.

Our work is based on the assumption that a tissue or cell type is important for a particular disease if and only if SNPs near genes with high specific expression in that tissue or cell type are enriched for heritability. This assumption leads to several limitations of our approach. First, when analyzing gene expression data from different tissues, cell type composition can confound the analysis, as we demonstrated in our comparison of brain regions; this makes enrichments of organs such as the esophagus or uterus hard to interpret. Second, tissues or cell types with similar gene expression profiles to a causal tissue or cell type will be identified as being relevant to disease, just as SNPs in LD with a causal SNP will be identified as being associated with disease in a GWAS; thus, significant tissues or cell types should be cautiously interpreted as the 'best proxy' for the truly causal tissue or cell type, which may be unobserved. Third, our focus on nearby SNPs prevents us from leveraging signals from regulatory SNPs that function at longer distances. Our approach is also fundamentally limited by the availability of gene expression data and cannot rule out the importance of a given cell type; for example, if the tissue or cell type that is most relevant for a disease occurs in a stage of development or under a stimulus that has not been assayed, then we may not identify enrichments in that tissue or cell type. We would also like to highlight that for most of these phenotypes there is likely not just one causal tissue or cell type, but many.

Our use of a heritability-based approach has advantages but also leads to some limitations. First, our approach will not detect strong but highly localized signals. Second, power increases only modestly with sample size at very large sample sizes (Supplementary Note). Also, because our approach uses stratified LD score regression, it cannot be applied to custom array data; it requires a sequenced reference panel that matches the population studied in the GWAS and can be affected by model misspecification[7]. Recent augmentations to the baseline model[65] have been shown to help ameliorate model misspecification, but we leave further investigation of this in the context of cell-type-specific analyses to future work.

Another limitation of our method is that its results may be difficult to validate. We undertook a type of validation using independent chromatin data, when there were comparable chromatin data available. However, this type of validation involves a number of challenges. First, we often do not have chromatin data for the same tissues and cell types as the gene expression data. Second, it is not clear that we should always expect results to replicate; for example, it is biologically plausible that SNPs near specifically expressed genes in the relevant tissue are enriched, whereas SNPs in the trimethylated Lys36 of histone H3 (H3K36me3) peaks called in the tissue are not. Third, our gene expression annotations represent relative activity—we select genes that have higher expression in the focal tissue than in other tissues—whereas the chromatin annotations that we use here represent absolute activity (although relative chromatin annotations are also possible[6,66]). Despite these limitations, replicating an enrichment for a particular system, tissue or cell type using independent chromatin data can provide strong validation for gene expression results.

Our power to identify disease-relevant tissues and cell types will improve as large GWAS sample sizes become available for more phenotypes, and as gene expression data are generated in new tissues and cell types. This will help advance understanding of disease biology and lay the groundwork for future experiments exploring specific variants and mechanisms.

URLs. LDSC software on GitHub, including LDSC-SEG, https://github.com/bulik/ldsc; gene sets and LD scores from this paper, https://data.broadinstitute.org/alkesgroup/LDSCORE/; GTEx, http://www.gtexportal.org/; Franke lab dataset, https://data.broadinstitute.org/mpg/depict/depict_download/tissue_expression; Cahoy et al. dataset (see Supplementary Tables 4–6), http://jneurosci.org/content/suppl/2008/01/03/28.1.264.DC1; PsychENCODE, https://www.synapse.org//#!Synapse:syn4921369/wiki/235539; ImmGen, https://www.immgen.org/; Roadmap Epigenomics, http://www.roadmapepigenomics.org/; GERA dataset (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.

Methods

Methods, including statements of data availability and any associated accession codes and references, are available at https://doi.org/10.1038/s41588-018-0081-4.

Received: 19 September 2016; Accepted: 29 January 2018;
Published online: 09 April 2018

References

1. 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. GTEx Consortium. The Genotype–Tissue Expression (GTEx) pilot analysis: multi-tissue gene regulation in humans. *Science* **348**, 648–660 (2015).
4. Ernst, J. et al. Mapping and analysis of chromatin-state dynamics in nine human cell types. *Nature* **473**, 43–49 (2011).
5. Trynka, G. et al. Chromatin marks identify critical cell types for fine mapping complex trait variants. *Nat. Genet.* **45**, 124–130 (2013).

6. Farh, K. K.-H. et al. Genetic and epigenetic fine-mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
7. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
8. Li, Y. & Kellis, M. Joint Bayesian inference of risk variants and tissue-specific epigenomic enrichments across multiple complex human diseases. *Nucleic Acids Res.* **44**, e144 (2016).
9. Maurano, M. T. et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
10. 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).
11. Kichaev, G. et al. Integrating functional data to prioritize causal variants in statistical fine-mapping studies. *PLoS Genet.* **10**, e1004722 (2014).
12. 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).
13. Ongen, H. et al. Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* **49**, 1676-1683 (2016).
14. 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).
15. 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).
16. Gormley, P. et al. Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nat. Genet.* **48**, 856–866 (2016).
17. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nat. Commun.* **6**, 5890 (2015).
18. Fehrmann, R. S. N. et al. Gene expression analysis identifies global gene-dosage sensitivity in cancer. *Nat. Genet.* **47**, 115–125 (2015).
19. 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).
20. Akbarian, S. et al. The PsychENCODE project. *Nat. Neurosci.* **18**, 1707–1712 (2015).
21. Heng, T. S. P. et al. The Immunological Genome Project: networks of gene expression in immune cells. *Nat. Immunol.* **9**, 1091–1094 (2008).
22. Auton, A. et al. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
23. 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).
24. Anttila, V. et al. Analysis of shared heritability in common disorders of the brain. Preprint at *bioRxiv* https://doi.org/10.1101/048991 (2016).
25. 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).
26. 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).
27. 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).
28. 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).
29. Traylor, M. et al. Genetic risk factors for ischemic stroke and its subtypes (the METASTROKE collaboration): a meta-analysis of genome-wide association studies. *Lancet Neurol.* **11**, 951–962 (2012).
30. Patsopoulos, N. A. et al. Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. *Ann. Neurol.* **70**, 897–912 (2011).
31. 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).
32. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
33. Okbay, A. et al. Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* **533**, 539–542 (2016).
34. 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).
35. Teslovich, T. M. et al. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* **466**, 707–713 (2010).
36. Schunkert, H. et al. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat. Genet.* **43**, 333–338 (2011).
37. 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).
38. Okada, Y. et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature* **506**, 376–381 (2014).
39. Jostins, L. et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
40. 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).
41. Dubois, P. C. A. et al. Multiple common variants for celiac disease influencing immune gene expression. *Nat. Genet.* **42**, 295–302 (2010).
42. 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).
43. 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).
44. 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).
45. 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).
46. 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).
47. Callicott, J. H. et al. Physiological dysfunction of the dorsolateral prefrontal cortex in schizophrenia revisited. *Cereb. Cortex* **10**, 1078–1092 (2000).
48. Medic, N. et al. Increased body mass index is associated with specific regional alterations in brain structure. *Int. J. Obes.* **40**, 1177–1182 (2016).
49. Maleki, N. et al. Migraine attacks the basal ganglia. *Mol. Pain* **7**, 71 (2011).
50. 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).
51. 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).
52. Benes, F. M. & Berretta, S. GABAergic interneurons: implications for understanding schizophrenia and bipolar disorder. *Neuropsychopharmacology* **25**, 1–27 (2001).
53. Dhirapong, A. et al. B cell depletion therapy exacerbates murine primary biliary cirrhosis. *Hepatology* **53**, 527–535 (2011).
54. Zhang, J. et al. Ongoing activation of autoantigen-specific B cells in primary biliary cirrhosis. *Hepatology* **60**, 1708–1716 (2014).
55. Raj, T. et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
56. Huang, K. L. et al. A common haplotype lowers PU.1 expression in myeloid cells and delays onset of Alzheimer's disease. *Nat. Neurosci.* **20**, 1052–1061 (2017).
57. Lloyd, C. M. & Hessel, E. M. Functions of T cells in asthma: more than just T_H2 cells. *Nat. Rev. Immunol.* **10**, 838–848 (2010).
58. 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).
59. Xavier, R. J. & Podolsky, D. K. Unravelling the pathogenesis of inflammatory bowel disease. *Nature* **448**, 427–434 (2007).
60. Sospedra, M. & Martin, R. Immunology of multiple sclerosis. *Annu. Rev. Immunol.* **23**, 683–747 (2005).
61. Barbosa, I. G., Machado-Vieira, R., Soares, J. C. & Teixeira, A. L. The immunology of bipolar disorder. *Neuroimmunomodulation* **21**, 117–122 (2014).
62. 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).
63. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
64. Corces, M. R. et al. Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nat. Genet.* **48**, 1193–1203 (2016).
65. Gazal, S. et al. Linkage-disequilibrium-dependent architecture of human complex traits reveals action of negative selection. *Nat. Genet.* **49**, 1421–1427 (2017).
66. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).

Acknowledgements

We are thankful to R. Herbst, E. Hodis, F. Hormozdiari, M. Kanai, T. Pers, S. Riesenfeld, J. Ulirsch and A. Veres for helpful comments. This research was conducted using the UK Biobank Resource (application number: 16549). This research was funded by NIH grants R01 MH107649 (H.K.F., S.G., B.M.N., A.L.P.), R01 MH109978 (A.G., A.L.P.), U01 CA194393 (H.K.F., A.L.P.) and U01 HG009379 (S.R., A.L.P.). H.K.F. was also supported by the Fannie and John Hertz Foundation and by Eric and Wendy Schmidt. Data on neuron types were generated as part of the PsychENCODE Consortium,

supported by: U01MH103392 (S. Akbarian, Icahn School of Medicine at Mount Sinai; P. Sklar, Icahn School of Medicine at Mount Sinai), U01MH103365 (F. Vaccarino, Yale University; M. Gerstein, Yale University; S. Weissman, Yale University), U01MH103346 (P. Farnham, University of Southern California; J. A. Knowles, University of Southern California), U01MH103340 (C. Liu, SUNY Upstate Medical University; K. White, University of Chicago), U01MH103339 (N. Sestan, Yale University; M. State, University of California, San Francisco), R21MH109956 (A. Jaffe, Lieber Institute for Brain Development), R21MH105881 (D. Pinto, Icahn School of Medicine at Mount Sinai), R21MH105853 (A. Jaffe, Lieber Institute for Brain Development; D. Weinberger, Lieber Institute for Brain Development), R21MH103877 (S. Dracheva, Icahn School of Medicine at Mount Sinai; S. Akbarian, Icahn School of Medicine at Mount Sinai), R21MH102791 (A. Jaffe, Lieber Institute for Brain Development), R01MH111721 (F. Goes, Johns Hopkins University; T. Hyde, Lieber Institute for Brain Development), R01MH110928 (M. State, University of California, San Francisco; S. Sanders, University of California, San Francisco; J. Willsey, University of California, San Francisco), R01MH110927 (D. Geschwind, University of California, Los Angeles), R01MH110926 (N. Sestan, Yale University), R01MH110921 (P. Sklar, Icahn School of Medicine at Mount Sinai), R01MH110920 (C. Liu, SUNY Upstate Medical University), R01MH110905 (K. White, University of Chicago), R01MH109715 (D. Pinto, Icahn School of Medicine at Mount Sinai), R01MH109677 (P. Roussos, Icahn School of Medicine at Mount Sinai), R01MH105898, (P. Zandi, Johns Hopkins University; T. M. Hyde, Lieber Institute for Brain Development), R01MH094714, (D. Geschwind, University of California, Los Angeles), P50MH106934, (N. Sestan, Yale University), R01MH105472 (G. Crawford, Duke University; P. Sullivan, University of North Carolina).

Author contributions

H.K.F. and A.L.P. designed the study; H.K.F., Y.A.R., K.S. and S.P. analyzed data; H.K.F. and A.L.P. wrote the manuscript with assistance from Y.A.R., V.A., K.S., A.G., A.B., S.G., P.-R.L., C.L., N.S., G.G., A.S., E.M., S.P., J.R.B.P., J.D.B., B.E.B., S.R., S.M. and B.M.N.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41588-018-0081-4.

Correspondence and requests for materials should be addressed to H.K.F. or A.L.P.

Methods

Computing *t*-statistics. When computing the *t*-statistic of each gene for a focal tissue, we excluded all samples from a similar tissue category (described for each dataset below). For example, when computing the *t*-statistic of specific expression for each gene in the 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 were more highly expressed in brain tissues than in non-brain tissues, even if they were not specific to the cortex region within the brain. This procedure resulted in a higher correlation among the *t*-statistics for the different brain regions; in a separate analysis, we computed within-brain *t*-statistics to disentangle this signal.

Thus, for a focal tissue (for example, cortex) in a larger tissue category (for example, brain), we computed the *t*-statistic for gene *g* as follows. We first constructed a design matrix *X*, where each row corresponded to a sample that was either in the cortex or outside of the brain. The first column of *X* had a '1' for every cortex sample and a '−1' for every non-brain sample. The remaining columns were an intercept and covariates (see below). The outcome *Y* in our model was expression. We fit this model via ordinary least-squares and computed a *t*-statistic for the first explanatory variable in the standard way

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

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

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

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

For visualization purposes and discussion of results, it is often useful to color tissues or cell types according to categories (categorization); the categorization for visualization is not always the same as the categorization for computing *t*-statistics. We gave the categorization for visualization in the Supplementary Tables listed in the respective figure captions.

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 dataset. 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 dataset[17,18]. Thus, it was not appropriate to compute *t*-statistics in this dataset. We used the original values in place of our *t*-statistics, then proceeded as described in Fig. 1.

Cahoy dataset. The dataset of Cahoy et al.[19] 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, by adding a 100-kb window and applying stratified LD score regression.

PsychENCODE dataset. The PsychENCODE dataset had available *t*-statistics for GABAergic neurons versus glutamatergic neurons. We used these *t*-statistics, rather than computing our own.

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

Application of stratified LD score regression. Stratified LD score regression[7] 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 modeled the causal effect of SNP *j* on phenotype *Y* as drawn from a distribution with mean 0 and variance

$$Var(\beta_i)=\sum_k \tau_k 1\{i\in C_k\} \quad (1)$$

(If the genomic annotations are real-valued rather than subsets of SNPs, then we can replace $1\{i\in C_k\}$ with any other function of the SNP indices[65].) We then modeled the phenotype *Y* as depending linearly on genotype: $Y=X\cdot\beta+\varepsilon$, where *X* is a vector of SNP values for an individual, each SNP has been standardized to mean 0 and variance 1 in the population, and ε represents environmental effects and noise. Because each SNP is standardized, and because β_i has a mean of 0, 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; so, we are estimating a fundamentally different quantity than that in some other methods[67].)

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[68], and $\ell(i,k)$ is the LD score of SNP *i* to category C_k, defined as $\ell(i,k)=\sum_j r^2(i,j)1\{j\in C_k\}$, 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 0 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 analysis[7], we computed *P* values that tested whether τ_k was positive. When reporting quantitative results, we normalized 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[7].

We analyzed autosomes only and excluded genes in the HLA region from all analyses. In each analysis, we jointly fit the following annotations: (i) the annotation created for our focal tissue by adding 100-kb windows around the top 10% of genes ranked by *t*-statistic; (ii) an identical annotation created for all genes included in the gene expression dataset being analyzed; (iii) the baseline model with 52 functional categories, described previously[7] and listed in Supplementary Table 1.

GTEx dataset. We downloaded the RNA sequencing (RNA-seq) read counts from GTEx v6p (see URLs), removed genes for which fewer than four 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 transcripts per million (TPM) normalization[69]. We analyzed 53 tissues with an average of 161 samples per tissue. We used the 'SMTSD' variable ('Tissue Type, more specific detail of tissue type') to define our tissues and the 'SMTS' variable ('Tissue Type, area from which the tissue sample was taken') to define the tissue categories for *t*-statistic computation (Supplementary Table 2). We used age and sex as covariates for our *t*-statistics.

Franke lab dataset. The Franke lab dataset is an aggregation of publicly available microarray gene expression datasets comprising 37,427 human samples[17,18]. We downloaded the publicly available gene expression data from the DEPICT website (see URLs). The available gene expression values already quantify relative expression for a tissue or cell type rather than absolute expression for a single sample[17,18], and so we used these values in place of our *t*-statistics. 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 (for example, 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 (for example, nose versus nasal mucosa; quadriceps muscle versus 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 dataset contained 152 tissues, which are listed in Supplementary Table 3.

UK Biobank data. We analyzed summary statistics from the full $N=500{,}000$ UK Biobank release[23] for 13 traits generated using BOLT-LMM v2.3[70].

Enrichment correlation. For a pair of phenotypes and a set of tissue or cell types, we defined the enrichment correlation to be the correlation between the regression coefficients that corresponded to each tissue or cell type. We estimated the enrichment correlation by correlating the estimates of the regression coefficients, and we quantified uncertainty via block jackknife over 200 sets of consecutive SNPs. We note that when the number of tissues or cell types included is small,

the true underlying enrichment correlation may be large even though there is no relationship between the two phenotypes; so, we only estimate enrichment correlations when there are at least ten tissues or cell types.

Distribution of *P* values. The correlation structure among annotations can lead to a distribution of *P* values that is highly non-uniform with many *P* values close to 0 or 1 (Fig. 2). This is caused by our one-sided test for enrichment, testing whether the regression coefficient—which represents the change in per-SNP heritability due to a given annotation, beyond what is explained by the set of all genes as well as the baseline model—is positive. The *P* values near 0 occur due to correlated annotations with true signal, and the *P* values near 1 occur due to annotations without true signal that, conditional on the baseline model, are negatively correlated to annotations with true signal as a consequence of our construction of sets of specifically expressed genes; these annotations thus have negative regression coefficients.

Chromatin-based annotations. We downloaded narrow peaks from the Roadmap Epigenomics consortium for DNase I hypersensitivity (DHS) 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 or tissues, for a total of 397 cell-type- or tissue-specific annotations. We also analyzed peaks called using Homer from EN-TEx, a subgroup of the ENCODE project, for four activating histone marks (H3K27ac, H3K4m3, H3K4me1 and H3K36me3). Each of these four features was present in a subset of 27 tissues that were also included in the GTEx dataset, for a total of 93 cell-type- or tissue-specific annotations. For each of these two datasets, for each of the annotations, we tested for enrichment by adding the annotation to the baseline model (Supplementary Table 1), 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- or 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 differed from our previous analysis of chromatin data[7] in three ways. First, we used a larger range of marks and of tissues or 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, in addition to recent EN-TEx data. Second, for our analysis of Roadmap data, we used narrow peaks from Roadmap for all of the marks. Previously, we analyzed H3K27ac data from one source[6] and H3K4me1, H3K4me3 and H3K9ac data from another source[5,12]; now that there was a single standard source with uniformly processed data for all of the Roadmap data, we switched to using these 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 tended to fall in areas more active overall were not falsely interpreted as being a cell-type-specific signal.

Classification of tissues or cell types for system-level validation of the results of the multiple-tissue analysis of gene expression. We used the classification for visualization used in Fig. 2, classifying the top tissue or cell type for each trait with a significant enrichment into one of the eight systems (excluding "Other") in the Fig. 2 legend. There were three phenotypes whose top tissue fell in the "Other" category; two of these we classified into a new "Reproductive" category. The last one, serous membrane, did not have any comparable tissues in our chromatin data, and we instead attempted to replicate the second most significant result for that phenotype.

Multiple-tissue validation results. The top enrichment from our multi-tissue analysis of gene expression was validated at the system level for 33 of 34 phenotypes, and at the tissue level for 13 of 20 (Results). If we allowed an enrichment of any artery sample in GTEx to be validated by an enrichment of any artery sample in EN-TEx (instead of requiring strict matching of aorta, tibial artery and coronary artery), then the number of validations increased from 13 to 16. Of the four remaining results that were not validated, three were an enrichment in lung for an immunological disease; for all three diseases, the top enrichment in the analysis of gene expression (not restricting to tissues shared between GTEx and EN-TEx) was an immune category from the Franke lab dataset, and the top enrichment in the analysis of chromatin data was an immune category in the Roadmap dataset. We hypothesize that the lung samples analyzed in GTEx contained substantial amounts of blood and thus exhibited a gene expression signature reflecting immune activity; this idea is supported by a Gene Ontology (GO) enrichment analysis of the lung gene set, in which the top three results were related to antigen presentation, immune response and cytokine-mediated signaling, respectively.

Heritability enrichments of chromatin-based annotations. After aggregating all of the results of the Roadmap and EN-TEx chromatin analyses, we found at least one tissue that was significant at FDR < 5% for 44 of the 48 traits (Supplementary Fig. 5 and Supplementary Tables 5 and 7). Averaging across the most significant annotation for each of these 44 traits, we found that the tissue-specific chromatin annotation spanned 3.3% of the genome and explained 43% of the SNP heritability (Supplementary Table 5). The sizes of the annotation ranged from 0.8% to 7.8%, and the estimates of enrichment varied from 3.5× to 33×, which represented much more variability than for the top annotations in the multiple-tissue gene expression analysis. Because the annotations were much smaller, the estimates of proportion of heritability tended to be much noisier.

Phenotypes with CNS enrichment. The following 12 traits had CNS enrichment at FDR < 5% in either the multiple-tissue analysis of gene expression or in the analysis of chromatin data described above: schizophrenia, bipolar disorder, Tourette syndrome, epilepsy, generalized epilepsy, attention-deficit hyperactivity disorder (ADHD), migraine, depressive symptoms, BMI, smoking status, years of education and neuroticism. The nervous system has been implicated, either with genetic evidence or with non-genetic evidence, for each of these traits[7,24,32,34,45,71–73].

Analysis of 13 brain regions using data from GTEx. Although 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 being enriched. For example, nearly every brain region in either the GTEx or Franke lab data was found to be enriched at FDR < 5% in individuals with schizophrenia (Fig. 2). To differentiate among brain regions, we restricted ourselves to gene expression data from only samples from the brain in the GTEx data. We computed *t*-statistics within the brain-only dataset; for example, we computed *t*-statistics for cortex versus other brain regions instead of cortex versus other tissues in GTEx, and we used these new *t*-statistics to construct and test gene sets as in the multiple-tissue analysis. In this analysis, we set each tissue to be its own category for the computation of *t*-statistics, and we used age and sex as covariates. Individual-level data were not available for the Franke lab dataset, and thus we could not compute within-brain *t*-statistics for this dataset.

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, whereas recomputing *t*-statistics within the brain allows us to construct new annotations with lower correlations (Supplementary Fig. 7), 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 used differential expression within the brain, rather than joint analysis of the original annotations, to differentiate among brain regions.

Data on three brain cell types from Cahoy et al. The authors of Cahoy et al.[19] purified neurons, astrocytes and oligodendrocytes from mouse forebrain and made lists of specifically expressed genes available for each of these three cell types, which we downloaded (see URLs). To obtain a list of all genes, we also downloaded a list of all of the genes that passed quality control in their analysis (Supplementary Table 3b in Cahoy et al.[19]). We mapped the mouse genes to human orthologs using Ensembl (see URLs).

Data on two neuron types from PsychENCODE. PsychENCODE[20] generated RNA-seq data from the nuclei of GABAergic and glutamatergic neurons from the dorsolateral prefrontal cortex of four neurotypical human donors and computed *t*-statistics using limma[74]. We used these *t*-statistics.

Phenotypes with immune enrichment. Twenty-five traits had immune enrichment at FDR < 5% in either the multiple-tissue analysis of gene expression or in the analysis of chromatin data. This included 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 included Alzheimer's and Parkinson's diseases, which are neurodegenerative diseases with an immune component that was previously identified from genetics[75,76], as well as several brain-related traits—ADHD, anorexia nervosa, bipolar disorder, schizophrenia, Tourette syndrome and neuroticism—and high-density lipoprotein (HDL), LDL, triglycerides, diastolic and systolic blood pressure, hypertension and BMI. Several of the brain-related traits have previously been suggested to have an immune component[32,77,78]; HDL, LDL and triglycerides have been linked to immune activation[79–82]; immune cells are causally involved in blood pressure and hypertension[83]; and obesity, in addition to contributing to inflammation[84], can also be induced in mice through alterations of the immune system[85].

Data on 292 immune cell types from ImmGen. We downloaded publicly available microarray gene expression data on 292 immune cell types from the ImmGen Consortium (see URLs). We used both phase 1 (GSE15907) and phase 2 (GSE37448) data. The data on Gene Expression Omnibus (GEO) were on an exponential scale, so we log-transformed the data and mapped it to human genes using ENSEMBL orthologs. We defined tissue categories for *t*-statistic computation using the classification on the main page of http://www.immgen.org of cell types

into categories: B cells, γδ T cells, αβ T cells, innate lymphocytes, myeloid cells, stromal cells and stem cells (Supplementary Table 10). The classification at http://www.immgen.org also has a 'T cell activation' category that we collapsed into the αβ T cell category because it had data on αβ T cells at different stages of activation. We did not include any covariates.

Validation of immune cell results. To validate the results of the ImmGen analysis, we analyzed ATAC-seq peaks from 13 cell types that spanned the hematopoietic hierarchy in humans[64]. The 13 cell types did not allow us to validate at a very high resolution; instead, we classified all of the cell types from ImmGen and from the hematopoiesis dataset using the classification for visualization of Fig. 5 into five categories: B cells, T cells, NK cells, myeloid cells and other cells. There were no stromal cells in the hematopoiesis dataset, and it was not possible to validate the enrichments for diastolic and systolic blood pressure; this left us with 14 phenotypes with an enrichment at FDR < 5% in the ImmGen analysis for which the top result fell into one of the first four categories (excluding 'Other'). We considered one of these 14 results to be validated if any cell type in the same category from the hematopoiesis dataset passed FDR < 5%. The four phenotypes whose top results did not replicate were lupus, schizophrenia, bipolar disorder and neuroticism.

Differences between LDSC-SEG and eQTL-based approaches. Our approach differs in several key ways from approaches that require eQTL data[3,13]. First, our approach can be applied to expression datasets such as the Franke lab dataset, the Cahoy dataset, the PsychENCODE dataset and the ImmGen dataset, which do not have genotypes or eQTLs available (Table 1). Second, 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, although subtler enrichments had decreasing levels of significance as the gene expression data were down-sampled (Supplementary Fig. 11 and Supplementary Note). 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 have a small number of samples per tissue or cell type, in contrast to that with eQTL-based methods. Finally, we note that a recent study[86] tested 30 phenotypes for tissue-specific enrichment in 44 tissues from GTEx using the TWAS approach[87] but concluded that their results "did not suggest tissue-specific enrichment at the current sample sizes". We share their hypothesis that this is because eQTLs are often shared across tissues even when overall expression levels are very different.

Comparison of gene expression and chromatin for cell-type-specific analysis. Our estimated enrichments were higher for the chromatin-based annotations than for the gene-expression-based annotations, but the gene-expression-based annotations were larger and had less LD to the rest of the genome. Some chromatin marks tend to be more cell type specific than overall gene expression, but our specifically expressed gene sets had low correlation across tissues (Supplementary Fig. 17). There were two instances in which we had gene expression and chromatin data on the same set of tissues or cell types, and we compared the *P* values in our analyses of these datasets. First, we compared our results from GTEx (gene expression) and EN-TEx (chromatin) for the tissues shared between these two datasets in the multiple-tissue analysis, and we found that the two datasets had comparable distributions of *P* values (Supplementary Fig. 4). In the second instance, the hematopoietic dataset that we analyzed[64] had matched ATAC-seq and RNA-seq data, and although our analysis of the ATAC-seq peaks led to significant enrichments for many traits (Fig. 5 and Supplementary Table 10), the RNA-seq dataset yielded only a single enrichment for a single trait (Supplementary Table 16).

Life Sciences Reporting Summary. Further information on experimental design is available in the Life Sciences Reporting Summary.

Code availability. Open source software implementing our approach is available on Github (see URLs).

Data availability. We have released all genome annotations derived from the publicly available gene expression data that we analyzed (see URLs). This includes all annotations used in Figs. 2–5 with the exception of the annotations derived from the PsychENCODE data in Fig. 4c, for which we did not have permission to release annotations.

References

67. 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).
68. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
69. Wagner, G. P., Kin, K. & Lynch, V. J. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.* **131**, 281–285 (2012).
70. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A. P. & Price, A. L. Mixed model association for biobank-scale data sets. Preprint at *bioRxiv* https://doi.org/10.1101/194944 (2017).
71. Backenroth, D. et al. Tissue-specific functional effect prediction of genetic variation and applications to complex trait genetics. Preprint at *bioRxiv* https://doi.org/10.1101/069229 (2016).
72. Wilens, T. E., Biederman, J. & Spencer, T. J. Attention deficit or hyperactivity disorder across the lifespan. *Annu. Rev. Med.* **53**, 113–131 (2002).
73. 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).
74. 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).
75. Gjoneska, E. et al. Conserved epigenomic signals in mice and humans reveal immune basis of Alzheimer's disease. *Nature* **518**, 365–369 (2015).
76. Gagliano, S. A. et al. Genomics implicates adaptive and innate immunity in Alzheimer's and Parkinson's diseases. *Ann. Clin. Transl. Neurol* **3**, 924–933 (2016).
77. 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).
78. Elamin, I., Edwards, M. J. & Martino, D. Immune dysfunction in Tourette syndrome. *Behav. Neurol.* **27**, 23–32 (2013).
79. 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).
80. Broedl, U. C. et al. Endothelial lipase promotes the catabolism of ApoB-containing lipoproteins. *Circ. Res.* **94**, 1554–1561 (2004).
81. Feingold, K. R. & Grunfeld, C. The role of HDL in innate immunity. *J. Lipid Res.* **52**, 1–3 (2011).
82. Lo, J. C. et al. Lymphotoxin-β-receptor-dependent control of lipid homeostasis. *Science* **316**, 285–288 (2007).
83. Harrison, D. G. The immune system in hypertension. *Trans. Am. Clin. Climatol. Assoc.* **125**, 130–138 (2014).
84. Hotamisligil, G. S. Inflammation and metabolic disorders. *Nature* **444**, 860–867 (2006).
85. Zlotnikov-Klionsky, Y. et al. Perforin-positive dendritic cells exhibit an immunoregulatory role in metabolic syndrome and autoimmunity. *Immunity* **43**, 776–787 (2015).
86. Mancuso, N. et al. Integrating gene expression with summary-association statistics to identify genes associated with 30 complex traits. *Am. J. Hum. Genet.* **100**, 473–487 (2017).
87. Gusev, A. et al. Integrative approaches for large-scale transcriptome-wide association studies. *Nat. Genet.* **48**, 245–252 (2016).

Corresponding Author: Hilary Finucane

Date: November 21, 2017

Life Sciences Reporting Summary

Nature Research wishes to improve the reproducibility of the work we publish. This form is published with all life science papers and is intended to promote consistency and transparency in reporting. All life sciences submissions use this form; while some list items might not apply to an individual manuscript, all fields must be completed for clarity.

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

Experimental design

1. Sample size

   Describe how sample size was determined.

   We analyzed available GWAS data, and did not do a new experiment in which we determined the sample size.

2. Data exclusions

   Describe any data exclusions.

   When computing t-statistics, we excluded samples from the same category as the focal tissue. We excluded the HLA from all analyses and analyzed only autosomes.

3. Replication

   Describe whether the experimental findings were reliably reproduced.

   There were no experimental findings. Where possible, we validated our computational results using gene expression data with chromatin data.

4. Randomization

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

   We did not allocate samples into experimental groups.

5. Blinding

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

   There was no group allocation.

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

6. Statistical parameters

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

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

See the web collection on statistics for biologists for further resources and guidance.

Software

Policy information about availability of computer code

7. Software

Describe the software used to analyze the data in this study.	We used the LDSC package, available on github. We also ran the SNPsea, MAGMA, and DEPICT software for comparison, using the 2016 versions of each tool.

For all studies, we encourage code deposition in a community repository (e.g. GitHub). Authors must make computer code available to editors and reviewers upon request. The *Nature Methods* guidance for providing algorithms and software for publication may be useful for any submission.

Materials and reagents

Policy information about availability of materials

8. Materials availability

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

9. Antibodies

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

10. Eukaryotic cell lines

a. State the source of each eukaryotic cell line used.	N/A
b. Describe the method of cell line authentication used.	N/A
c. Report whether the cell lines were tested for mycoplasma contamination.	N/A
d. If any of the cell lines used in the paper are listed in the database of commonly misidentified cell lines maintained by ICLAC, provide a scientific rationale for their use.	N/A

Animals and human research participants

Policy information about studies involving animals; when reporting animal research, follow the ARRIVE guidelines

11. Description of research animals

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

Policy information about studies involving human research participants

12. Description of human research participants

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