nature genetics

ARTICLES

https://doi.org/10.1038/s41588-021-00987-9

Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits

Biao Zeng[1,2,3,4,5], Jaroslav Bendl[1,2,3,4,5], Roman Kosoy[1,2,3,4,5], John F. Fullard[1,2,3,4,5], Gabriel E. Hoffman[1,2,3,4,5] ✉ and Panos Roussos[1,2,3,4,5,6,7] ✉

While large-scale, genome-wide association studies (GWAS) have identified hundreds of loci associated with brain-related traits, identification of the variants, genes and molecular mechanisms underlying these traits remains challenging. Integration of GWAS with expression quantitative trait loci (eQTLs) and identification of shared genetic architecture have been widely adopted to nominate genes and candidate causal variants. However, this approach is limited by sample size, statistical power and linkage disequilibrium. We developed the multivariate multiple QTL approach and performed a large-scale, multi-ancestry eQTL meta-analysis to increase power and fine-mapping resolution. Analysis of 3,983 RNA-sequenced samples from 2,119 donors, including 474 non-European individuals, yielded an effective sample size of 3,154. Joint statistical fine-mapping of eQTL and GWAS identified 329 variant–trait pairs for 24 brain-related traits driven by 204 unique candidate causal variants for 189 unique genes. This integrative analysis identifies candidate causal variants and elucidates potential regulatory mechanisms for genes underlying schizophrenia, bipolar disorder and Alzheimer's disease.

Genome-wide association studies have associated hundreds of loci with neuropsychiatric and neurodegenerative traits[1–5]. Nevertheless, elucidation of the molecular mechanisms underlying these traits remains challenging since most risk variants are noncoding and highly correlated due to linkage disequilibrium (LD)[3,6]. Integration of risk loci with expression quantitative trait loci (eQTL) has been widely adopted to identify genes and candidate causal variants[7–9]. Recent work by the Genotype-Tissue Expression (GTEx) consortium across 838 individuals and 49 tissues detected eQTLs for 95% of protein-coding and >60% of long noncoding RNA genes[8]. While the power to detect primary (that is, the most significant association) eQTLs is very high, advances in identification of tissue- and cell-type-specific effects, conditionally independent effects and candidate causal variants in trait-relevant tissues and cell types promise to further inform the molecular etiology of disease[8–12].

Large-scale efforts have been undertaken to catalog human brain eQTLs[8,13–16]. All these efforts focus on homogenate brain tissue, which is composed of multiple cell types[17–20], and therefore cell-type-specific eQTLs are not fully captured[21–23]. This is an important limitation given that disease variants act through cell-type-specific biological effects[21,24,25]. Initial efforts have included cell-type-specific eQTL analysis in the human brain by experimental purification of specific cell types[26–28], but the sample size of such studies is necessarily limited by the increased experimental costs and data quality can be affected by additional experimental steps. An alternative strategy to capture cell-type-specific effects is to statistically define conditional- or context-dependent eQTLs[10,11]. While existing studies have sufficient power to detect primary eQTLs, identification of conditionally independent eQTLs that capture more subtle cell-type-specific effects requires large sample sizes[29,30].

Following eQTL detection, statistical fine-mapping can identify candidate causal variants likely to drive variation in expression[6,9,31,32]. Going one step further, joint statistical fine-mapping by integration of GWAS and gene expression traits can define those candidate causal variants that increase disease risk through alterations of gene expression[9]. Interpretion and validation of such variants can pinpoint genes such as *FURIN*[33], *BIN1* (ref. [34]) and *C4* (ref. [35]), along with molecular mechanisms that can be further studied in experimental systems. Nevertheless, the resolution of statistical fine-mapping for eQTL and GWAS is incomplete due to limited sample sizes and lack of transancestry analysis[6]. A sample size of >2,000 donors is needed to detect eQTLs and perform GWAS colocalization for identification of causal variants explaining 1% of heritability[9]. The large human brain eQTL mega-analysis by PsychENCODE included 1,387 unique donors from multiple cohorts[15]. Moreover, most eQTL analyses have been limited to European populations despite the fact that much shorter LD in individuals of African or African-American ancestry can substantially increase the resolution of statistical fine-mapping[6,36–38].

Given the limited availability of human brain samples, it is critical to maximize power and fine-mapping resolution by combining existing datasets. However, differences in study designs have thus far hindered such efforts. Multi-ancestry studies have long been challenging in genetics, but linear mixed models can control the

[1]Center for Disease Neurogenomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Pamela Sklar Division of Psychiatric Genomics, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]Icahn Institute for Data Science and Genomic Technology, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]Mental Illness Research, Education and Clinical Centers, James J. Peters VA Medical Center, Bronx, NY, USA. [7]Center for Dementia Research, Nathan Kline Institute for Psychiatric Research, Orangeburg, NY, USA.
✉e-mail: gabriel.hoffman@mssm.edu; panagiotis.roussos@mssm.edu

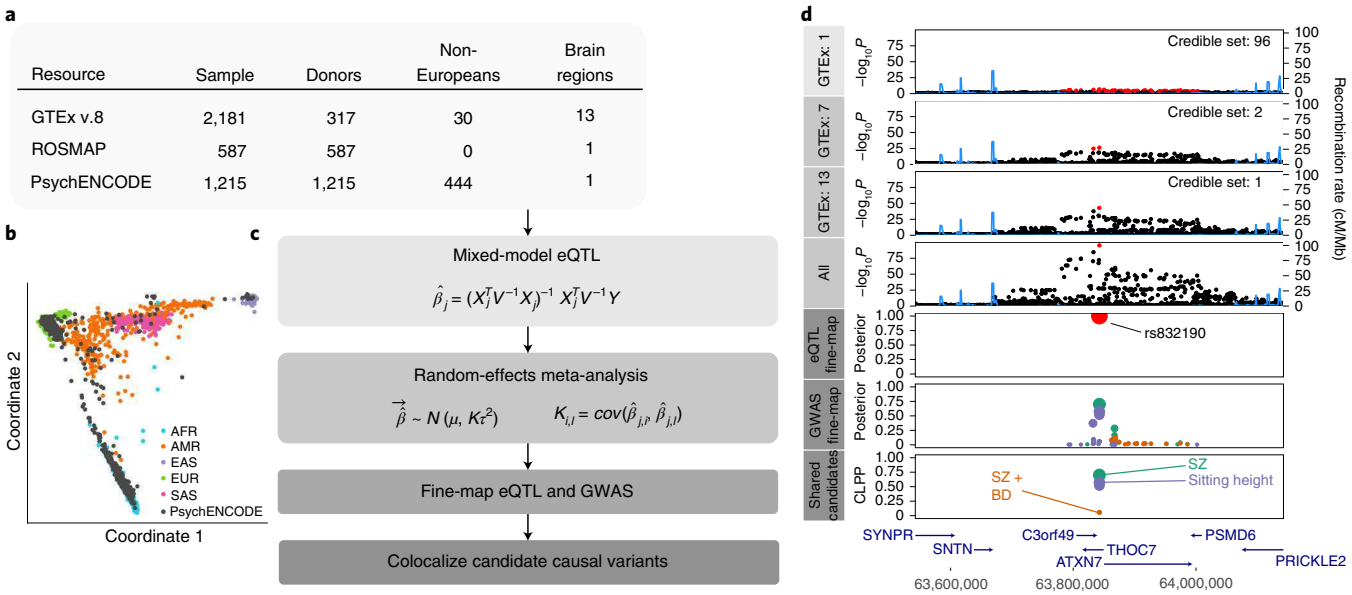


Fig. 1 | Workflow for multi-ancestry eQTL meta-analysis. a, RNA-seq datasets with details on ancestry and repeated measures. **b**, Multidimensional scaling illustrating diverse ancestry of donors from PsychENCODE resource in addition to donors of African (AFR), American (AMR), East Asian (EAS), European (EUR), and South Asian (SAS) ancestry. **c**, mmQTL workflow comprises eQTL analysis within each brain region for each resource using a linear mixed model to account for population stratification. Each analysis is then combined using a random-effects meta-analysis that accounts for repeated measures from GTEx samples and effect size heterogeneity across brain regions and resources. Statistical fine-mapping is performed on GWAS and combined eQTL results separately. Finally, fine-mapping posterior probabilities from the eQTL analysis and each GWAS are combined to produce CLPP. **d**, Analysis of data for *THOC7* from one, seven and 13 GTEx brain tissues, with the addition of PsychENCODE and ROSMAP, reduces the size of the 95% credible sets indicated by red points. Statistical fine-mapping for this gene and integration with GWAS nominates a single-candidate causal variant, rs832190, affecting SZ, a combined risk for SZ and BD and sitting height in this region.

false-positive rate in the presence of complex population structure[39–41]. Moreover, expression measurements from multiple brain regions in GTEx are not statistically independent, so combining these data entails explicit modeling of these correlated measurements from the same set of individuals[42].

To realize the potential of multi-ancestry eQTL fine-mapping and integration with brain-related GWAS results, we developed the multivariate multiple QTL (mmQTL) pipeline and applied it to a combined analysis of brain tissues from PsychENCODE, the Religious Orders Study and Memory and Aging Project (ROSMAP) and GTEx. Our pipeline performs eQTL detection with a linear mixed model, identifies conditionally independent eQTLs and combines results across datasets with a random-effects meta-analysis that models the correlation between multiple brain regions from a shared set of individuals. Joint fine-mapping then identifies candidate causal variants shared between gene expression and GWAS traits. This integrative analysis identifies candidate causal variants and elucidates potential regulatory mechanisms for genes underlying schizophrenia (SZ), bipolar disorder (BD) and Alzheimer's disease (AD).

Results

Analysis overview. We performed a multi-ancestry eQTL meta-analysis on RNA-sequenced (RNA-seq) gene expression data from nonoverlapping samples from the dorsolateral prefrontal cortex from PsychENCODE[15] and ROSMAP[43], and from 13 brain regions from GTEx[44] (Fig. 1a). We accounted for diverse ancestry (Fig. 1b) by applying a linear mixed model to the full dataset within each resource and then combined summary statistics from these 15 eQTL analyses using a random-effects meta-analysis to account for effect size heterogeneity and donor overlap between brain regions in GTEx (Fig. 1c). This statistical framework is implemented in our mmQTL software (Methods). Statistical fine-mapping of the eQTL meta-analysis was integrated with GWAS fine-mapping from CAUSALdb[45] to identify candidate causal variants shared between gene expression and neuropsychiatric traits.

For example, results for *THOC7* illustrate that increasing the number of GTEx tissues from one to seven to 13 enhances power and decreases the size of the 95% credible sets, while integration with PsychENCODE and ROSMAP nominates a single-candidate causal variant (Fig. 1d). Integration of GWAS and eQTL results produces colocalization posterior probabilities (CLPP) > 0.05 for SZ, BD and sitting height, and identifies rs832190 and *THOC7* as the candidate causal variant and gene, respectively, for this locus.

Biologically motivated simulations. Simulations motivated by the scenarios considered here (that is, diverse ancestry and repeated-measures design of the human brain datasets) were used to evaluate mmQTL performance in terms of (1) controlling the false-positive rate, (2) leveraging eQTL effects shared across multiple tissues and (3) reducing the size of the credible set from statistical fine-mapping (Fig. 2). For the eQTL analysis we considered a linear regression model including five genotype principal components (PCs) and a linear mixed model that accounts for the genetic similarity between all pairs of samples[39–41]. The summary statistics for each single-nucleotide polymorphism (SNP)–gene pair were aggregated across tissues using a fixed- or random-effects meta-analysis, or simply with the minimum *P* value with Sidak correction to account for the number of tissues. The first two explicitly account for the repeated-measures design by modeling the correlation between summary statistics under the null, while Sidak-corrected minimum *P* values assume independence.

We simulated genotypes for 500 individuals in each of three distinct populations—European, African and Asian. A single causal

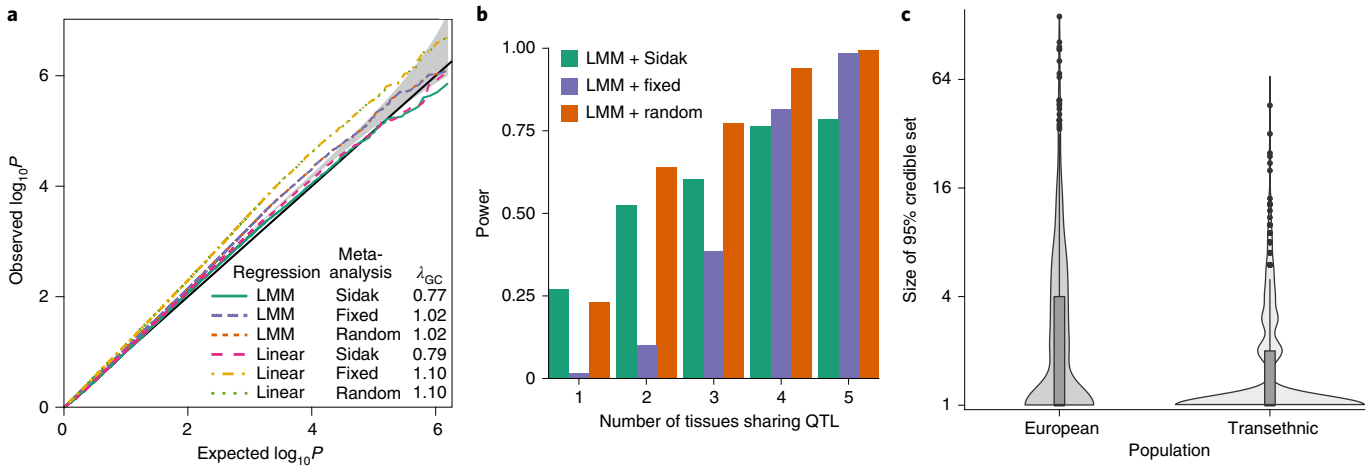


Fig. 2 | Biologically motivated simulations demonstrate performance of mmQTL workflow: high-correlation scenario. a, Q–Q plot of results from null simulation shows that the linear mixed model (LMM) with either fixed- or random-effects meta-analysis accurately controlled the false-positive rate, while linear regression with five genotype PCs did not. The Sidak method was very conservative in both cases. λ_{GC} indicates the genomic control inflation factor, and gray band indicates 95% confidence interval under the null. **b**, Power from LMM followed by three types of meta-analysis versus the number of tissues sharing an eQTL. **c**, Size of the 95% credible sets from statistical fine-mapping for a dataset of European samples versus a multi-ancestry dataset of the same size. Box plot indicates median, interquartile range (IQR) and 1.5 × IQR.

eQTL explaining 1–2% of expression variation in up to five tissues for these 1,500 individuals was simulated for 800 randomly chosen genes where the number of tissues with a shared effect varied from one to five. Correlation between the same gene expression trait measured in two tissues was simulated to be either low ($r = 0.12$) or high ($r = 0.45$) (Methods).

In a null simulation with all genetic effects set to zero in both the low- and high-correlation scenarios, the linear mixed model accurately controlled the false-positive rate when summary statistics from multiple tissues were aggregated using the Sidak method, and for fixed- or random-effects meta-analysis (Fig. 2a and Extended Data Fig. 1). As expected, because the linear model did not adequately account for the complex population structure and showed an inflated false-positive rate, it was not included in subsequent simulations.

Power analyses were performed on the same set of samples of diverse ancestry where the number of tissues with a shared eQTL effect varied between one and five (Fig. 2b). Using a *P* value cutoff of 10^{-6}, the random-effects meta-analysis following a linear mixed-model eQTL analysis had the highest power under most levels of eQTL sharing across tissues, because it models heterogeneity in effect sizes across tissues. The fixed-effects meta-analysis was less powerful because it assumes a shared effect size across tissues. The Sidak-corrected minimum *P* value performed best only when the eQTL was tissue specific (that is, no cross-tissue sharing), since it assumes statistical independence of the results from each tissue.

The mmQTL workflow with the linear mixed model followed by a random-effects meta-analysis demonstrated accurate control of the false-positive rate while retaining high power under biologically motivated simulations. With the goal of identifying candidate causal variants shared with brain-related traits, we evaluated the benefit of using a dataset of diverse ancestry. A dataset of 1,500 European individuals was simulated in addition to the multi-ancestry cohort above. One causal variant with effect size 1.00 ± 0.09% was used to simulate gene expression traits. Statistical fine-mapping of eQTL results from the multi-ancestry cohort produced 95% credible sets containing a mean of 2.0 SNPs compared to a mean of 4.8 for the European-only cohort (Fig. 2c). In the multi-ancestry cohort, 73.0% of genes have a single-candidate causal variant compared to 51.6% in the European cohort. Moreover, random-effects meta-analysis reduces the credible set by 10.0% compared to fixed-effects meta-analysis.

Evaluation of mmQTL workflow on real data. Here we evaluate the empirical performance of our mmQTL workflow on real data by analyzing an increasing number of brain regions ($k = 1$, 4, 7, 13) from GTEx (Fig. 3). As expected, mmQTL is able to borrow information across multiple brain regions using a random-effects meta-analysis so that increasing k substantially increases the empirical effective sample size (n_{eff}) (Fig. 3a). With $k = 13$, there are 2,181 RNA-seq samples from 317 individuals producing empirical $n_{eff} = 1{,}070$. Moreover, increasing k decreases the median size of the 95% credible sets from statistical fine-mapping (Fig. 3b).

The benefit of adding each successive study to the meta-analysis was evaluated for both primary eQTLs and secondary and tertiary conditional eQTLs using a conservative *P* value cutoff of 10^{-6} (Fig. 3c and Methods). The PsychENCODE study included a large cohort and yielded 50.4% of genes having genome-wide significant primary eQTLs. The addition of data from GTEx and ROSMAP produced a combined eQTL analysis comprising 3,983 RNA-seq samples from 2,119 donors to give $n_{eff} = 3{,}154$. Powered by this substantial increase in n_{eff}, eQTLs were detected for 76% of genes analyzed in the final meta-analysis.

Properties of brain eQTL meta-analysis. Our brain eQTL meta-analysis identified 10,769 genes with a genome-wide significant eQTL, including 5,336 with at least one conditional eQTL, using a conservative *P* value threshold of 10^{-6} (Fig. 4a). These eQTL results are highly reproducible, with an estimated replicated rate of $\pi_1 = 73.6\%$ when evaluated in an independent dataset of bulk brain tissue[46] using Storey's π_1 statistic[47]. The increased power from our meta-analysis enables detection of cell-type-specific eQTLs not detectable in smaller studies of bulk brain tissue. eQTLs detected in the granule cell layer of the dentate gyrus enriched for excitatory neurons[27] are replicated in our analysis at $\pi_1 = 77.8\%$ compared to $\pi_1 = 65.2\%$ in the PsychENCODE analysis (one-sided Mann–Whitney *U*-test $P < 1 \times 10^{-16}$), and eQTLs detected in purified microglia (Kosoy et al., in preparation) are replicated in our analysis at $\pi_1 = 68.4\%$ compared to $\pi_1 = 55.0\%$, from the PsychENCODE analysis (one-sided Mann–Whitney *U*-test $P < 1 \times 10^{-16}$) (Fig. 4b).

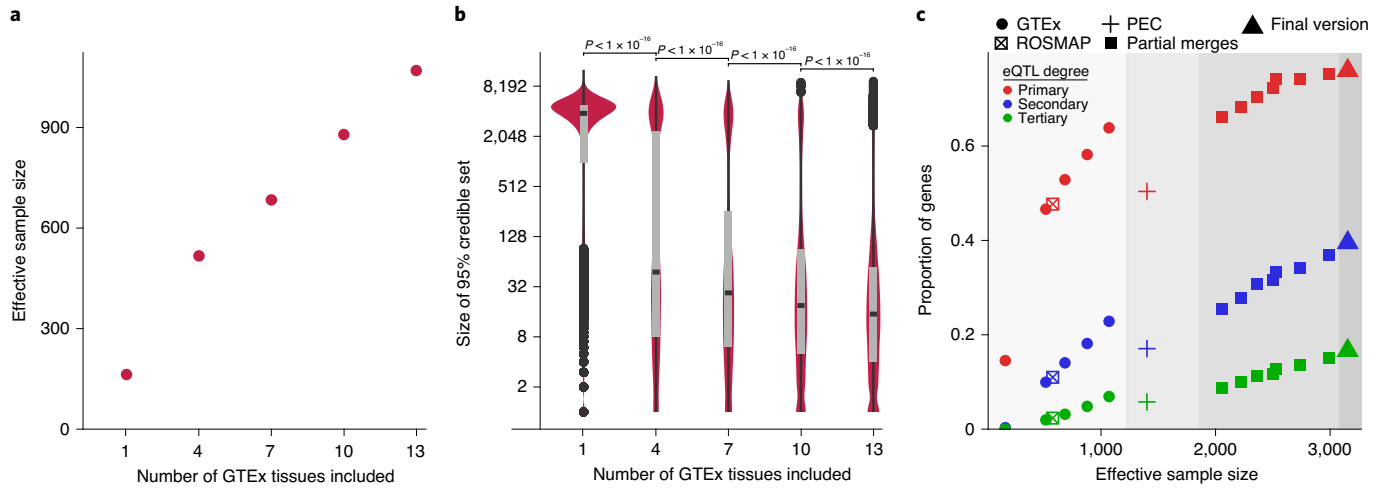


Fig. 3 | Evaluation of mmQTL workflow on real data. a, Increasing the number of brain regions from GTEx increases n_{eff}. **b**, Increasing the number of brain regions from GTEx decreases the median 95% credible set size. *P* values are shown from a one-sided Kolmogorov–Smirnov test between adjacent categories. Box plot indicates median, IQR and 1.5 × IQR. **c**, Inclusion of additional datasets increases the proportion of genes with a detectable primary or conditional eQTL. Colors indicate degree of eQTL. Panel is divided into regions showing (1) GTEx and ROSMAP results; (2) PsychENCODE (PEC) data analyzed here, and published PEC summary statistics[15]; (3) addition of an increasing number of GTEx brain tissues to PEC + ROSMAP results; and (4) the final version, merging PEC + ROSMAP + GTEx.

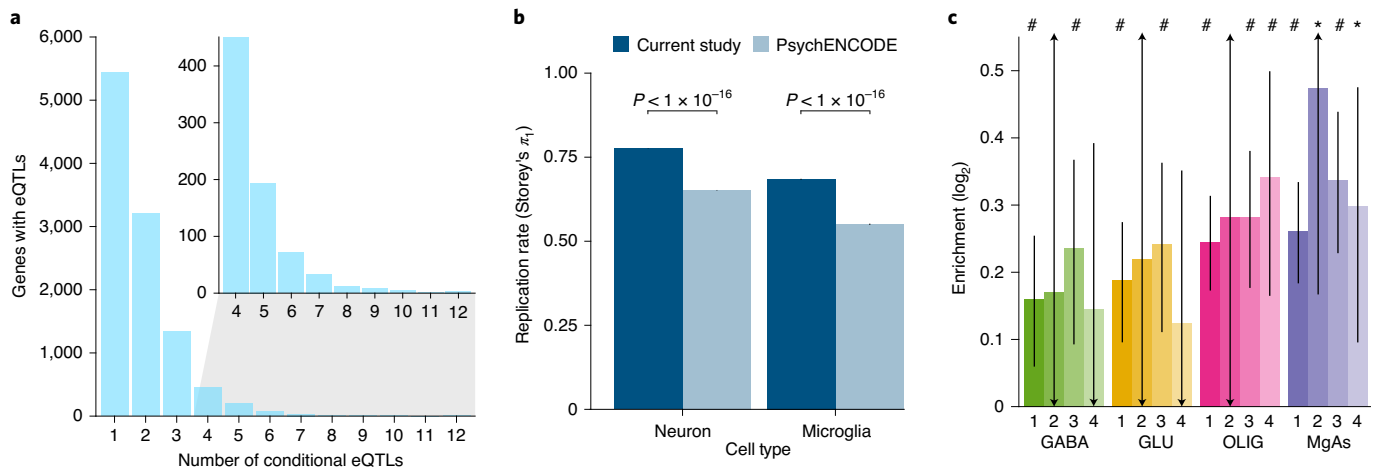


Fig. 4 | Properties of brain eQTL meta-analysis. a, Number of genes having a significant primary or conditional eQTL for degrees 1–12. Inset shows number of genes for eQTL degrees 4–12. **b**, Replication rate as measured by Storey's π_1 in the current study and PsychENCODE for eQTLs discovered in the granule cell layer of the dentate gyrus enriched for excitatory neurons[27], and in purified microglia (Kosoy et al., in preparation). Error bar indicates s.e.m. from 100 bootstrap samplings. *P* value indicates one-sided Mann–Whitney *U*-test. **c**, Bar plot indicating $\log_2$ enrichment of variants in the 95% causal sets for each gene in open chromatin regions assayed in each of four cell populations. Results are shown for eQTL degrees 1–4. Error bars indicate s.d., # indicates Bonferroni adjusted $P < 0.05$ and * indicates nominal $P < 0.05$ for Fisher's exact test.

Moreover, concordance in the sign of the estimated effect sizes between our meta-analysis and cell-type-specific analyses increased with stricter *P* value cutoffs (Extended Data Fig. 2). Overlaying variants in 95% credible sets with assay for transposase-accessible chromatin using sequencing (ATAC–seq) regions identified by fluorescence activated nuclei sorting (FACS) for four cell populations—GABAergic neurons (GABA), glutamatergic neurons (GLU), oligodendrocytes (OLIG) and a mixture of microglia and astrocytes (MgAs)[48]—identified significant enrichment within open chromatin regions for each cell population (Fig. 4c).

The test statistic for random-effects meta-analysis used here comprises the sum of statistics testing the mean (S_{mean}) and variance ($S_{variance}$) of the estimated effect sizes across datasets[49]. Statistical power to detect eQTLs is thus dependent on both effect size and effect size heterogeneity across brain regions. In our analysis an average of 72.2% of power for primary eQTL analysis is attributable to effect size while the remainder is attributable to heterogeneity (Extended Data Fig. 3). Considering only cortical brain regions reduces both effect size heterogeneity and the number of genes with detected eQTLs (from 10,769 to 9,431) but, more importantly, also reduces the number of genes with conditional eQTLs from 5,336 to 3,533.

Conditional eQTLs have different properties to primary eQTLs. While primary eQTLs are a median of 22.8 kb from the transcription start site, conditional eQTLs are more distal with median distances of 35.7 kb for secondary, 50.6 kb for tertiary and 52.7 kb for quaternary

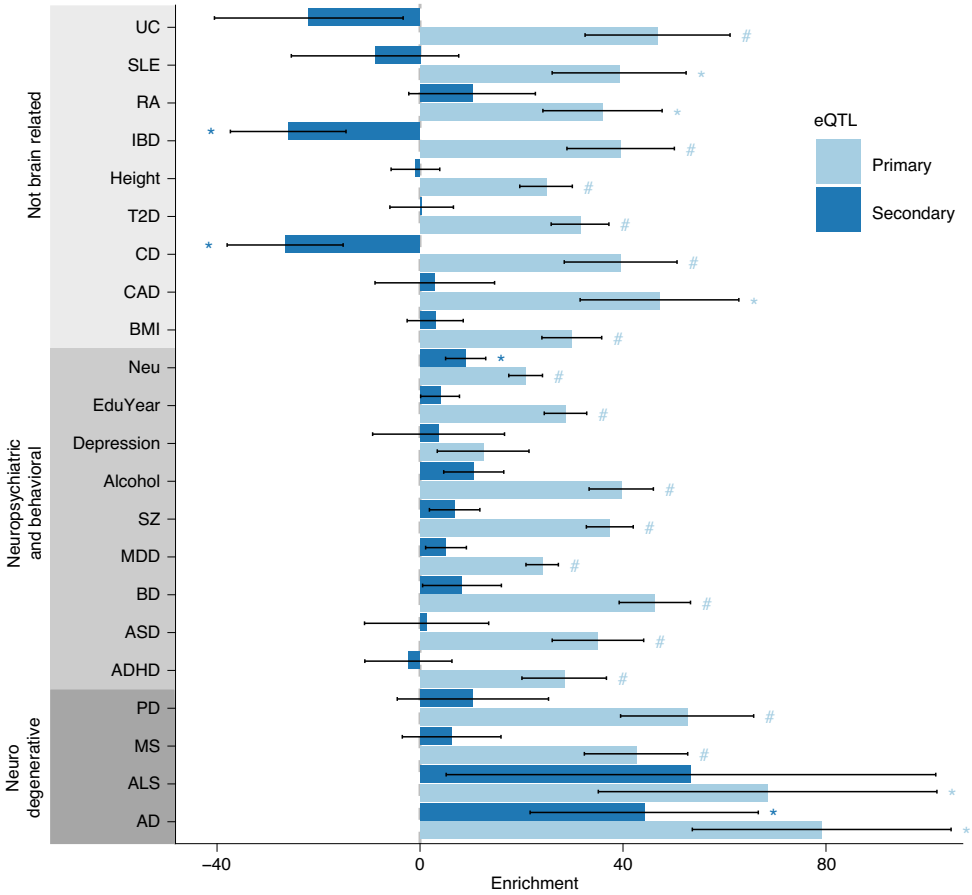


Fig. 5 | Heritability enrichment of variants in the 95% causal set for 22 complex traits. Bar plot indicates LDSC enrichments for variants in the 95% causal set for primary and secondary eQTLs. Error bars indicate standard errors. # indicates that the *P* value passes 5% Bonferroni cutoff for 44 tests, and * indicates $P < 0.05$ from two-sided, one-sample *z*-test. Traits are abbreviated for Ulcerative colitis (UC), Systemic lupus erythematosus (SLE), Rheumatoid arthritis (RA), Inflammatory bowel disease (IBD), Type 2 diabetes (T2D), Crohn's disease (CD), Coronary artery disease (CAD), Body mass index (BMI), Neuroticism (Neu), Education Years (EduYears), Major depressive disorder (MDD), Autism spectrum disorder (ASD), Attention deficit hyperactivity disorder (ADHD), Parkinson's disease (PD), Multiple sclerosis (MS), and Amyotrophic lateral sclerosis (ALS). Supplementary Table 1 shows trait references.

eQTLs ($P < 1.6 \times 10^{-14}$ for all comparisons among primary, secondary and tertiary eQTLs using one-sided Mann–Whitney *U*-test; Extended Data Fig. 4a). This is consistent with primary eQTLs often affecting promoters and conditional eQTLs more often affecting enhancers. In addition, genes with more independent eQTLs have higher cell-type specificity in brain tissue from humans[18] (Spearman rho = 0.0621, $P = 1.93 \times 10^{-10}$) and mice[50] (Spearman rho = 0.0617, $P = 2.50 \times 10^{-10}$) (Extended Data Fig. 4b). Finally, genes with more conditional eQTLs tend to be under lower evolutionary constraint as measured by the probability of loss intolerance (pLI) calculated from large-scale exome sequencing[51]. While 29.9% of genes with no detectable eQTLs are highly constrained (pLI > 0.9), only 6.5% with four eQTLs exceed this cutoff (Extended Data Fig. 4c). While the distribution of estimated effect sizes is similar for increasing conditional eQTL degree, MAF decreases markedly (Extended Data Fig. 5). Interpretation of the estimated effect sizes from bulk and cell-type-specific data is challenging and is affected by multiple factors (Extended Data Fig. 6 and Methods).

Credible set variants are enriched for risk to brain-related traits. Integration of variants in the 95% credible set for primary and conditional eQTLs with large-scale GWAS summary statistics using stratified LD scores regression[52] found significant enrichments across 22 complex traits after accounting for baseline annotations (Fig. 5). Variants in the 95% credible set for primary eQTLs were enriched for 21 traits, including eight neuropsychiatric and behavioral traits and four neurodegenerative diseases. Meanwhile, enrichment for conditional eQTLs was limited to AD, BD and alcohol use. These enrichments indicate that our meta-analysis and statistical fine-mapping captures risk variants for brain-related phenotypes.

Identification of candidate causal risk variants for brain-related traits. Integration of our eQTL fine-mapping results with candidate causal variants from large-scale GWAS[45] using a joint fine-mapping approach[9] identified 7,564 variant–trait pairs (CLPP > 0.01), including 2,102 unique candidate causal variants and 1,666 unique genes among 668 complex traits (Extended Data Fig. 7). These results include 329 variant–trait pairs for 24 brain-related traits for 204 and 189 unique candidate causal variants and genes, respectively (Fig. 6a). Analysis of SZ and BD, two neuropsychiatric diseases with high genetic coheritability[53–56], identified candidate causal variants for 20 genes predicted to confer risk for one or both diseases (Fig. 6b). The top genes with CLPP > 0.5 for either of these diseases include *ZNF823*, *THOC7* and *FURIN*. While these genes have previously been implicated in SZ or bipolar (BP)—and, in fact, the candidate causal variant for *FURIN*, rs4702, has been validated experimentally[33]—candidate causal variants for the other two genes have not previously been identified. Moreover, integration of results from

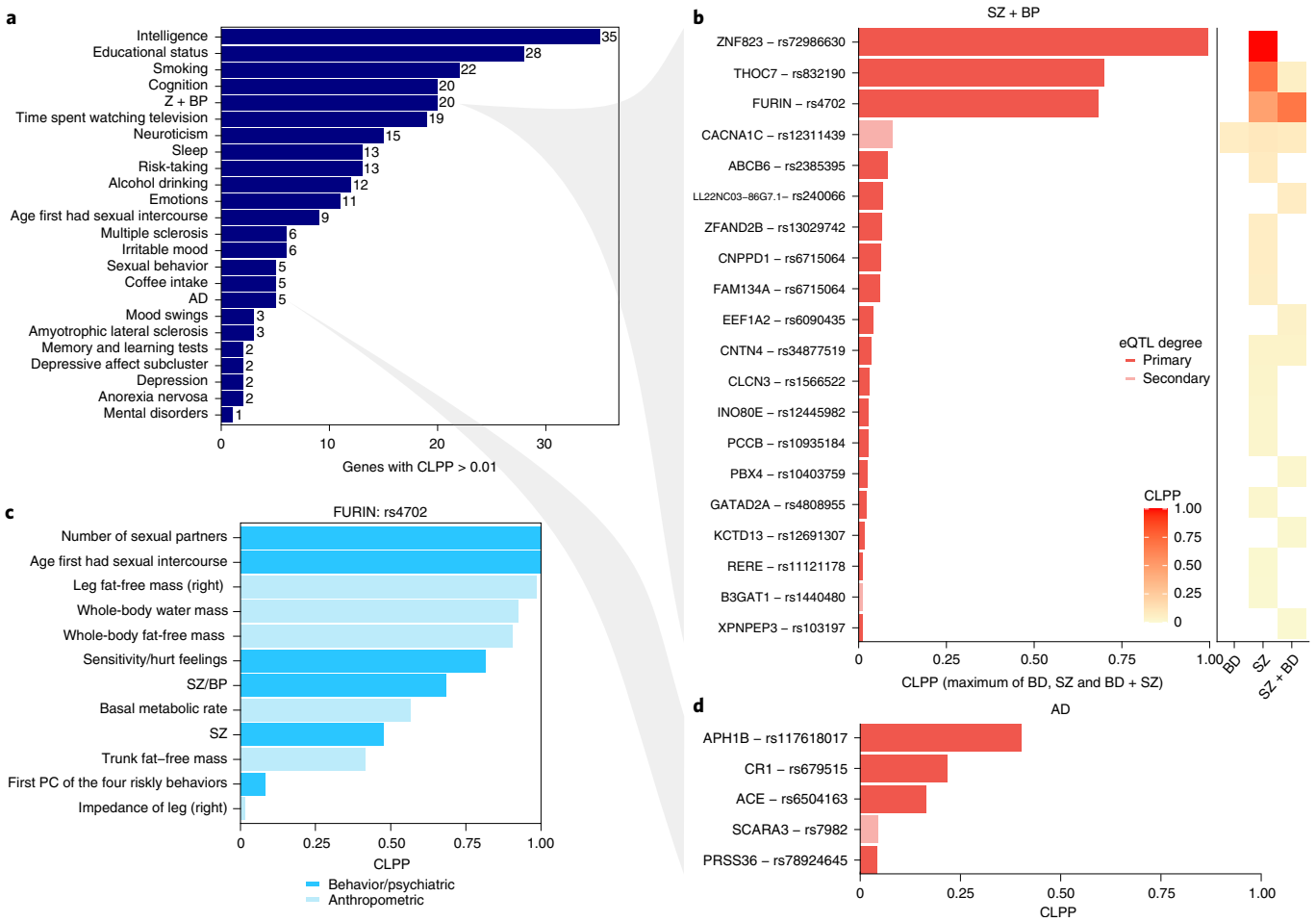


Fig. 6 | Summary of joint fine-mapping colocalization with brain-related traits. a, Number of genes with CLPP > 0.01 for 24 brain-related traits. **b**, For each gene, the maximum CLPP value across SZ, BP and SZ + BP is shown. Right: CLPP for BP, SCZ and SZ + BP compared to controls. **c**, A validated casual variant, rs4702,which affects expression of *FURIN*, is predicted to affect risk for multiple complex behavioral, psychiatric and anthropometric traits. **d**, Genes with CLPP > 0.01 for AD.

analysis of SZ, BP and SZ + BP versus controls indicated the specificity of these candidate causal variants. *ZNF823* is predicted to confer risk to SZ, but not to BD. *THOC7* has a substantially larger CLPP score for SZ compared to the joint SZ + BP GWAS. Conversely, *FURIN* has a higher CLPP for the joint SZ + BP GWAS than for SZ alone. Notably, the candidate causal variant driving colocalization with SZ and BD for *CACNA1C* is in fact a secondary eQTL, emphasizing the importance of including conditional eQTL analysis.

In addition, analysis of candidate causal variables across many phenotypes enables insight into pleiotropy. *FURIN* and rs4702 are also implicated in the number of sexual partners, age at first sexual intercourse, risk-taking behavior and emotional sensitivity/hurt feelings and multiple anthropometric traits (Fig. 6c and Extended Data Fig. 8). Sharing of a candidate causal variant and gene between SZ + BP and these risk-taking behavior traits is particularly interesting, given that impulsiveness is a clinical feature of both SZ and BD[57,58] and is associated with more severe psychiatric symptoms and decreased level of functioning[59].

Analysis of AD identified candidate causal variants for five genes. While these genes have previously been highlighted[1], our analysis highlights variants and their mechanistic link to disease (Fig. 6d).

Candidate causal variants elucidate potential molecular mechanisms. It is known that rs117618017 is the top causal variant for AD and drives the expression of *APH1B*, a subunit of the gamma-secretase complex, which includes multiple AD risk genes as components (Fig. 7a). This missense coding variant was identified in a GWAS meta-analysis for AD[1], but an experimental attempt to validate a functional effect from this single amino acid change yielded only negative results[60]. Nevertheless, our analysis indicates an alternative molecular mechanism whereby, rather than acting by changing protein sequence, the minor allele of rs117618017 increases AD risk by directly increasing gene expression of *APH1B*.

The top hit for SZ is rs72986630, which is predicted to drive expression of *ZNF823*, a zinc finger protein with little additional annotation (Fig. 7b). This C/T SNP is located in the 5' untranslated region of the gene and the minor allele, T (frequency ~6%), is protective against SZ. This variant is predicted to disrupt a binding site for the RE1 silencing transcription factor (REST), also known as neuron-restrictive silencing factor. REST is upregulated during neurogenesis and in adult non-neuronal cells, and acts by silencing neuron-specific genes[61,62]. Analysis of chromatin accessibility in this region using a large-scale ATAC–seq dataset from purified neuronal and non-neuronal nuclei from the anterior cingulate cortex of post mortem brains of 368 donors elucidated the molecular mechanism (Bendl et al., in preparation) (Fig. 7c). In non-neuronal cells, but not in neuronal cells, individuals heterozygous at this site have higher chromatin accessibility at both the 644-bp ATAC–seq peak ($P = 0.016$) and the 21-bp motif ($P = 0.026$), and this corresponds to decreased binding of REST at this site. Since REST is a transcriptional

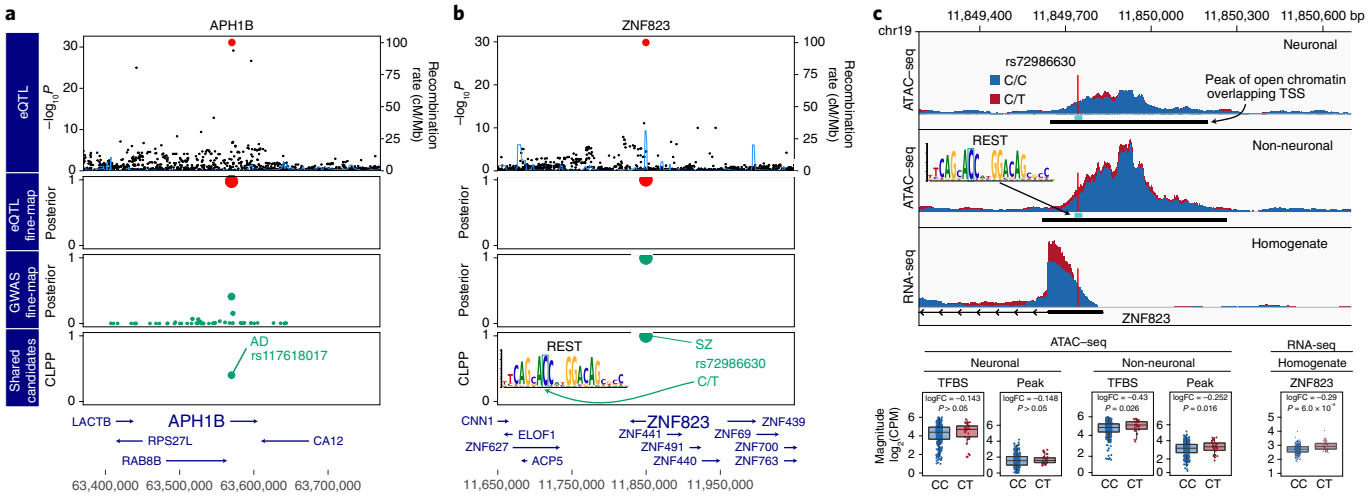


Fig. 7 | GWAS-eQTL colocalization by joint fine-mapping. a,**b**, From the top, the plot shows $-\log_{10}P$ from eQTL analysis, poster probabilities from statistical fine-mapping of eQTL results, poster probabilities from statistical fine-mapping of GWAS results and CLPP for a combination of eQTL and GWAS fine-mapping. **a**, Expression of *APH1B* and AD risk share rs117618017 as a candidate causal variant. **b**, Expression of *ZNF823* and SZ risk share rs72986630 as a candidate causal variant. This variant is predicted to disrupt a REST binding site motif. **c**, Individuals heterozygous for rs72986630 have increased chromatin accessibility at the peak and REST binding site in non-neuronal cells. Genome plot shows chromatin accessibility for neuronal (top) and non-neuronal (middle) nuclei, and gene expression from brain homogenate (bottom). The lower panel shows boxplots comparing chromatin accessibility and gene expression between individuals with two reference alleles (that is, CC) compared to CT heterozygotes. Box plots indicate median, IQR and 1.5 × IQR. Hypothesis tests were performed with limma/voom. FC, fold change; TFBS, transcription factor binding site.

silencer, decreased binding of REST should lead to increased expression of *ZNF823*. Querying RNA-seq data from brain homogenate from these samples confirmed that heterozygous individuals have increased expression of *ZNF823* ($P = 6.01 \times 10^{-9}$).

Discussion

Integration of eQTL and GWAS is a powerful method toward understanding the molecular mechanism influencing complex traits. While transcript-wide association studies aim to identify genes underlying a complex trait, correlated expression and coregulation can be challenging to overcome[63,64]. Joint fine-mapping focuses instead on identifying variants that drive both gene expression and a downstream trait[9]. Despite recent successes, fine-mapping is often limited by statistical power and LD[6,9]. Our mmQTL workflow addresses both of these issues by performing a multi-ancestry eQTL meta-analysis of 3,983 RNA-seq samples from 2,119 donors, with an effective sample size of 3,154, to produce a large resource characterizing the genetics of gene expression in the human brain. This analysis has substantially boosted the catalog of genes with detected conditional eQTLs while increasing the resolution of statistical fine-mapping.

Despite being performed on bulk RNA-seq data, our analysis is able to replicate eQTLs discovered in both purified microglia (Kosoy et al., in preparation) and neurons[27], and the replication rate is substantially higher than for PsychENCODE[15]. Moreover, we identify candidate causal variants enriched in cell-type-specific open chromatin regions. While much recent work has pursued the generation of eQTLs from purified cell populations[26–28], and eQTL discovery from single-cell/nucleus RNA-seq is becoming tractable[65,66], our eQTL meta-analysis from bulk tissue illustrates that large sample size and sophisticated statistical modeling have substantial power to replicate eQTLs from smaller studies of purified cell types.

While the number of genes with detectable eQTLs approaches saturation, there is substantial value in increasing sample size. Here we use individuals of diverse ancestry paired with a linear mixed model in our mmQTL workflow to increase the resolution of statistical fine-mapping. Moreover, we perform conditional eQTL analysis to identify genes with up to 12 independent eQTLs. These conditional eQTLs tend to have lower MAF, be further from transcription start sites and affect genes that are more cell-type specific. The number of genes with secondary and tertiary eQTLs does not appear close to saturation, underscoring the regulatory variation that remains to be identified.

Integration of statistical fine-mapping for eQTLs and GWAS across hundreds of complex traits enabled insight into candidate causal variants, mechanisms of disease genetics and pleiotropy. Focusing on regulatory mechanisms for genes underlying brain-related traits, we identified 20 genes and candidate causal variants predicted to drive risk for SZ and BD, plus a further five for AD. While other methods focus on discovering disease genes, here we focus on discovering gene–variant pairs underlying disease risk to elucidate the molecular mechanisms that convey risk.

Here we highlighted two examples. The SNP rs117618017 is a candidate causal variant causing a single amino acid change in *APH1B*. While experimental results of the impact of this amino acid change were negative[60], our analysis instead supports a mechanism where this variant increases disease risk by increasing expression of APH1B. Our analysis predicts that rs72986630 drives expression of *ZNF823* and is protective against SZ. By integration of chromatin accessibility data from post mortem brains, we traced the predicted chain of causality and found that the minor allele disrupts binding of REST in non-neuronal cells, which then increases expression of *ZNF823*. The lack of an effect in neuronal nuclei is consistent with the higher expression of REST in non-neuronal cells during adulthood, silencing neuron-specific genes[61,62].

While we focused on regulatory mechanisms for genes underlying SZ, BD and AD, all results are available from the Brain eQTL meta-analysis resource (icahn.mssm.edu/brema).

Further integration of multiomics data with multi-ancestry fine-mapping and large-scale GWAS promises to yield further insight into the molecular mechanisms underlying disease risk. Future studies are poised to perform multiple genomic assays, namely RNA-seq and ATAC–seq, on multiple tissues or brain regions, and to target multiple cell types either by sorting or single-cell/nucleus methods[65,66].

Moreover, these studies will increasingly include individuals of diverse ancestry[67]. Our mmQTL method will enable the field to take advantage of these repeated-measures datasets while modeling effect size heterogeneity and controlling the false-positive rate. Efforts to trace the chain of causality from variants and molecular mechanisms to pleiotropy across complex phenotypes are poised to yield insight into novel therapeutic targets.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-021-00987-9.

Received: 18 January 2021; Accepted: 17 November 2021;
Published online: 20 January 2022

References

1. Jansen, I. E. et al. Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
2. Schizophrenia Working Group of the Psychiatric Genomics Consortium. Biological insights from 108 schizophrenia-associated genetic loci. *Nature* **511**, 421–427 (2014).
3. Visscher, P. M. et al. 10 Years of GWAS discovery: biology, function, and translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).
4. Nalls, M. A. et al. Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).
5. Wray, N. R. et al. Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nat. Genet.* **50**, 668–681 (2018).
6. Schaid, D. J., Chen, W. & Larson, N. B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat. Rev. Genet.* **19**, 491–504 (2018).
7. Gallagher, M. D. & Chen-Plotkin, A. S. The post-GWAS era: from association to function. *Am. J. Hum. Genet.* **102**, 717–730 (2018).
8. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
9. Hormozdiari, F. et al. Colocalization of GWAS and eQTL signals detects target genes. *Am. J. Hum. Genet.* **99**, 1245–1260 (2016).
10. Kim-Hellmuth, S. et al. Cell type-specific genetic regulation of gene expression across human tissues. *Science* **369**, eaaz8528 (2020).
11. Dobbyn, A. et al. Landscape of conditional eQTL in dorsolateral prefrontal cortex and co-localization with schizophrenia GWAS. *Am. J. Hum. Genet.* **102**, 1169–1184 (2018).
12. Hormozdiari, F. et al. Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nat. Genet.* **50**, 1041–1047 (2018).
13. Fromer, M. et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat. Neurosci.* **19**, 1442–1453 (2016).
14. Ng, B. et al. An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. *Nat. Neurosci.* **20**, 1418–1426 (2017).
15. Wang, D. et al. Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**, eaat8464 (2018).
16. Jaffe, A. E. et al. Developmental and genetic regulation of the human cortex transcriptome illuminate schizophrenia pathogenesis. *Nat. Neurosci.* **21**, 1117–1125 (2018).
17. Habib, N. et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. *Nat. Methods* **14**, 955–958 (2017).
18. Darmanis, S. et al. A survey of human brain transcriptome diversity at the single cell level. *Proc. Natl Acad. Sci. USA* **112**, 7285–7290 (2015).
19. Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
20. Cao, J. et al. A human cell atlas of fetal gene expression. *Science* **370**, eaba7612 (2020).
21. Raj, T. et al. Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes. *Science* **344**, 519–523 (2014).
22. van der Wijst, M. G. P. et al. Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
23. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
24. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).
25. Farh, K. K.-H. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
26. Young, A. M. H. et al. A map of transcriptional heterogeneity and regulatory variation in human microglia. *Nat. Genet.* **53**, 861–868 (2021).
27. Jaffe, A. E. et al. Profiling gene expression in the human dentate gyrus granule cell layer reveals insights into schizophrenia and its genetic risk. *Nat. Neurosci.* **23**, 510–519 (2020).
28. de Paiva Lopes, K. et al. Atlas of genetic effects in human microglia transcriptome across brain regions, aging and disease pathologies. Preprint at https://www.biorxiv.org/content/10.1101/2020.10.27.356113v1?rss=1 (2020).
29. Jansen, R. et al. Conditional eQTL analysis reveals allelic heterogeneity of gene expression. *Hum. Mol. Genet.* **26**, 1444–1451 (2017).
30. Zhernakova, D. V. et al. Identification of context-dependent expression quantitative trait loci in whole blood. *Nat. Genet.* **49**, 139–145 (2017).
31. Benner, C. et al. FINEMAP: efficient variable selection using summary data from genome-wide association studies. *Bioinformatics* **32**, 1493–1501 (2016).
32. Hormozdiari, F., Kostem, E., Kang, E. Y., Pasaniuc, B. & Eskin, E. Identifying causal variants at loci with multiple signals of association. *Genetics* **198**, 497–508 (2014).
33. Schrode, N. et al. Synergistic effects of common schizophrenia risk variants. *Nat. Genet.* **51**, 1475–1485 (2019).
34. Nott, A. et al. Brain cell type-specific enhancer-promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
35. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
36. Zaitlen, N., Paşaniuc, B., Gur, T., Ziv, E. & Halperin, E. Leveraging genetic variability across populations for the identification of causal variants. *Am. J. Hum. Genet.* **86**, 23–33 (2010).
37. Asimit, J. L., Hatzikotoulas, K., McCarthy, M., Morris, A. P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur. J. Hum. Genet.* **24**, 1330–1336 (2016).
38. Morris, A. P. Transethnic meta-analysis of genomewide association studies. *Genet. Epidemiol.* **35**, 809–822 (2011).
39. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
40. Yang, J., Zaitlen, N. A., Goddard, M. E., Visscher, P. M. & Price, A. L. Advantages and pitfalls in the application of mixed-model association methods. *Nat. Genet.* **46**, 100–106 (2014).
41. Sul, J. H., Martin, L. S. & Eskin, E. Population structure in genetic studies: confounding factors and mixed models. *PLoS Genet.* **14**, e1007309 (2018).
42. Han, B. et al. A general framework for meta-analyzing dependent studies with overlapping subjects in association mapping. *Hum. Mol. Genet.* **25**, 1857–1866 (2016).
43. Bennett, D. A. et al. Religious orders study and rush memory and aging project. *J. Alzheimers Dis.* **64**, S161–S189 (2018).
44. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
45. Wang, J. et al. CAUSALdb: a database for disease/trait causal variants identified using summary statistics of genome-wide association studies. *Nucleic Acids Res.* **48**, D807–D816 (2020).
46. Wang, M. et al. The Mount Sinai cohort of large-scale genomic, transcriptomic and proteomic data in Alzheimer's disease. *Sci. Data* **5**, 180185 (2018).
47. Storey, J. D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl Acad. Sci. USA* **100**, 9440–9445 (2003).
48. Hauberg, M. E. et al. Common schizophrenia risk variants are enriched in open chromatin regions of human glutamatergic neurons. *Nat. Commun.* **11**, 5581 (2020).
49. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).
50. Zeisel, A. et al. Brain structure. Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq. *Science* **347**, 1138–1142 (2015).
51. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
52. Finucane, H. K. et al. Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
53. Pardiñas, A. F. et al. Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nat. Genet.* **50**, 381–389 (2018).
54. Bipolar Disorder and Schizophrenia Working Group of the Psychiatric Genomics Consortium. Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell* **173**, 1705–1715 (2018).
55. Cross-Disorder Group of the Psychiatric Genomics Consortium. Genomic relationships, novel loci, and pleiotropic mechanisms across eight psychiatric disorders. *Cell* **179**, 1469–1482 (2019).

56. Roussos, P. et al. A role for noncoding variation in schizophrenia. *Cell Rep.* **9**, 1417–1429 (2014).
57. Najt, P. et al. Impulsivity and bipolar disorder. *Eur. Neuropsychopharmacol.* **17**, 313–320 (2007).
58. Ouzir, M. Impulsivity in schizophrenia: a comprehensive update. *Aggress. Violent Behav.* **18**, 247–254 (2013).
59. Cerimele, J. M. & Katon, W. J. Associations between health risk behaviors and symptoms of schizophrenia and bipolar disorder: a systematic review. *Gen. Hosp. Psychiatry* **35**, 16–22 (2013).
60. Zhang, X. et al. Negative evidence for a role of APH1B T27I variant in Alzheimer's disease. *Hum. Mol. Genet.* **29**, 955–966 (2020).
61. Hwang, J.-Y. & Zukin, R. S. REST, a master transcriptional regulator in neurodegenerative disease. *Curr. Opin. Neurobiol.* **48**, 193–200 (2018).
62. Schoenherr, C. J. & Anderson, D. J. The neuron-restrictive silencer factor (NRSF): a coordinate repressor of multiple neuron-specific genes. *Science* **267**, 1360–1363 (1995).
63. Mancuso, N. et al. Probabilistic fine-mapping of transcriptome-wide association studies. *Nat. Genet.* **51**, 675–682 (2019).
64. Wainberg, M. et al. Opportunities and challenges for transcriptome-wide association studies. *Nat. Genet.* **51**, 592–599 (2019).
65. van der Wijst, M. et al. The single-cell eQTLGen consortium. *eLife* **9**, e52155 (2020).
66. Mandric, I. et al. Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nat. Commun.* **11**, 5504 (2020).
67. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).

Methods

Obtaining and processing of RNA-seq datasets. Imputed genotypes from GTEx v.8 were downloaded from dbGAP (accession no. phs000424.v8.p2). For ROSMAP, imputed genotypes were downloaded from the Synapse website (id: syn3157329). Imputed genotypes for each cohort in the PsychENCODE study were downloaded from the Synapse website (id: syn21052530) and were then filtered to retain variants with imputation quality ≥0.3. Filtered genotypes from each cohort were merged, and variants with MAF ≥1% and Hardy–Weinberg equilibrium $P \geq 10^{-6}$ were retained.

The original PsychENCODE analysis performed eQTL detection using 1,387 individuals[15]. In the current work we exclude a small fraction of these individuals. First, the full PsychENCODE dataset contains GTEx samples, which we excluded to avoid redundancy with our separate GTEx analysis. Second, the original analysis used ~5 million imputed SNPs. Since accurate statistical fine-mapping depends on inclusion of the true causal variant in the analysis, we included additional, well-imputed SNPs at the cost of excluding a small set of samples. Exclusion of samples with <8 million well-imputed (info score >0.3) variants yielded the 1,215 individuals used in this study.

The normalized gene expression of GTEx v.8 was downloaded from GTEx Portal (GTEx_Analysis_v8_eQTL_expression_matrices.tar, https://gtexportal.org/), and we regressed out covariates from the companion file `GTEx_Analysis_v8_eQTL_covariates.tar` with linear regression. Normalized data from PsychENCODE (`DER-01_PEC_Gene_expression_matrix_normalized.txt`) were downloaded from http://resource.psychencode.org/ and, as the downloaded gene expression is already normalized, to regress out the effect of covariates no further normalization was undertaken. Data from ROSMAP (syn3388564, `ROSMAP_RNAseq_FPKM_gene.tsv`) were downloaded from https://adknowledgeportal.synapse.org. The values of fragments per kilobase exon per million mapped reads abundance provided were quantile normalized, $\log_2$ transformed and standardized to normal distribution, and 20 PCs of the gene expression matrix were regressed out.

Linear mixed-model eQTL analysis. Given the expression abundance of a gene measured in n tissues from the same set of individuals, gene expression in tissue t can be modeled as:

$$y_{i,t} = x_{i,j}\beta_{j,t} + \sum_{k=1}^{m} x_{i,k}\alpha_{k,t} + \varepsilon_{i,t} \quad (1)$$

where $y_{i,t}$ is the measured gene expression value for individual i in tissue t, which has been normalized so that it has mean 0 and variance 1; $x_{i,j}$ is the genotype dosage for individual i at variant j, normalized so that it has mean 0 and variance 1; and $\beta_{j,t}$ is effect size for variant j and tissue t. The next term models the polygenic background across m variants, where $x_{i,k}$ is the genotype dosage value for individual i at variant k and $\alpha_{k,t}$ is the effect size for variant k and tissue t with distribution $\mathcal{N}\left(0, \sigma^2_{g_t}\right)$, where $\sigma^2_{g_t}$ is the tissue-specific parameter for genetic background. Finally, $\varepsilon_{i,t}$ is the normally distributed error variance for individual i and tissue t with distribution $\mathcal{N}\left(0, \sigma^2_{\varepsilon_t}\right)$, where $\sigma^2_{\varepsilon_t}$ is the tissue-specific parameter for random noise.

This linear mixed model can be transformed for practical estimation of effect size $\beta_{j,t}$. Equation (1) can be rewritten as

$$y_{i,t} = x_{i,j}\beta_{j,t} + \widehat{\varepsilon_{i,t}} \quad (2)$$

where $\widehat{\varepsilon_{i,t}} = \sum_{k=1}^{m} x_{k,j}x_{k,j}\alpha_{k,j} + \varepsilon_{i,t}$ and has a distribution $\mathcal{N}\left(0, K\sigma^2_{g_t} + \sigma^2_{\varepsilon_t}\right)$, with K is a genetic relatedness matrix estimated based on genome-wide genotypes.

Considering that the phenotype was collected among l individuals, we can write equation (2) into a vector format:

$$Y_t = X_j\beta_{j,t} + \widehat{\varepsilon_t} \quad (3)$$

where Y_t, X_j and $\widehat{\varepsilon_t}$ are l-dimensional vectors and contain normalized phenotype, normalized genotype of variant j and noise, respectively.

From equation (3), $\beta_{j,t}$ can be estimated as

$$\widehat{\beta_{j,t}} = \left(X_j^T V^{-1} X_j\right)^{-1}\left(X_j^T V^{-1} Y_t\right) \quad (4)$$

where $V = K\sigma^2_{g_t} + \sigma^2_{\varepsilon_t}$ and produces an unbiased estimator since

$$\begin{aligned} E\left(\widehat{\beta_{j,t}}\right) &= E\left(\left(X_j^T V^{-1} X_j\right)^{-1}\left(X_j^t V^{-1}\right)\left(X_j\beta_{j,t} + \widehat{\varepsilon_t}\right)\right) \\ &= \beta_{j,t} + E\left(\left(X_j^T V^{-1} X_j\right)^{-1} X_j^t V^{-1}\widehat{\varepsilon_t}\right) \\ &= \beta_{j,t} + E\left(\widetilde{\varepsilon_t}\right) \\ &= \beta_{j,t}. \end{aligned}$$

Modeling covariance across tissues. While standard meta-analysis assumes that effect size estimates are statistically independent, analysis of multiple tissues from the same set of subjects produces covariance between the coefficient estimates. Here we explicitly model this covariance to control the false-positive rate.

Denote the estimate for variant j across all tissues as the vector $\widehat{\vec{\beta}_j} = \left[\widehat{\beta_{j,1}}, \widehat{\beta_{j,2}}, \ldots, \widehat{\beta_{j,l}}\right]$. Since individuals overlap across multiple tissues, the entries of $\widehat{\vec{\beta}_j}$ will be correlated. Estimation of coefficients for tissues 1 and 2 using equation (4) gives

$$\widehat{\beta_{j,1}} = \left(X_1^T V_1^{-1} X_1\right)^{-1}\left(X_1^T V_1^{-1} Y_1\right) \quad (5)$$

$$\widehat{\beta_{j,2}} = \left(X_2^T V_2^{-1} X_2\right)^{-1}\left(X_2^T V_2^{-1} Y_2\right) \quad (6)$$

where an index is added to distinguish two tissues that may have partial sample overlapping. These estimates are not statistically independent, since

$$\begin{aligned} E\left(\widehat{\beta_{j,1}}\widehat{\beta_{j,2}}\right) &= E\left(\left(X_1^T V_1^{-1} X_1\right)^{-1}\left(X_1^T V_1^{-1} Y_1\right)\left(X_2^T V_2^{-1} X_2\right)^{-1}\left(X_2^T V_2^{-1} Y_2\right)\right) \\ &= E\left(\left(X_1^T V_1^{-1} X_1\right)^{-1}\left(X_1^T V_1^{-1}\left(X_1\beta_{j,1} + \widehat{\varepsilon_1}\right)\right)\right. \\ &\quad \left.\times\left(X_2^T V_2^{-1} X_2\right)^{-1}\left(X_2^T V_2^{-1}\left(X_2\beta_{j,2} + \widehat{\varepsilon_2}\right)\right)\right) \\ &= E\left(\beta_{j,1}\beta_{j,2}\right) + E\left(C\widehat{\varepsilon_1}\widehat{\varepsilon_2}\right) \end{aligned}$$

where $C = \left(X_1^T V_1^{-1} X_1\right)^{-1} X_1^T V_1^{-1} V_2^{-1} X_2\left(X_2^T V_2^{-1} X_2\right)^{-1}$ is involved only with transformed genotypes projected by a covariance matrix. Noting that $\widehat{\varepsilon_1}$ and $\widehat{\varepsilon_2}$ are the summed contribution from polygenic background and noise, if there is sample overlapping and the phenotypes share causal variants in two tissues, then

$$E\left(C\widehat{\varepsilon_1}\widehat{\varepsilon_2}\right) = Cn_{\text{shared}}\sigma_{g,1}\sigma_{g,2} \neq 0,$$

where n_{shared} is the number of shared individuals and $\sigma_{g,1}$ and $\sigma_{g,2}$ are the genetic component for polygenic background in tissues 1 and 2, respectively. Finally, we note that

$$\text{cov}\left(\widehat{\beta_{j,1}}, \widehat{\beta_{j,2}}\right) = E\left(\left(\widehat{\beta_{j,1}} - \beta_{j,1}\right)\left(\widehat{\beta_{j,2}} - \beta_{j,2}\right)\right) = E\left(C\widehat{\varepsilon_1}\widehat{\varepsilon_2}\right),$$

explicitly indicating that there is non-zero covariance between estimators. Our mmQTL method estimated the covariance matrix among n tissues based on the nonsignificant z-score in tissues, and set it to be $\hat{C}$. This matrix is defined so that the covariance between tissues i and j is estimated by the covariance between z-scores from nonsignificant variants ($P > 0.05$) according to:

$$\widehat{C} = \text{cov}\left(Z_i, Z_j\right),$$

where Z_i and Z_j are vectors containing statistical z-scores.

Fixed- and random-effects meta-analysis. The results from multiple analyses are aggregated using either a fixed- or random-effects meta-analysis. The true effects sizes are assumed to be drawn from a normal distribution $\mathcal{N}\left(\beta, \sigma^2_\beta\right)$ centered on the true effect size β with variance σ^2_β. For a fixed-effect model, the true effect size is fixed at a constant value equivalent to setting $\sigma^2_\beta = 0$ and, for the random-effects model, $\sigma^2_\beta \geq 0$. From this hierarchical framework we obtain estimators for variant j among tissues, denoted as a vector $\hat{\beta}_j = \beta_{j1} + \tilde{\varepsilon}_j$ which has a distribution $\mathcal{N}\left(\beta_{j1}, \sigma^2_{g_j} I + \hat{C}\right)$. We applied the Brent method implemented in the C++ Boost library to estimate β_j and $\sigma^2_{g_j}$. To test the difference with null hypothesis, we applied the random-effects model[42,49] to obtain a P value.

Statistically, standard fixed-effects mega-analysis combines all data into a single regression model and assumes a fixed-effects size, as well as constant error variance, across all studies. These assumptions are not satisfied in multitissue eQTL analyses due to variation in effect size and variation in error variance across tissues[68]. Use of a random-effects meta-analysis addresses both of these issues to retain control of the false-positive rate while leveraging effect size heterogeneity to increase power.

Detection of conditional eQTLs. We applied a stepwise selection strategy to explore the *cis*-region and identify conditionally independent eQTL associations. An iterative strategy was applied to determine conditional independent eQTLs: previously detected eQTL signals were regressed out and a further round of eQTL detection initiated. If one or more variants had $P < 0^{-6}$, the variant with the lowest P value was added to the list of conditionally independent effects. The process is repeated until no additional variant has $P < 10^{-6}$. If a high-order eQTL is in high LD with a low-order eQTL ($r^2 \geq 0.3$), the former will be excluded to avoid attenuating the estimated effect size of the latter.

Importantly, we demonstrate statistically that the order in which conditional eQTLs is detected is biologically meaningful: large-effect eQTLs shared among tissues are likely to be detected first, while small-effect eQTLs or tissue-specific effects will be detected as higher-order eQTLs.

Consider two true causal variants i and j, where the estimated effect has the distribution around the true value according to $\widehat{\beta}_i \sim \mathcal{N}\left(\beta_i, \sigma_i^2 I + \widehat{C}_i\right)$, where $\widehat{C}_i$ is defined above. The noncentrality parameter (NCP) reflecting the statistical power for this variant is

$$\mathrm{NCP}_i = \frac{\beta_i}{\sqrt{\left(tr\left(\left(\sigma_i^2 I + \widehat{C}_i\right)^{-1}\right)\right)^{-1}}}.$$

The ratio between the NCP of variants i and j is

$$\frac{\mathrm{NCP}_i}{\mathrm{NCP}_j} = \frac{\beta_i}{\beta_j} \Big/ \sqrt{\frac{tr\left(\left(\sigma_j^2 I + \widehat{C}_j\right)^{-1}\right)}{tr\left(\left(\sigma_i^2 I + \widehat{C}_i\right)^{-1}\right)}} \quad (7)$$

From empirical observation that the effect size (with the genotype and response normalized) of primary eQTLs is larger than that of nonprimary eQTLs, $\widehat{C}_i \approx \widehat{C}_j$ and both are positive definite and can be decomposed into $U\Sigma U$, in which U consists of the eigenvectors and Σ is a diagonal matrix with elements being eigenvalues, denoted as diag $(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_N)$, $\sigma_i^2 I + \widehat{C}_i$ can be decomposed into $U\left(\Sigma + I\sigma_i^2\right)U$ and $\sigma_j^2 I + \widehat{C}_j$ to be $U\left(\Sigma + I\sigma_j^2\right)U$. Therefore, equation (7) can be rewritten as

$$\frac{\mathrm{NCP}_i}{\mathrm{NCP}_j} = \frac{\frac{\beta_i}{\beta_j}}{\frac{\sum_k^N\left(\sigma_j^2+\lambda_k\right)^{-1/2}}{\sum_k^N\left(\sigma_i^2+\lambda_k\right)^{-1/2}}}.$$

It is apparent that the difference in statistical power for variants i and j is mainly determined by effect size and its variance. For a variant with larger effect size and smaller variance, it has a higher statistical power consistent with the empirical that primary eQTL has a much larger normalized effect size ($\beta_i > \beta_j$) and smaller variance, because of the sharing among tissues $\left(\sigma_j^2 > \sigma_i^2\right)$, so $\frac{\mathrm{NCP}_i}{\mathrm{NCP}_j} > 1$. It follows that, on average, mmQTL will pick the independent eQTL signal in a biologically meaningful manner so that eQTLs with a larger influence on expression abundance among conditions tend to be selected first.

Multiple testing. Multiple testing correction was performed at the locus level as well as genome wide. Empirically we performed the locus-level control applying Bonferroni correction, which is a most conservative strategy, and the Benjamini–Hochberg method[69] on genome-wide correction; we found $P = 10^{-6}$ sufficient for two-level, multiple-test correction. While studies often use more liberal multiple testing cutoffs because of limited statistical power, the statistical fine-mapping that is the focus of this analysis can perform poorly on genes that pass only a liberal cutoff[9,12].

We note that Wang et al.[15] performed a mega-analysis of PsychENCODE data and used a permutation method to compute false-discovery rate (FDR) and P value cutoff. In their analysis, FDR 5% cutoff empirically corresponds to $P < 8.3 \times 10^{-4}$. Nevertheless, the complexity of applying a permutation approach to linear mixed models[70], and the use of a random-effects meta-analysis afterwards in this analysis, made a computationally efficient permutation approach impractical here.

Compution of empirical effective sample size. In the linear regression model for QTL analysis, given that both phenotype and genotype were normalized, the estimator for the allelic effect size is $\hat{\beta} = \left(X^T X\right)^{-1} XY$ and its variance is $\mathrm{var}\left(\hat{\beta}\right) = \left(X^T X\right)^{-1}\sigma_e^2$. Letting R_i^2 be the variance explained by the explored variant and n_i the (effective) sample size for study i, the variance of the effect size estimate is $\mathrm{var}\left(\widehat{\beta}_i\right) = \frac{1-R_i^2}{n_i}$. Consider two studies, where the (effective) sample size of the first is easy to estimate simply by using the number of samples and the second has some confounding factors such as repeat measurements or population structure. Assuming that the effect size of a given causal variant is constant in the two studies, the ratio of variances is determined only by n_1 and n_2:

$$\frac{\mathrm{var}\left(\widehat{\beta}_2\right)}{\mathrm{var}\left(\widehat{\beta}_1\right)} = \frac{n_1}{n_2}.$$

Therefore, the effective sample size, n_2, can be computed from known values by $n_2 = \frac{n_1 \mathrm{var}\left(\widehat{\beta}_1\right)}{\mathrm{var}\left(\widehat{\beta}_2\right)}$.

We used individual brain tissues in the GTEx dataset as study 1 to define n_1, and eQTL results from fixed-effects meta-analysis as study 2. The genome-wide variance ratio was set to be the median ratio of variances based on all variants with abs (z-score) ≥10 in the fixed-effects meta-analysis. When evaluating the effective size of a meta-analysis, effective sample size was computed by treating each brain tissue in GTEx as baseline and then taking the mean estimated effective sample size over 13 brain regions.

Replication of eQTLs from purified cell types. To assess the replication of eQTLs discovered in independent datasets, we considered the lead SNP for each gene with a genome-wide significant eQTL in the granule cell layer of the dentate gyrus enriched for excitatory neurons[27] and purified microglia (Kosoy et al., in preparation). For the set of lead SNPs from each dataset, P values were extracted from both the current eQTL and PsychENCODE analyses[15], and Storey's π_1 was evaluated using q-value[47]. PsychENCODE P values were obtained from http://resource.psychencode.org/Datasets/Derived/QTLs/Full_hg19_cis-eQTL.txt.gz. Uncertainty in π_1 estimates was evaluated using 100 bootstraps where SNPs were sampled with replacement and π_1 was recomputed each time. A P value comparing the replication rate for the current and PsychENCODE analyses was computed using a one-sided z-test using the estimated π_1 values and their bootstrap variances.

Interpretation of estimated effect sizes. In the scenario where a SNP has a large cell-type-specific effect on gene expression, the true biological effect will be attenuated in bulk data comprising multiple cell types. Nevertheless, testing this biological intuition through eQTL analysis is challenging for a number of technical reasons. Unfortunately, eQTL analysis does not directly estimate biological effect size because gene expression is typically $\log_2$ transformed, scaled to have variance 1 and often quantile normalized. In addition, the inclusion of PEER factors or other covariates can account for cell-type heterogeneity across samples in the data. Therefore, the estimated eQTL effect size reflects the association between SNP and (transformed) gene expression after accounting for other variables.

Furthermore, the technical process of obtaining gene expression from bulk tissue versus cell-type-specific samples is susceptible to different noise profiles based on differing protocols and the biological condition of the physical samples. In fact, Young et al.[26] found that cell-type-specific samples from microglia are noisier than those from bulk samples, and that reported estimated effect size in purified cells is attenuated.

We performed an empirical analysis of the estimated allelic effect sizes from the lead eQTL variants for each gene in our meta-analysis of bulk data, compared to estimates from cell-type-specific data from neurons[27] and microglia from Kosoy et al. (in preparation) and Young et al.[26]. Comparison between bulk and neuron-enriched data gives a slope of 0.59, indicating that the slope is on average actually smaller in cell-type data. Comparison to the microglia data of Kosoy et al. and Young et al.[26] gives a slope of 1.15 and 0.59, respectively. These results are difficult to interpret, especially given the caveats above.

These results are not unexpected given the statistical and technical challenges outlined above; in fact, these findings are not unique to our data. Ota et al.[71] generated eQTLs of four immune cell types and compared the estimated effect sizes to bulk immune data from Ishigaki et al.[72]. Our analysis recapitulates their finding, that effect size estimates are lower in cell-type-specific data.

Rigorous analysis of effect size estimates is challenging, both statistically and due to the different noise profiles of bulk and cell-type data. Mohammadi et al.[73] developed a method for estimation of a biologically interpretable allelic effect size. Further research on this challenge in the field could yield further insight into cell-type-specific gene regulation.

Simulation pipeline for evaluation of mmQTL performance. Genotype and gene expression data were simulated for comparison of the empirical performance of eQTL analysis using a linear model with five genotype PCs compared to a linear mixed model. Results from eQTL analysis of five simulated tissues were then aggregated using either Sidak correction or fixed- or random-effects meta-analysis.

Biologically realistic genotype data reflecting real human populations were simulated with a sampling-based simulation package, hapgen2 (ref. [74]), and haplotype information for European, African and Asian populations from the 1000 Genomes Project (https://mathgen.stats.ox.ac.uk/impute/data_download_1000G_2010_interim.html). We simulated 500 individuals for each population and merged these into a single transancestry dataset with a sample size of 1,500. We also simulated 1,500 individuals solely based on European haplotype information.

Based on these genotypes, we adapted the phenotypesimulator pipeline[75] to perform 800 simulations for each scenario, simulating one gene expression trait per simulation. For each gene, a single eQTL was simulated to affect expression abundance explaining 1% phenotypic variance and the contribution due to polygenic background was set to 30%. We applied phenotypesimulator's simulating strategy to account for shared environmental factors, measurement noise and polygenic background to create correlated phenotypes. In the simulation, we simulated phenotypes in five tissues and set the number of tissues affected by the causal genetic variant to 1, 2, 3, 4, 5. To demonstrate the robustness of mmQTL in controlling for population structure and batch effect, we set two different levels of phenotype correlation, a low level ($r = 0.12$) and a high level ($r = 0.45$). For power analysis, any simulated causal variants located in high LD ($r^2 \geq 0.8$), where a variant passing the multiple testing cutoff was considered to be detected.

We also performed a null simulation with no true causal variants, where all effect sizes were set to zero. Results from 50 simulations were aggregated, and we used genomic inflation factor[76] and quantile–quantile (Q–Q) plots to assess the false-positive rate.

Comparison of fine-mapping resolution between a European and a multi-ancestry population was performed using simulated pure 1,500 European individuals and multi-ancestry 1,500 individuals, with 500 individuals each of European, African and Asian ancestry. Gene expression phenotype was simulated in a single tissue, and a causal variant was randomly chosen to explain 2% phenotypic variance. For 1,500 European individuals we applied a standard linear regression model to detect eQTL and then fine-mapping was conducted to obtain a 95% credible set candidate for causal variants, while for 1,500 multi-ancestry individuals we used a mixed linear model to detect eQTL and then fine-mapping was used to find a 95% credible set. The size of the 95% credible set was used to compare fine-mapping resolution, a lower number indicating higher fine-mapping resolution.

Integration with ATAC–seq data. Variants in the 95% credible set were overlaid with open chromatin regions from four distinct populations of cells (GLU, GABA, OLIG and MgAs) identified by ATAC–seq[48]. To reduce the influence of the low fine-mapping resolution of conditional eQTL, if the size of the 95% credible set for a single gene contained more than ten variants only the ten variants with highest posterior inclusion probability (PIP) were included. Enrichment of variants within open chromatin regions was evaluated using Fisher's exact test implemented in QTLTools[77].

Evaluation of GWAS enrichment for variants in credible sets. We applied a strategy developed by Hormozdiari et al.[12]: for each eQTL we performed fine-mapping and computed the causal posterior probability (CPP) of each *cis*-SNP, retaining only variants in the fine-mapped 95% credible set. For each SNP in *cis*-regions we assigned an annotation value based on the maximum value of CPP across all molecular phenotypes; SNPs not belonging to any 95% credible set were assigned an annotation value of 0, which is referred to as MaxCPP in ref. [12]. Stratified LD score regression[52] was then used to partition trait heritability using the constructed functional annotations, and estimated enrichment was used to measure the importance of each eQTL category on human complex traits or diseases. To rule out the potential influences of correlation among eQTL categories we aggregated the baseline LD model, which includes a set of 75 functional annotations and functional annotations for eQTL categories, and ran stratified LD score regression simultaneously.

GWAS summary statistics were obtained for 22 human complex traits or diseases, including both brain and nonbrain traits (Supplementary Table 1).

eQTL detection in cell-type-specific datasets. Microglia from fresh human brain specimens (101 samples, including 27 non-European) were prepared using the Adult Brain Dissociation Kit (Miltenyi Biotech). Tissue homogenates were incubated in antibody (CD45: BD Pharmingen, Clone HI30 and CD11b: BD Pharmingen, Clone ICRF44) at 1:500 for 1 h in the dark at 4 °C with end-over-end rotation. Before FACS, DAPI (Thermoscientific) was added at 1:1,000 to facilitate identification of dead cells. Viable (DAPI-negative) $CD45^+/CD11b^+$ cells were isolated by FACS using a FACSAria flow cytometer (BD Biosciences) (Kosoy et al., in preparation). RNA was extracted from FACS sorted cells (Arcturus PicoPure RNA Isolation Kit, Life Technologies), and sequencing libraries were generated using the SMARTer Stranded RNA-seq kit (Clontech) according to the manufacturer's instructions. Variants with MAF >5% and Hardy–Weinberg equilibrium test $P > 10^{-6}$ were retained and analyzed using a linear mixed model implemented in mmQTL. Gene expression was normalized using $\log_2$ counts per million (CPM), and eQTL analysis was performed on residuals after regression out of 15 PCs of gene expression. For each gene, a Benjamini–Hochberg FDR correction was applied across all variants tested in the *cis*-regulatory region to obtain the minimum q-value. Then, minimum q-values across all genes were again adjusted by the Benjamini–Hochberg FDR method to compute genome-wide FDR. Limited by the small sample size, we chose a less conservative FDR cutoff of 10%.

Statistical fine-mapping. For each detected eQTL we conducted fine-mapping analysis applying the CAVIAR method[32] implemented in mmQTL to find a 95% credible set for causal variants. Briefly, the meta-analysis P value based on a random-effects model in each round of conditional eQTL detection was first converted to z-score, which was then used as input for fine-mapping. CAVIAR will calculate the PIP of each variant to causal, and a set of variants prioritized by PIP score were outputted with summed PIP equal to 0.95.

Detection of colocalization between eQTL and GWAS signals. Joint statistical fine-mapping of eQTL and GWAS signals[9] was performed by multiplying the estimated PIP for a given variant from the eQTL analysis by the PIP for this variant from GWAS of traits compiled in CausalDB[45] to obtain CLPP. A gene is considered to share a candidate causal variant with a GWAS trait if at least one variant has CLPP > 0.01 (ref. [9]).

Trait classification. CausalDB[45] provided the Medical Subject Headings (MeSH) category for each GWAS trait. However, brain-related traits fall within multiple MeSH categories and there is no single criterion to identify such traits. We performed manual inspection of traits in CausalDB that could be considered neuropsychiatric, neurodegenerative or behavioral. and termed them 'brain-related'.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Brain eQTL meta-analysis resource: http://icahn.mssm.edu/brema

Code availability

mmQTL: https://github.com/jxzb1988/mmQTL and Zenodo[79] (https://doi.org/10.5281/zenodo.5560014).

References

68. Sul, J. H., Han, B., Ye, C., Choi, T. & Eskin, E. Effectively identifying eQTLs from multiple tissues by combining mixed model and meta-analytic approaches. *PLoS Genet.* **9**, e1003491 (2013).
69. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
70. Joo, J. W. J., Hormozdiari, F., Han, B. & Eskin, E. Multiple testing correction in linear mixed models. *Genome Biol.* **17**, 62 (2016).
71. Ota, M. et al. Dynamic landscape of immune cell-specific gene regulation in immune-mediated diseases. *Cell* **184**, 3006–3021 (2021).
72. Ishigaki, K. et al. Polygenic burdens on cell-specific pathways underlie the risk of rheumatoid arthritis. *Nat. Genet.* **49**, 1120–1125 (2017).
73. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017).
74. Su, Z., Marchini, J. & Donnelly, P. HAPGEN2: simulation of multiple disease SNPs. *Bioinformatics* **27**, 2304–2305 (2011).
75. Meyer, H. V. & Birney, E. PhenotypeSimulator: a comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. *Bioinformatics* **34**, 2951–2956 (2018).
76. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
77. Delaneau, O. et al. A complete tool set for molecular QTL discovery and analysis. *Nat. Commun.* **8**, 15452 (2017).
79. Zeng, B. jxzb1988/MMQTL: mmQTL v1.5.0. https://doi.org/10.5281/zenodo.5771105 (2021).
78. Glassberg, E. C., Gao, Z., Harpak, A., Lan, X. & Pritchard, J. K. Evidence for weak selective constraint on human gene expression. *Genetics* **211**, 757–772 (2019).

Acknowledgements

This project was supported by the National Institute of Mental Health (NIH grants nos. R01-MH109677, U01-MH116442, R01-MH125246 and R01-MH109897), the National Institute on Aging (NIH grants nos. R01-AG050986, R01-AG067025 and R01-AG065582) and the Veterans Affairs Merit (no. BX004189) to P.R. G.E.H. was supported in part by NARSAD Young Investigator Grant no. 26313 from the Brain & Behavior Research Foundation. J.B. was supported in part by NARSAD Young Investigator Grant no. 27209 from the Brain & Behavior Research Foundation. Research reported in this paper was supported by the Office of Research Infrastructure of the National Institutes of Health under award nos. S10OD018522 and S10OD026880. The content herein is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Author contributions

B.Z., G.E.H. and P.R. conceived and designed the study. B.Z. designed and implemented the statistical method. B.Z. and G.E.H. performed analyses. J.F.F. generated cell-type-specific expression and chromatin accessibility data. J.B. and R.K. preprocessed and analyzed cell-type-specific expression and chromatin accessibility data. J.F.F. and P.R. supervised data generation. G.E.H. and P.R. supervised data analyses. G.E.H., B.Z. and P.R. wrote the manuscript with the help of all authors.

Competing interests

The authors declare no competing interests.

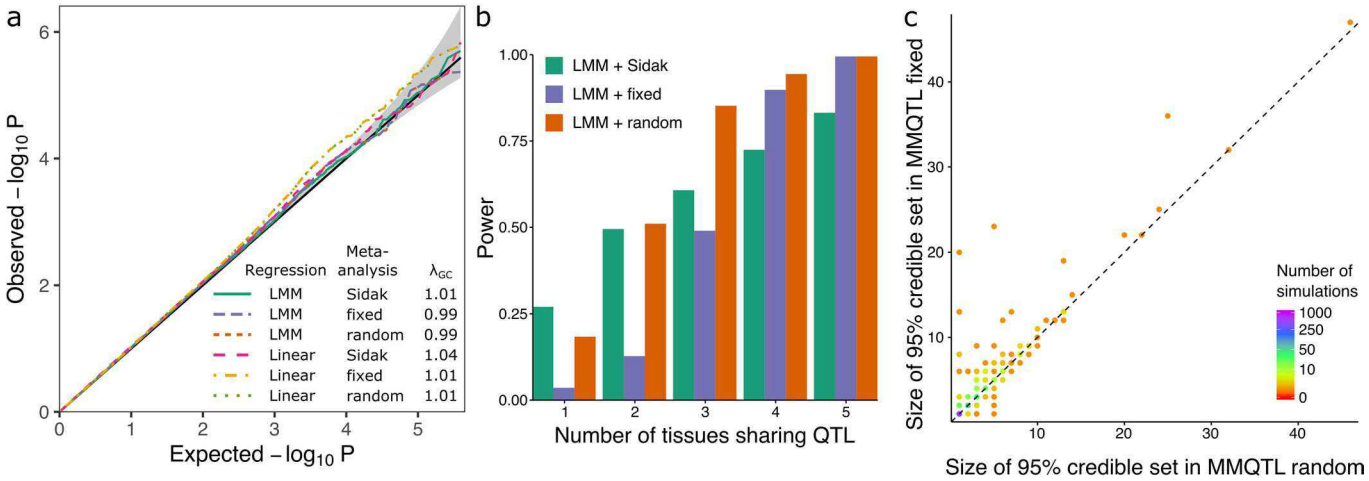
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41588-021-00987-9.

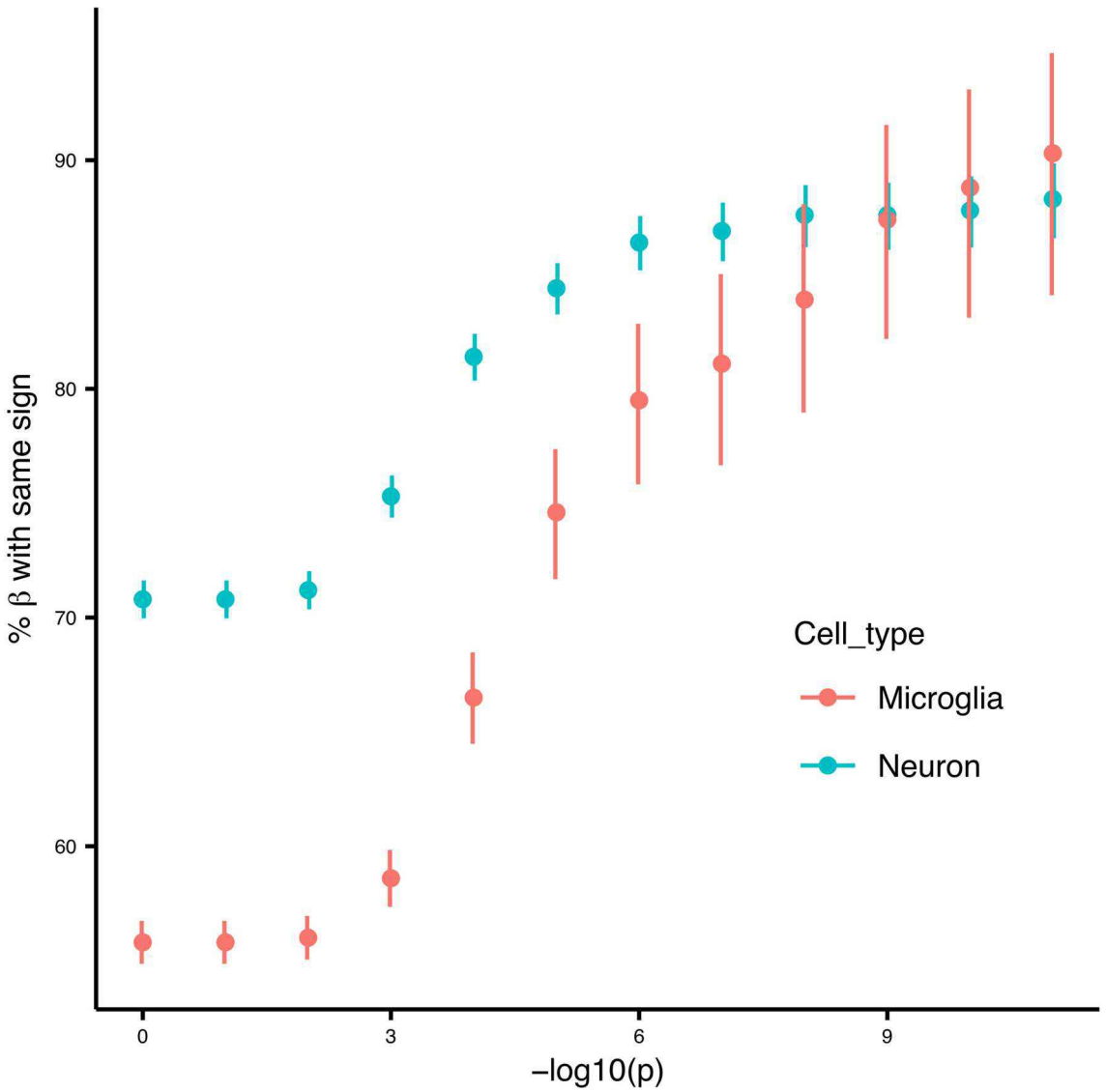
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-021-00987-9.

Correspondence and requests for materials should be addressed to Gabriel E. Hoffman or Panos Roussos.

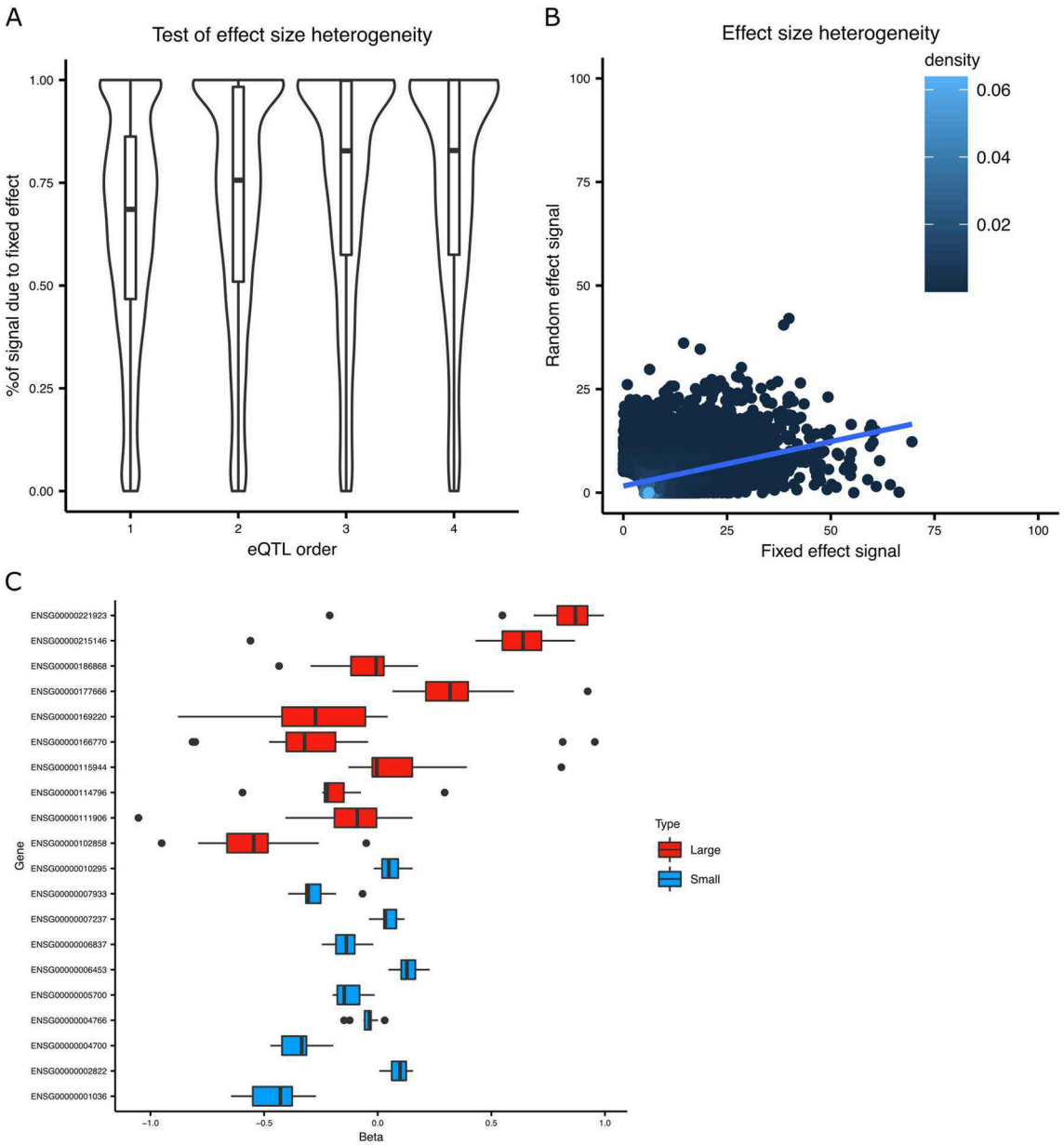
Peer review information *Nature Genetics* thanks Andrew Jaffe and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.



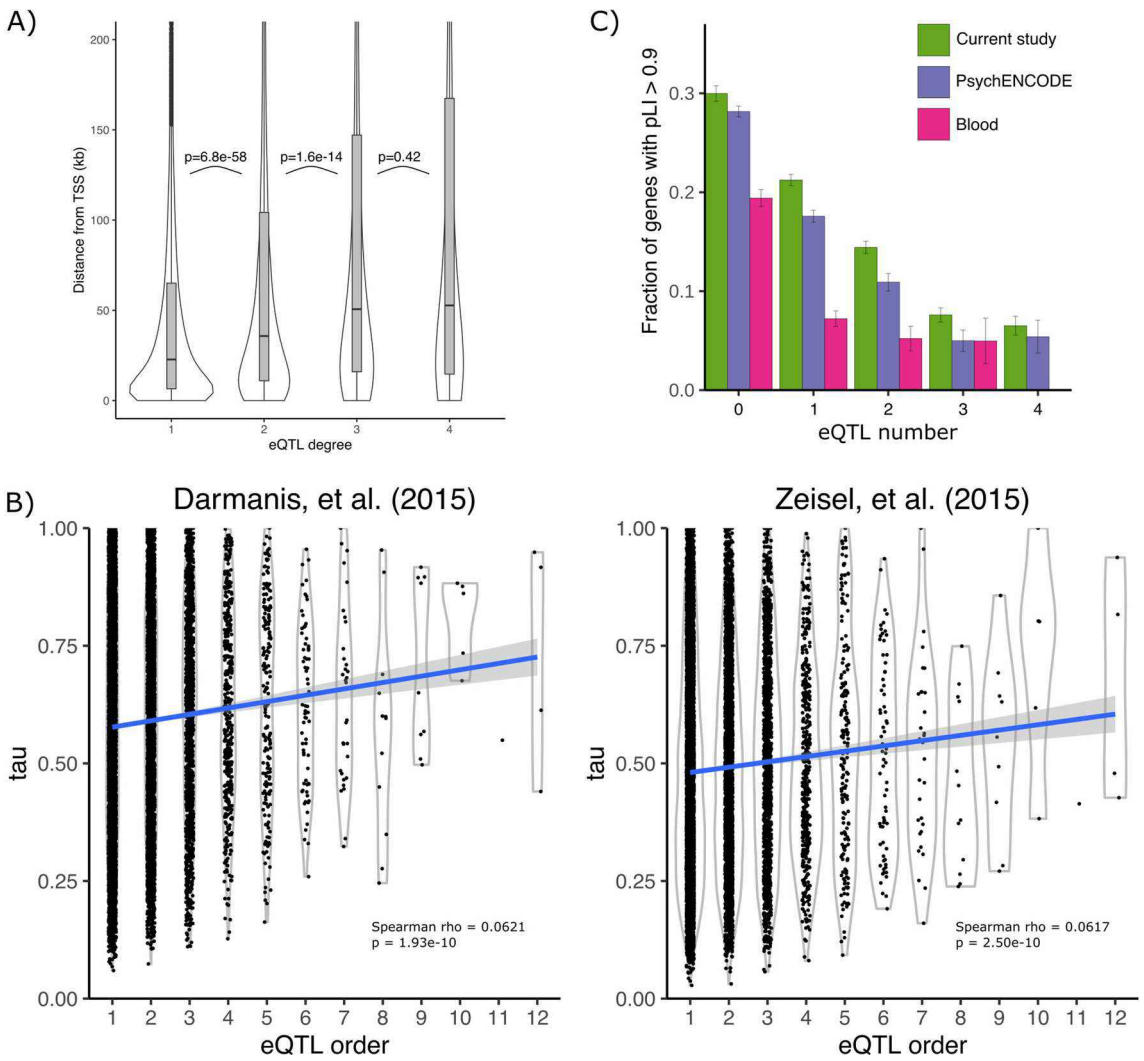
Extended Data Fig. 1 | Biologically motivated simulations demonstrate performance of mmQTL workflow: low correlation scenario. a) QQ plot of results from null simulation shows that the linear mixed model (LMM) with fixed or random effect meta-analysis accurately controls the false positive rate for, while linear regression with 5 genotype principal components did not. The Sidak method was very conservative in both cases. λ_{GC} indicates the genomic control inflation factor. Gray band indicates 95% confidence interval under the null. **b)** Power from LMM followed by 3 types of meta-analysis versus the number of tissues sharing an eQTL. **c)** Size of the 95% credible sets from fixed- (y-axis) and random- (x-axis) effects meta-analysis from simulations in Fig. 2c.



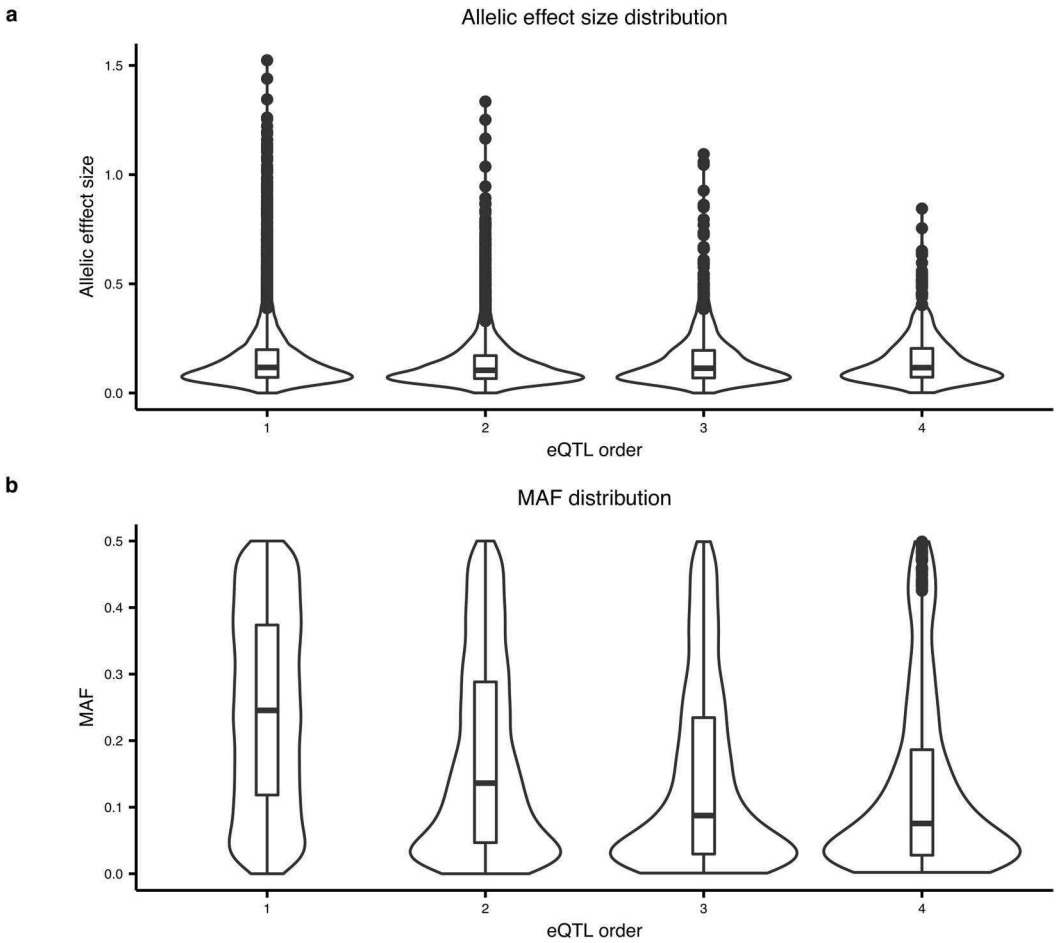
Extended Data Fig. 2 | Lead eQTL SNP sign concordance. For the lead eQTL SNP of each gene in the meta-analysis, the sign of the mean estimated effect size is compared to the estimated effect sign from neuron and microglia eQTL analyses. The concordance rate increases with the strictness of the p-value cutoff, so a smaller p-value indicates a higher concordance rate. Error bars indicate 95% confidence interval for a binomial proportion. Analysis included 11,709 variants for neuron, and 10,865 variants for microglia.



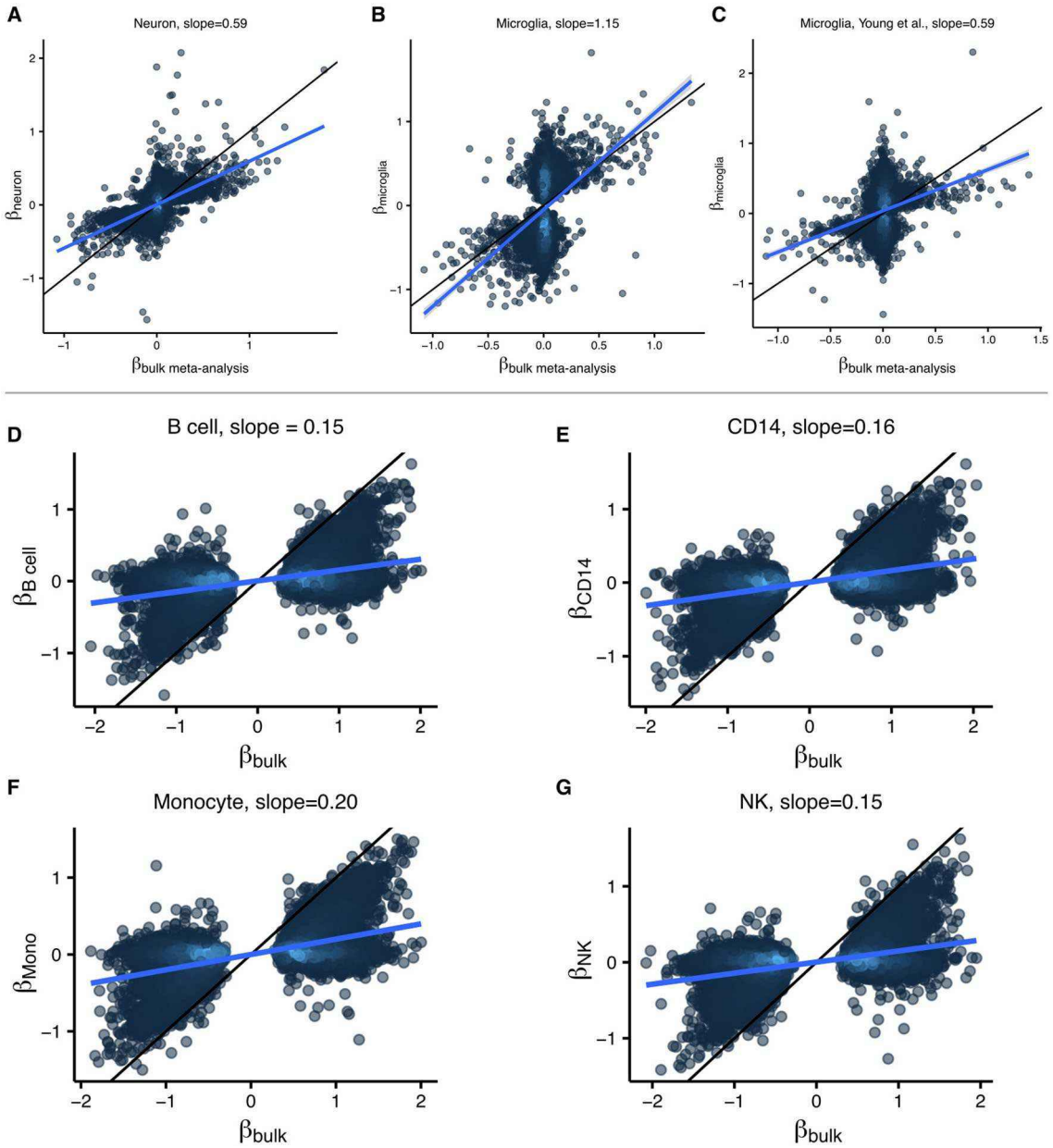
Extended Data Fig. 3 | Impact of effect size heterogeneity. The test statistic from the random effect meta-analysis used here (Han and Eskin, 2011) is the sum of statistics testing the mean (S_{mean}) and variance ($S_{variance}$) of the estimated effect sizes. **a)** The percent of total signal contributed by the fixed effect (that is S_{mean} / (S_{mean} + $S_{variance}$)) is shown for the lead eQTL SNP for multiple orders of conditional analysis. Box plot indicates median, interquartile range (IQR) and 1.5*IQR. **b)** The relationship between the test statistics is visualized by plotting sqrt $S_{variance}$ against sqrt S_{mean} from the lead eQTL SNP from the primary eQTL analysis. **c)** The estimated effect sizes from the lead eQTL SNP for genes with high and low levels of effect size heterogeneity is shown. Box plot indicates median, interquartile range (IQR) and 1.5*IQR.



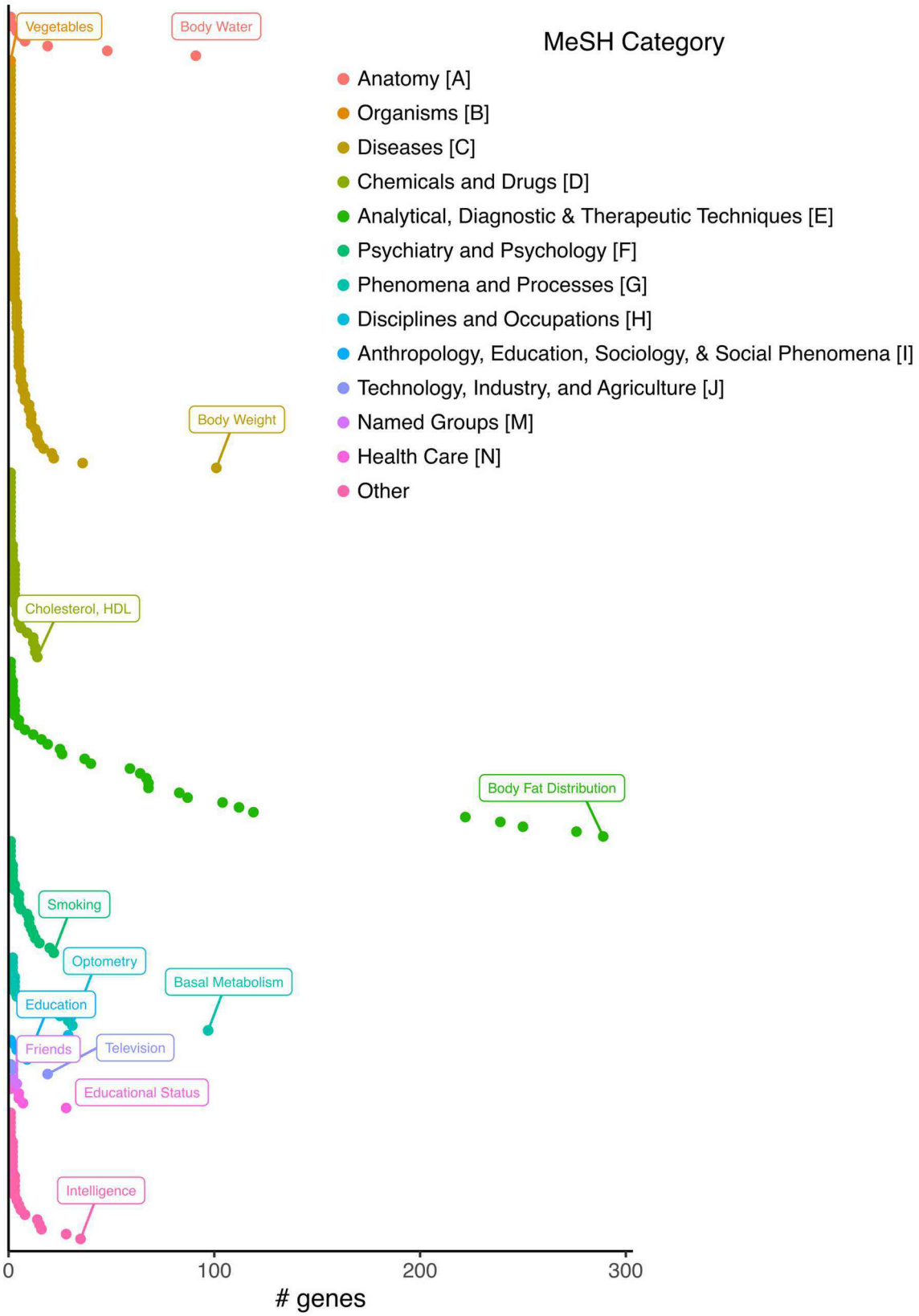
Extended Data Fig. 4 | Properties of conditional eQTLs. a) The distribution of the distance to the transcription start site is shown for the lead variant for eQTL analysis of increasing degree. P-values indicate significance of one-sided Mann–Whitney U test between adjacent groups. Box plot indicates median, interquartile range (IQR) and 1.5*IQR. **b)** Cell type specificity metric tau plotted against the number of independent eQTLs discovered for each gene. Gray band indicates 95% confidence interval. **c)** Bar plot shows that the fraction of genes with high evolutionary constraint (pLI > 0.9) decreases with eQTL degree for the current study, PsychENCODE[15], and whole blood[78]. Error bars indicate standard error based on asymptotic estimate of binomial proportion. Analysis included 10769 genes with eQTLs.



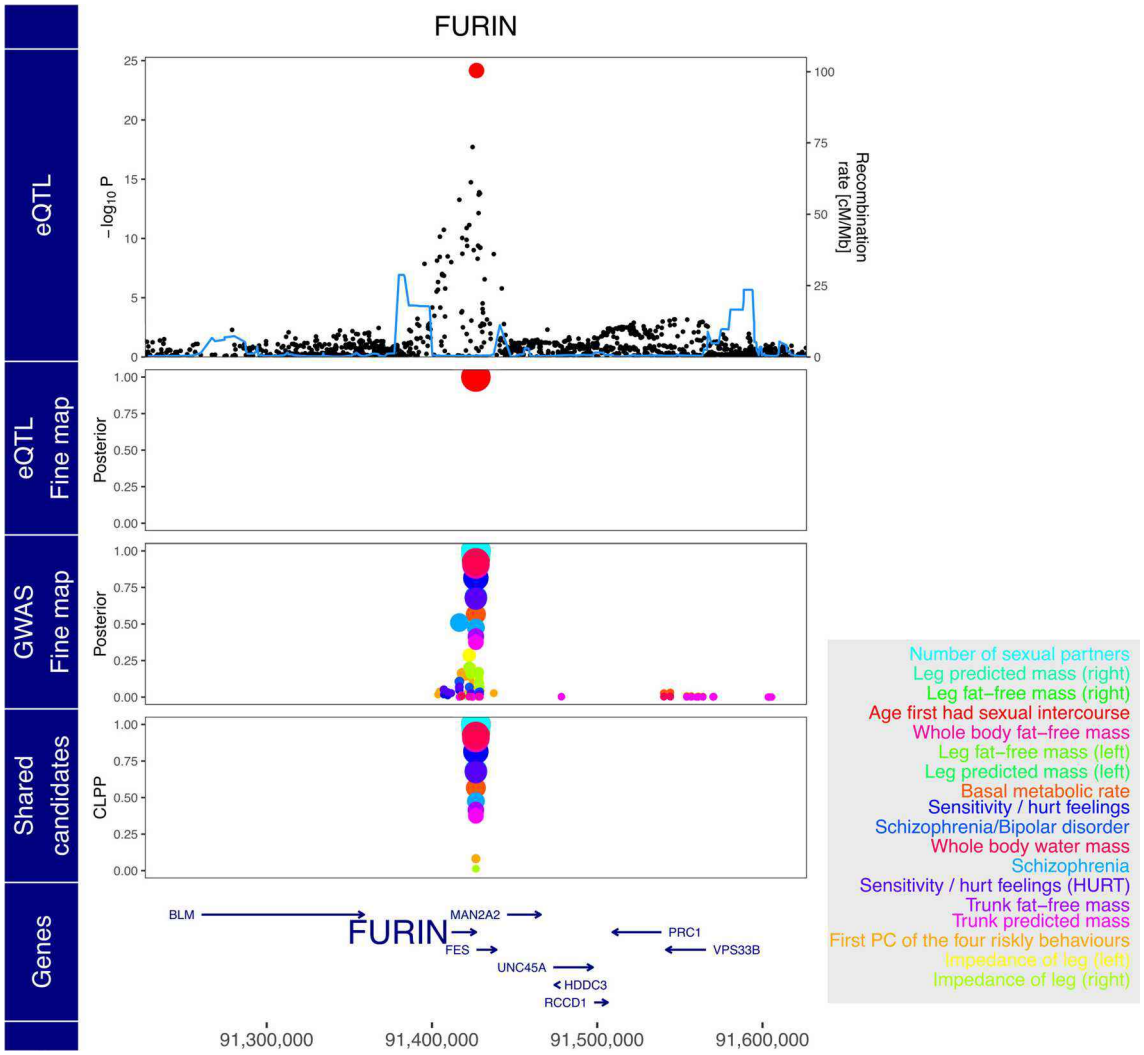
Extended Data Fig. 5 | Estimated effect size and minor allele frequencies from conditional eQTL analysis. The estimated effect size **(a)** and MAF **(b)** are shown for the lead eQTL SNP of significant genes for increasing order to conditional eQTL analysis. **a)** The distribution of estimated effect size is similar for all conditional analyses. **b)** The MAF shows a marked decrease with increasing order of conditional analysis. Box plot indicates median, interquartile range (IQR) and 1.5*IQR.



Extended Data Fig. 6 | Comparison of estimated effect size for bulk and cell-type specific data. (a-c) Estimated allelic effect size for eQTL lead in (**a**) neurons (Jaffe, et al. 2020), (**b**) microglia from Kosoy, et al. (in preparation) and (**c**) microglia from Young, et al. (2021) compared to effect size estimates from meta-analysis of bulk data from the current study. (**d–g**) Estimated allelic effect size for eQTL lead SNP in four immune cell types including (**d**) B cells, (**e**) CD14, (**f**) monocytes, (**g**) NK cells from Ota, et al. (2021) compared to estimates from bulk samples (Ishigaki, et al. 2017).



Extended Data Fig. 7 | Number of genes colocalizing for each MeSH category with CLPP > 0.01. The phenotype with the highest number of colocalized genes for each MeSH category is indicated.



Extended Data Fig. 8 | Expression of *FURIN* and risk for multiple complex traits share rs4702 as a candidate causal variant. Starting from the top, the plot shows -$\log_{10}$ p-values from eQTL analysis, poster probabilities from statistical fine-mapping of eQTL results, poster probabilities from statistical fine-mapping of GWAS results, and colocalization posterior probabilities (CLPP) for combining eQTL and GWAS fine-mapping. Traits are shown in the box on the right in decreasing order to CLPP value.

nature portfolio

Corresponding author(s): Gabriel Hoffman, Panos Roussos

Last updated by author(s): Nov 11, 2021

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data used for this analysis comes from the AD Knowledge Portal (adknowledgeportal.synapse.org) and the PsychENCODE Knowledge Portal (synapse.org/pec) and is hosted on the Sage Bionetworks' Synapse platform for access by qualified investigators. Data was generated from post-mortem tissue and has be de-identified according to the Synapse terms of use, and is available through the submission of an AD Knowledge Portal Data Use Certificate (https://adknowledgeportal.synapse.org/DataAccess/DataUseCertificates) or by application to the NIMH Repository and Genomics Resources (http://nimhgenetics.org), respectively. We also used publicly available data from GTEX (gtexportal.org).
Data analysis	eQTL discovery was performed with a linear mixed model using the mmQTL software developed in this work (https://github.com/jxzb1988/mmQTL and https://doi.org/10.5281/zenodo.5560014). All downstream analysis was performed with R v4.0.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Analysis results are available at http://icahn.mssm.edu/brema

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We included samples with genotype and brain gene expression data from GTEx, PsychENCODE, and ROSMAP datasets. The sample size was based on the availability of the existing data. No statistical method was used to determine what sample size to use
Data exclusions	Samples with < 8 million well imputed SNPs (info < 0.3) were excluded from PsychENCODE. All other samples with both genotype and brain gene expression were retained.
Replication	We evaluated the replication of our eQTL analyses in independent datasets of 1) bulk brain tissue (Wang, et al), 2) granule cell layer of the dentate gyrus enriched for excitatory neurons (Jaffe, et al, 2020) and 3) purified microglia (Kosoy, et al. in prep). The genome-wide replication rate is consistent with the rest of the field.
Randomization	We performed analysis on existing data, so randomization of the experimental design does not apply
Blinding	We performed analysis on existing data, so blinding of the experimental design does not apply

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Data generated by Bendl, et al (in revision, https://doi.org/10.1101/2021.01.11.426303) was show in Figure 7 involves use of NeuN antibody. However no data was generated as part of the current study. We refer the reader to that work for experimental details.
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*