Genetic control of RNA splicing and its distinctive role in complex trait variation

Jian Yang (jian.yang@westlake.edu.cn)
Westlake University https://orcid.org/0000-0003-2001-2474
Ting Qi
Westlake University
Yang Wu
The University of Queensland https://orcid.org/0000-0002-0128-7280
Futao Zhang
The University of Queensland
Jian Zeng
University of Queensland https://orcid.org/0000-0001-8801-5220

Article

Keywords: genome-wide association studies (GWAS), gene regulation

Posted Date: February 10th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-155233/v1

Version of Record: A version of this preprint was published at Nature Genetics on August 18th, 2022. See the published version at https://doi.org/10.1038/s41588-022-01154-4.

Genetic control of RNA splicing and its distinctive role in complex trait variation

Ting Qi[1,2,3], Yang Wu[3], Futao Zhang[3,4,5], Shouye Liu[1,6], Jian Zeng[3], Jian Yang[1,2,3,*]

[1]School of Life Sciences, Westlake University, Hangzhou, Zhejiang 310024, China
[2]Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, Zhejiang 310024, China
[3]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia
[4]Neuroscience Research Australia, Sydney, New South Wales 2031, Australia
[5]Clinical Genetics and Genomics, NSW Health Pathology Randwick, Sydney, New South Wales 2031, Australia
[6]Institute for Advanced Research, Wenzhou Medical University, Wenzhou, Zhejiang 325027, China
[*]Correspondence: Jian Yang (jian.yang@westlake.edu.cn)

Abstract

Most genetic variants identified from genome-wide association studies (GWAS) in humans are noncoding, indicating their role in gene regulation. Prior studies have shown considerable links of GWAS signals to expression quantitative trait loci (eQTLs), but the links to other genetic regulatory mechanisms such as splicing QTLs (sQTLs) are underexplored. Here, we introduce a transcript-based sQTL method (named THISTLE) with improved power for sQTL detection. Applying THISTLE along with LeafCutter, an event-based sQTL method, to brain transcriptomic data ($n = 1{,}073$), we identified 7,491 genes with sQTLs with $P < 5 \times 10^{-8}$ (the largest brain cis-sQTL collection to date), ~68% of which were distinct from eQTLs. Integrating the sQTL data into GWAS for ten brain-related complex traits (including diseases), we identified 107 genes associated with the traits through the sQTLs, ~68% of which could not be discovered using eQTL data. Our study demonstrates the distinctive role of most sQTLs in genetic regulation of transcription and complex trait variation.

Introduction

Genome-wide association studies (GWAS) have led to the discovery of tens of thousands of genetic variants associated with human complex traits (including diseases)[1,2]. However, the majority of the trait-associated variants are of uncharacterised function, and the mechanisms through which the genetic variants exert their effects on the traits remain largely elusive. Considering that most of the GWAS signals are located in non-coding regions of the genome, one hypothesis is that the genetic variants affect the traits through genetic regulation of gene expression[3,4]. The effects of genetic variation on mRNA abundance (also known as expression quantitative trait loci or eQTLs) have been studied for more than a decade[5-7], and nearly all genes have been found with one or more genetic variants associated with their mRNA abundance[6-8]. These advances have propelled the development of methods[9-13] to integrate eQTL data with GWAS to prioritize genes responsible for the GWAS signals. However, only a moderate proportion of the GWAS signals have been attributed to cis-eQTLs[14-18], likely because of various reasons including limited power, spatiotemporal eQTL effects, the focus on genomic regions in cis, and mechanisms beyond genetic control of mRNA abundance.

Genetic control of pre-mRNA splicing (also called splicing quantitative trait locus or sQTL) is another fundamental mechanism of gene regulation but remains heavily underexplored compared to eQTL. Currently, there is no consensus in the literature regarding the relationship between eQTL and sQTL. For example, there are observations showing that sQTLs are largely independent of eQTLs[7,19,20] and hypothesized to be one of the major contributors to genetic risk of disease[20], whereas a recent study shows that sQTLs do not have statistically significant contribution to trait heritability conditional on eQTLs[21]. These results motivated us to investigate the role of sQTLs in complex traits using a larger data set with a rich collection of sQTLs and a broader spectrum of splicing events in a single tissue. We focused our study on the brain because of data availability (see below) and the considerable links of sQTLs to neurodegenerative diseases such as Alzheimer's disease[22] and Parkinson's disease[23] reported recently.

There are two major classes of methods for sQTL mapping[24], transcript-based methods[25-29] and event-based methods[30-33], each favouring different types of splicing events. The transcript-based methods typically involve the generation of a splicing phenotype such as isoform ratio (i.e., a ratio of mRNA abundance of an isoform to the overall mRNA abundance of the gene), and the test of associations between genetic variants and the splicing phenotype[27-29]. sQTLseekeR[29] is one of the most commonly used transcript-based methods. It can detect sQTLs arising from complex splicing events because it is sensitive to splicing events involving main isoform changes. However, sQTLseekeR often reports a modest number of sQTLs and suffers from computational burden. On

the other hand, event-based methods[30-33] such as LeafCutter[33] typically involve the quantification of splicing events by counting RNA-seq reads aligned to specific exons or splice junctions, the estimation of a splicing phenotype (e.g., exon-inclusion level), and the association tests for the splicing phenotype. More specifically, LeafCutter[33], which does not require read assembly or inference of isoforms, allows the identification and quantification of novel splicing events by focusing on intron excision. LeafCutter can detect sQTLs arising from subtle splicing patterns (e.g., exon skipping) because it is sensitive to local splicing events.

In this study, we aim to investigate the genetic control of RNA splicing by generating the largest collection of sQTLs to date in brain and to understand the role of sQTLs in complex trait variation. To harvest as many sQTLs as possible, we used both the transcript- and event-based methods for sQTL analysis. The limited number of sQTLs detected by sQTLseekeR motivated us to develop a more powerful and efficient transcript-based sQTL method (dubbed THISTLE) that requires only isoform-level eQTL (isoform-eQTL) summary data. We show by simulation that THISTLE has well-calibrated test-statistics and improves power over sQTLseekeR. We applied THISTLE together with LeafCutter to the PsychENCODE data[34,35] (the largest publicly available brain transcriptomic data, n = 1,073), integrated the sQTL data into GWAS for ten brain-related traits (including diseases) of large sample sizes (n ranging from 46,350 to 766,345), and compared the results with those for eQTLs identified using the same data.

Results

Overview of THISTLE

Details of the THISTLE method can be found in the Methods section. For ease of understanding, let us take a gene with two transcript isoforms as an example. If a genetic variant is associated with a splicing event (e.g., exon skipping), the variant is expected to be associated with the difference in mature mRNA abundance between the two transcript isoforms (**Fig. 1**). Hence, an sQTL test can be performed by assessing the association of the variant with the difference in mRNA abundance between the isoforms, which is equivalent to a test of the difference in variant effect on mRNA abundance (i.e., isoform-eQTL effect) between the isoforms. In other words, it is a test of heterogeneity in isoform-eQTL effect between isoforms. Without loss of generality, let m be the number of transcript isoforms for a gene and $\hat{\mathbf{b}} = \{\hat{b}_1, \dots \hat{b}_j, \dots, \hat{b}_m\}$, where $\hat{b}_j$ is the estimated isoform-eQTL effect for isoform j. We have $\hat{\mathbf{b}} \sim MVN(\mathbf{b}, \mathbf{S})$ with $\mathbf{S}$ being the variance-covariance matrix of $\hat{\mathbf{b}}$. The difference between b_j and b_k ($j \neq k$) can be estimated as $\hat{d}_{jk} = \hat{b}_j - \hat{b}_k$. Then, if we let $\hat{\mathbf{d}} = \{\hat{d}_{jk}\}_{j,k\in\{1,\dots,m\},j<k}$, we have $\hat{\mathbf{d}} \sim MVN\ (\mathbf{d}, \mathbf{V})$ with $\mathbf{V}$ being the variance-covariance matrix of $\hat{\mathbf{d}}$. The covariance between $\hat{d}_{gh}$ and $\hat{d}_{jk}$ is $cov(\hat{d}_{gh}, \hat{d}_{jk}) =$

$cov(\hat{b}_g, \hat{b}_j) - cov(\hat{b}_h, \hat{b}_j) - cov(\hat{b}_g, \hat{b}_k) + cov(\hat{b}_h, \hat{b}_k)$, and the covariance between $\hat{b}_g$ and $\hat{b}_j$ is $cov(\hat{b}_g, \hat{b}_j) = \theta_{gj} S_g S_j$, where S_g^2 and S_j^2 are the variance of $\hat{b}_g$ and $\hat{b}_j$, respectively, and θ_{gj} is the correlation between $\hat{b}_g$ and $\hat{b}_j$. We have shown previously[11,36] that all the parameters in $\mathbf{V}$ can be estimated from summary-level data if individual-level data are unavailable (**Methods**). Under the null hypothesis of no sQTL effect, $\mathbf{d} = \mathbf{0}$. We use an approximate multivariate approach, similar to the heterogeneity in dependent instruments (HEIDI) test[11], to test against the null hypothesis given $\mathbf{V}$. We name the method THISTLE, which stands for testing for heterogeneity between isoform-eQTL effects, and have implemented it in the OSCA[37] software.

Calibration of THISTLE

We first used simulations to calibrate THISTLE (**Supplementary Note**) and compare it with the competing method, sQTLseekeR. The simulation results showed that there was no inflation in false-positive rate (FPR) under the null of no sQTL effect for both THISTLE and sQTLseekeR, in either the absence or the presence of eQTL effect (**Supplementary Fig. 1**). With respect to statistical power, the area under the receiver operating characteristic curve (AUC) for THISTLE was larger than that for sQTLseekeR, and the difference in power between the two methods increased with the increase of -$\log_{10}$(p-value) threshold (**Supplementary Fig. 2**). The difference in power was more prominent in the analysis of the PsychENCODE data (see below for more details), with 699,445 sQTLs discovered by THISTLE for 8,241 genes vs. 178,961 sQTLs discovered by sQTLseekeR for 4,887 genes using a false discovery rate (FDR) threshold of 0.01 (**Supplementary Note**), predominantly owing to a substantial number of genes filtered out by sQTLseekeR as being non-informative. Nevertheless, for the genes tested in both methods, 70% of the sQTLs discovered by THISTLE with $P_{\text{sQTL}} < 5 \times 10^{-8}$ were replicated in sQTLseekeR with $P_{\text{sQTL}} < 0.05$ (**Supplementary Fig. 3**). Note that the reciprocal analysis (selecting sQTLs from sQTLseekeR for replication with THISTLE) was infeasible because the sQTLseekeR p-values were capped at 1×10^{-6} owing to the use of permutation to compute the p-values. In addition, the performance of THISTLE using individual-level SNP genotype and RNA-seq data is similar to that using summary-level isoform-eQTL data (**Supplementary Fig. 4**), demonstrating the flexibility of the method.

Identifying cis-sQTLs in brain

We applied THISTLE along with LeafCutter to a brain transcriptomic data set from the PsychENCODE Consortium. We processed the PsychENCODE data from six cohorts: UCLA-ASD, BrainGVEX, the Lieber Institute for Brain Development (LIBD), the CommonMind Consortium (CMC), the CMC's NIMH Human Brain Collection Core (HBCC), and the Bipseq of Bipolar cohorts (Bipseq). Generation and processing of the genotype and RNA-seq data have been detailed

elsewhere[34,35]. Further quality control (QC) in each cohort and the combined data set has been summarized in the Methods section (**Supplementary Figs. 5** & **6**). After QC, we included in the analysis ~3.8 million genotyped or imputed common variants (minor allele frequency, MAF > 0.01) on 1,073 unrelated individuals of European ancestry, including 348 subjects with schizophrenia (SCZ), 164 subjects with bipolar disorder (BIP), 17 subjects with autism spectrum disorder (ASD), and 544 controls. For THISTLE analysis, we excluded isoforms with low expression level (i.e., Transcripts Per Million (TPM) < 0.1 in more than 20% of individuals) or genes with only one isoform, retaining 77,511 isoforms of 13,366 genes (**Methods**). To remove confounding factors, VariancePartition[38] was applied to decompose the variation in isoform abundance into components attributable to multiple known biological and technical factors such as study, bank, RNA quality (RIN), and age at death (**Supplementary Fig. 7**). To account for other unknown biological and technical factors such as hidden batch effects, Probabilistic Estimation of Expression Residuals (PEER) method[39] was applied to obtain a set of latent covariates. The adjusted isoform abundances for each gene were used for an isoform-eQTL analysis in OSCA[37] with the first 5 genetic principal components (PCs) and 50 PEER factors fitted as covariates, followed by an sQTL analysis using THISTLE (**Supplementary Fig. 8**). Recognising the likely insufficient power to detect trans-sQTLs and for ease of computing, we limited the sQTL test to SNPs within 2 Mb of each gene on either side. In total, we identified 610,039 significant cis-sQTLs for 4,881 genes with $P_{\mathrm{sQTL}} < 5 \times 10^{-8}$, comprising 433,220 unique SNPs (**Supplementary Fig. 9a**). The use of a p-value threshold of 5×10^{-8} is because it is the genome-wide significance level used in GWAS and also the default threshold used in SMR to select instrument SNPs[11] (see below). Note that if we use a less stringent p-value threshold (e.g., FDR < 0.01), there were 1,114,064 significant cis-sQTLs for 8,241 genes, comprising 699,445 unique SNPs. Of the 4,881 genes with at least one sQTL with $P_{\mathrm{sQTL}} < 5 \times 10^{-8}$, 87 (1.8%) were lowly expressed in brain with median Fragments Per Kilobase of transcript per Million mapped reads (FPKM) ranging from 0.2 to 1 (**Supplementary Fig. 10a**).

We also used LeafCutter, the state-of-the-art event-based sQTL method, to identify sQTLs for splicing events that may not be well captured by a transcript-based method like THISTLE. Using LeafCutter, we quantified excision ratios of 202,344 introns in 46,868 intronic excision clusters (**Methods**) with 46,199 clusters uniquely mapped to 16,518 genes and the remaining clusters either mapped to more than one gene or not located in the gene body. In the following analyses, we focused only on the 46,199 uniquely mapped clusters. To define the set of covariates for adjustment, VariancePartition[38] was applied to decompose the variance in intron excision ratio into components attributable to biological and technical factors (**Methods**). The results showed that a substantial proportion of variance in gene expression level could be explained by multiple

biological and technical factors such as study, bank, library batch, and RIN (**Supplementary Fig. 7b**). In contrast, variance in intron excision ratio explained by these factors were, on average, minor (**Supplementary Fig. 7a**), likely because these factors affect both the numerator and denominator of the intron excision ratio so that the effects are mostly cancelled out each other. Each of the intron excision ratio was then adjusted for the selected covariates (i.e., bank, study, RIN, age of death, PMI, library batch, platform, sex, and diagnosis) and used as a quantitative trait for association analysis with SNPs within 2 Mb of each intron using FastQTL[40] for sQTL test, with the first 5 genetic PCs and 15 PEER factors (estimated from the adjusted intron excision ratio) fitted as covariates (**Supplementary Fig. 8**). In total, we identified 1,860,296 cis-sQTLs for 7,451 LeafCutter intron clusters in 5,629 genes with $P_{\mathrm{sQTL}} < 5 \times 10^{-8}$, comprising 494,283 unique SNPs (**Supplementary Fig. 9b**). Of note, if we use an FDR threshold of 0.01, there were 3,479,048 significant cis-sQTLs for 16,518 intron clusters in 10,186 genes, comprising 787,277 unique SNPs. Of the 5,629 genes with at least one sQTL with $P_{\mathrm{sQTL}} < 5 \times 10^{-8}$, 786 genes (14.0%) were lowly expressed in brain (**Supplementary Fig. 10c**).

We define genes with at least one significant sQTL with $P_{\mathrm{sQTL}} < 5 \times 10^{-8}$ as sGenes. Combining the sQTL results from THISTLE and LeafCutter, there were a total of 632,481 sQTLs associated with 7,491 sGenes, of which 1,862 (25%) and 2,610 (35%) sGenes were detected only by THISTLE and LeafCutter, respectively, with 3,019 sGenes (40%) detected by both methods (**Supplementary Fig. 11a**). Although the proportion of sGenes over the tested genes for THISTLE ($4{,}881/13{,}366 = 36.5\%$) was slightly higher than that for LeafCutter ($5{,}629/16{,}518 = 34.1\%$), LeafCutter identified more sGenes (5,629 vs. 4,881) than THISTLE (**Supplementary Fig. 11a**), partly because THISTLE focuses only on genes with more than one annotated isoform while LeafCutter does not have such limitation. For example, 602 LeafCutter sGenes were filtered out in the THISTLE analysis because they have only one annotated isoform (**Supplementary Fig. 10d**). In addition, we have developed an online R shiny application to query the sQTLs with FDR < 0.01 (https://yanglab.westlake.edu.cn/qi-sqtl/).

Of the 3,019 sGenes detected by both methods, there were 1,954 sGenes for which the THISTLE sQTL signal was shared with the LeafCutter sQTL signal as indicated by a COLOC[9] PP4 value of > 0.8, in line with the observation that the THISTLE top sQTL SNP was in high linkage disequilibrium (LD), $r^2 > 0.8$, with the LeafCutter top sQTL SNP for 1,723 sGenes (**Supplementary Fig. 11b**). Note that PP4 is the posterior probability for the 4$^{\mathrm{th}}$ hypothesis of COLOC that a trait is associated with another trait through a shared genetic variant[9]. These results mean that apart from the 1,862 THISTLE-specific sGenes and 2,610 LeafCutter-specific sGenes, there were additionally 1,065 sGenes for which the sQTL signals detected by the two methods were distinct,

demonstrating the benefit of using a combination of transcript- and event-based methods for sQTL detection.

Quantifying the relationship between sQTLs and eQTLs

To assess the relationship between sQTLs and eQTLs, we performed an eQTL analysis using the same data as used in the sQTL analysis above and identified 663,326 cis-eQTLs with $P_{\mathrm{eQTL}} < 5 \times 10^{-8}$ for 9,524 genes. Similar as above, we define eGenes as genes with at least one significant eQTL with $P_{\mathrm{eQTL}} < 5 \times 10^{-8}$. We found that 65% (4,893/7,491) of the sGenes were also eGenes (**Fig. 2a**), a proportion higher than that (50%) reported in a recent study using transcriptomic data from fetal brain[19]. We hypothesize that the difference is due to the larger sample size (n) used in our study than that used in the fetal brain study (n = 233). To test this hypothesis, we used a down-sampling strategy to assess the sGene-eGene overlap in several subsets of data with n ranging from 100 to 1,073. The result showed that the power of either sGene or eGene discovery was proportional to n and that the difference in discovery power between sGene and eGene increased with increasing n (**Fig. 2c**), in line with the observation from prior work[7]. We also found that the sGene-eGene overlap was positively correlated with n (**Fig. 2e**), which is expected if most genes have both eQTL and sQTL. Of the 4,893 overlapping genes, there were 2,418 genes (32% of all sGenes) for which the sQTL signal was shared with the eQTL signal as indicated by a COLOC PP4 value of > 0.8, in line with the observation that LD r^2 between the top sQTL SNP and the top eQTL SNP was > 0.8 for only 2,094 genes (**Fig. 2b**). In summary, our results suggest that although a large proportion of sGenes are expected to be eGenes with large n and this proportion will increase with increasing n, most sQTLs (an estimate of ~68% in this study) are distinct from eQTLs.

sQTLs are enriched for splicing sites

Having shown that sQTLs were largely distinct from eQTLs (**Fig. 2**), we then asked whether the sQTLs and eQTLs show different patterns of functional enrichment. To do this, we annotated the 7,187 top sQTL SNPs and 9,106 top eQTL SNPs (only one top sQTL or eQTL per gene) using SnpEff[41] alongside the functional annotation data from the brain samples of the Roadmap Epigenomics Mapping Consortium (REMC)[42] (**Methods**). The fold enrichment was computed as the proportion of sQTLs or eQTLs in a functional category divided by the mean of a null distribution generated by resampling "control" cis-SNPs with MAF matched. It is reassuring that sQTLs were more enriched in "splice regions" (i.e., splice acceptor, splice donor, and splice region) than eQTLs (on average, 26.7-fold of enrichment for sQTLs vs. 9.9-fold for eQTLs; **Fig. 3a** & **3b**). eQTLs were more enriched in transcription start site (TSS) than sQTLs (11.5-fold, s.e. = 0.25, for eQTLs vs. 7.8-fold, s.e. = 0.28, for sQTLs; **Fig. 3c** & **3d**), consistent with our observation that eQTLs

were much closer to TSS than sQTLs (**Fig. 3e**). The results remained largely unchanged if we further matched the control SNPs with the top cis-sQTL (or cis-eQTL) SNPs by distance to TSS (**Supplementary Fig. 12**). It is of note that the enrichment of eQTLs in "splice regions" and the enrichment of sQTLs in TSS are not unexpected because of the shared sQTL and eQTL signals for about one third of the sGenes.

Enrichment of the sQTLs for heritability of complex traits

As described above, an sQTL is a genetic variant that affects transcript abundance because of its effect on splicing event. Hence, if an sGene is functionally related to a trait, the sQTL of the gene would exert an effect on the trait. We tested whether the brain cis-sQTLs are enriched for genetic variants associated with complex traits and disorders related to the brain. We acquired GWAS summary statistics for ten brain-related traits, i.e., fluid intelligence score (IQ), educational attainment (EA), smoking initiation (SmkInit), schizophrenia (SCZ), Alzheimer's disease (AD), Parkinson's disease (PD), autism spectrum disorder (ASD), insomnia, bipolar disorder (BIP), and major depression (MD) from published studies[43-52] (**Methods** & **Supplementary Table 1**). Both the sQTLs and eQTLs showed more inflated GWAS test-statistics compared to the other SNPs for all the ten traits, and the levels of inflation were indistinguishable between the sQTLs and the eQTLs (**Supplementary Fig. 13**). Nevertheless, if a phenotype is polygenic, the GWAS test-statistics are expected to be inflated even in the absence of confounding factors, and the level of inflation depends on sample size, trait heritability, and LD between SNPs and causal variants[53,54]. To account for the LD, we performed a stratified LD score regression (S-LDSC)[55,56] analysis to quantify the enrichment of the top cis-sQTL SNPs for SNP-based heritability (h^2_{SNP}) in comparison with that of the top cis-eQTL SNPs when fitted together with 53 other functional categories in the "baselineLD model"[55] (**Methods**). The result showed that both the sQTLs and eQTLs were highly enriched for heritability of all the traits except AD, and the fold enrichment for sQTLs (median enrichment = 65) was similar to that of eQTLs (median enrichment = 55) on average across traits (**Fig. 4a**). It is of note that the top SNPs were ascertained in the cis-regions, and SNPs in or near genes are known to explain disproportionately more trait variation than those in intergenic regions[57]. To avoid such an ascertainment bias, we re-computed the enrichment by comparing the per-SNP heritability of the top sQTL (or eQTL) SNPs to a set of MAF-matched "control SNPs" resampled from the SNPs used in the sQTL (or eQTL) analysis (i.e., the SNPs in cis). In this case, the overall levels of enrichment decreased, but the fold enrichment of sQTLs (median = 1.85) remained similar to that of eQTLs (median = 1.73) on average across traits (**Fig. 4b**). We further quantified the heritability enrichment at all the significant cis-sQTL (or cis-eQTL) SNPs, clumped cis-sQTL (or cis-eQTL) SNPs, or fine-mapped cis-sQTL (or cis-eQTL) SNPs with maximum causal posterior probabilities[21], and the conclusion that sQTLs have a similar level of heritability

enrichment to eQTLs still held (**Supplementary Figs. 14a, 15a** & **16a**). In addition, to account for overlap between the sQTLs and eQTLs, we estimated the contributions of the top cis-sQTL and the top cis-eQTL SNPs to trait heritability in a joint model and determined that the annotation effect size (τ) remained statistically significant for both the sQTLs and eQTLs for most traits (**Fig. 4c, Supplementary Figs. 14b, 15b** & **16b**).

We have shown that both the sQTLs and the eQTLs were enriched for trait heritability. However, it is unclear whether or not the enrichment is driven by mediated genetic effects on trait through pre-mRNA splicing (or mRNA abundance). To address this question, we applied a recently developed method, mediated expression score regression (MESC)[18], to quantify the proportion of heritability mediated (h^2_{med}/h^2_g) by the cis-sQTLs and the cis-eQTLs. We included only the sQTLs discovered by LeafCutter because the MESC analysis requires estimated sQTL effects which are not provided by THISTLE. Across the ten traits, we observed a median estimate of h^2_{med}/h^2_g of 0.13 and 0.11 from expression scores and splicing scores, respectively (**Fig. 4d**), where the estimate for expression scores is similar to that reported by a previous study[18]. All the results from the heritability enrichment and mediation analyses suggest that sQTLs play a unique role in explaining complex trait variation, at a level that is comparable to that for eQTLs (see below for more discussion).

Identifying complex trait genes using sQTL data

To leverage the cis-sQTLs to prioritize functional genes for the ten brain-related traits, we applied the summary-data-based Mendelian Randomization (SMR) approach[11] to test if an sGene is associated with a trait through sQTL and the COLOC approach[9] to assess whether the sGene-trait association is driven by the same set of causal variants. We used a combination of SMR and COLOC for this analysis because while SMR is more powerful than COLOC in detecting sGene-trait associations[11], the HEIDI test in SMR (for filtering out sGene-trait associations not driven by the same set of causal variants) is inapplicable to the THISTLE sQTL summary data owing to the unavailability of sQTL effects. We identified 212 sGenes in total for the ten traits at a genome-wide significance level ($P_{\text{SMR}} < 2.1 \times 10^{-6}$), correcting for ~23,400 SMR tests for each trait, 107 of which showed a COLOC PP4 value of > 0.8 (**Fig. 4e** & **Supplementary Table 2**), consistent with a plausible mechanism that genetic variants affect a trait through genetic control of splicing. *MPHOSPH9* is a notable example where *MPHOSPH9* was identified to be associated with SCZ using sQTL summary data from either THISTLE or LeafCutter (**Supplementary Fig. 17**), and the top cis-sQTL SNP, rs1727307, for the intronic excision event, 12:123707674:123710835:clu_71282_-, is located in the introns that are actively retained in mature mRNAs (intron retention). We also included the eQTL data in the SMR analysis for the ten traits and identified 317 eGenes ($P_{\text{SMR}} <$

5.3×10^{-6}), 149 of which reached PP4 > 0.8 in COLOC (**Fig. 4e** & **Supplementary Table 2**). Between the two sets of trait-associated genes discovered through sQTLs and eQTLs respectively, 34 genes were in common, meaning that we identified 73 more genes through sQTLs on top of the 149 genes identified through eQTLs, a ~49% increase (**Fig. 4e**). One of the examples is *DGKZ* (**Fig. 5**), the functional relevance of which to SCZ has been implicated in prior work using gene expression data from blood[58] or multiple tissues[59]. In this study, *DGKZ* was identified to be associated with SCZ using the sQTL data, implying a possible mechanism that the SNP effects on SCZ are mediated by genetic regulation of RNA splicing of *DGKZ*. Notably, no genome-wide significant eQTL was identified to be associated with the overall expression level of *DGKZ* in PsychENCODE (**Fig. 5b**), meaning that *DGKZ* would have been missed if we analyzed the eQTL data only. In summary, ~68% (73/107) of the trait-associated genes identified through sQTLs were not detected using eQTL data, which again suggests the distinctive role of sQTL in complex traits.

For each of the 34 trait-associated genes that could be identified by using either the sQTL or the eQTL data, we further performed a COLOC analysis to test whether the sQTL and eQTL signals are driven by the same or distinct causal variants (i.e., sharing vs. linkage). We found evidence for sharing (PP4 > 0.80) for 25 genes (23% of all trait-associated sGenes). One typical example is *SETD6*, for which the COLOC analysis (PP4 = 0.94) suggested that the sQTL, eQTL and SCZ GWAS signals are driven by the same causal variant(s) (**Supplementary Fig. 18**). However, this result also means that there were 9 genes, each of which was associated with a trait through two distinctive genetic regulatory mechanisms (see **Supplementary Fig. 19** for a typical example). Taken all together, our results show the distinctiveness of most sQTLs in mediating the polygenic effects for complex traits and demonstrate the substantial gain of power in gene discovery for complex traits by integrating sQTL data into GWAS.

Discussion

In this study, we have generated, to date, the largest catalog of genetic variants associated with a broad spectrum of alternative splicing events in human brain, significantly expanding our understanding of genetic control of RNA splicing. By comparing the sQTLs with the eQTLs identified in this study, we showed that ~68% of the sQTLs are distinct from the eQTLs, propelling research in the future to increase sample sizes for sQTL detection. By integrating sQTLs with GWAS for ten brain-related traits, ~68% of the trait-associated genes identified through sQTLs could not be discovered through eQTLs, demonstrating the unique role of sQTLs in dissecting the genetic architecture underpinning complex trait variation. Moreover, the trait-associated genes identified through sQTLs in this study provide important leads for further

mechanistic work to elucidate their functions in the development of the brain-related traits and disorders.

We applied both THISTLE and LeafCutter to the PsychENCODE data for sQTL discovery. These two methods complement to each other in several ways. First, deriving transcript abundance from short-read RNA-seq data in a transcript-based method like THISTLE is more challenging than directly quantifying intron excision ratios from split reads in an event-based method like LeafCutter, because each sequencing read only represents a small part of the transcript isoform, and alternative transcripts often have substantial overlaps. Second, local splicing events may not be able to capture the complete landscape of gene regulation but have more ability to detect sQTLs. Third, although LeafCutter identified a larger number of sGenes than THISTLE (5,629 vs. 4,881; **Supplementary Fig. 11a**), the number of genome-wide significant trait-associated genes discovered by SMR through the THISTLE sQTLs is larger than that through the LeafCutter sQTLs (103 vs. 74; **Supplementary Table 2** & **Supplementary Fig. 21**). This is partly due to that the Bonferroni correction in the SMR analysis using the LeafCutter sQTL summary data (m = ~18,500 tests for each trait) was more conservative than that using the THISTLE sQTL summary data (m = ~4,800). These results implicate that THISTLE and LeafCutter are complementary and should be used together for sQTL detection in practice. Of note that THISTLE requires only summary-level isoform-eQTL data, a feature that can facilitate sQTL analyses using data from multiple cohorts when sharing individual-level data is prohibitive.

With individual-level bulk RNA-seq and SNP genotype data on 1,073 prefrontal cortex samples from the PsychENCODE consortium, we identified 632,481 sQTLs associated with splicing levels of 7,491 genes and 663,326 eQTLs associated with expression levels of 9,524 genes at p-value < 5×10^{-8}. We have developed an online tool to query the sQTLs and eQTLs with FDR < 0.01. The full sQTL and eQTL summary data without p-value thresholding will also be publicly available upon publication. These data sets may be useful for future studies to understand the molecular mechanisms underpinning the genetic regulation of splicing in brain, to identify functional genes and variants for other brain-related phenotypes (e.g., cognitive and neuroimaging phenotypes and neuropsychiatric disorders), and to improve genomic risk prediction, etc. Our study also informs future analyses to quantify the relationship between sGenes and eGenes, between sQTLs and eQTLs, or more generally between any two types of molecular QTLs. We show that the low to moderate sGene-eGene overlap as observed in previous studies[7,20] are due to small *n* because the overlap is a function of *n* (**Fig. 2**). For example, only ~8% of sGenes are eGenes when $n = 100$, and this proportion increases to 64% as *n* increases to 1,073. If every gene has sQTL and eQTL (the latter has almost been proved to be true[6]), we predict that this proportion will eventually

approach to 1 with ever increasing *n*. Nevertheless, even if this is the case, the underlying sQTL and eQTL causal variants are largely distinct. In this study, we estimated that the sQTL and eQTL causal variants are shared for only ~32% of the sGenes (**Fig. 2**). This 32% may even be an overestimation because the limited power of COLOC in distinguishing close linkage (a scenario where the sQTL and eQTL causal variants are distinct but in high LD) from sharing (a scenario where the sQTL and eQTL share the same causal variant(s)), especially when *n* is not sufficiently large. In an extreme scenario where the sQTL and eQTL causal variants are in perfect LD, there is no power to distinguish linkage from sharing.

By integrating the sQTLs with GWAS data, we confirmed that sQTLs were enriched for genetic variants associated with complex traits, as in previous studies[19,20,22,23,60]. We note that a previous study by Hormozdiari et al.[21] shows that sQTLs do not have a significant contribution to disease heritability in a joint analysis of five BLUEPRINT molecular QTLs, namely eQTL, sQTL, H3K27ac histone QTL (hQTL), H3K4me1 hQTL, and DNA methylation QTL (mQTL)[21], which does not seem consistent with our results. The discrepancy is likely due to the small sample sizes of the molecular QTL data in Hormozdiari et al. (*n* = ~200 per cell type). In the present study, we have provided multiple lines of evidence that the role of sQTLs in complex trait variation is largely distinct from that of eQTLs. First, as discussed above, only up to 32% of the sQTL and eQTL signals are shared (**Fig. 2**). Second, sQTLs contributed significantly to trait heritability conditional on eQTLs (**Fig. 4c**). Third, ~68% of the trait-associated genes detected by integrating GWAS data with the sQTLs were not detected by using the eQTLs (**Fig. 4e**). Our results also imply that the contribution of sQTLs in mediating polygenic effects is comparable to that of eQTLs. For example, we observed that the inflation in GWAS test-statistics at the sQTLs was indistinguishable from that at the eQTLs (**Supplementary Fig. 13**), that heritability enrichment of the sQTLs was similar to that of the eQTLs (median fold enrichment of 1.85 for sQTLs vs. 1.73 for eQTLs across the ten traits), that the proportion of mediated heritability for the traits through cis-sQTLs was on a par with that through cis-eQTLs (median h^2_{med}/h^2_g of 0.11 for sQTLs vs. 0.13 for eQTLs across the ten traits), and that the number of trait-associated genes identified through the sQTLs was of the same magnitude as that through the eQTLs (107 vs. 149). Hence, large-scale sQTL studies in blood and other tissues or specific cell types are urgently needed to make more discovery of sQTLs to improve our understanding of genetic regulation of RNA splicing and to facilitate the translation of GWAS maps into mechanisms.

Our study has several limitations. First, the THISTLE method relies on accurate assembly and quantification of mRNA isoforms, which are limited by the short-read RNA-seq technology but could be improved in the future by long-read RNA-seq that allows direct identification of isoforms.

For example, the Pacific Biosciences (PacBio) isoform sequencing has successfully identified many novel transcripts and alternative splicing events in a range of tissues and cell types with well-characterized transcriptomes[61]. Second, the eQTLs and sQTLs identified in this study only explain a small fraction of the GWAS loci for the ten traits, and the majority of the loci remain to be unexplained, which is likely due to the insufficient sample size of the transcriptome data and mechanisms beyond genetic control of transcriptional regulation. Thus, a comprehensive set of genetic variants associated with other molecular phenotypes (e.g., chromatin accessibility, DNA methylation, histone modification, and protein abundance) from studies with very large sample sizes is essential to unveil the molecular basis of polygenic trait variation. Third, we have identified a large number of genetic variants associated with mRNA abundance and pre-mRNA splicing, but the causative functional variants are mostly unknown. Fine-mapping analyses and functional experiments are required to pinpoint the causal variants and elucidate the mechanisms behind genetic control of transcriptional regulation and the following cascades of events that lead to phenotypic changes. Last but not least, we focus this study only on brain transcriptomic data and brain-related traits. Although it is likely that our conclusions can be generalised to other tissues and complex traits given the strong correlation of eQTL effects across tissues[7,36,62] and the extremely similar pattern of genetic architecture of brain phenotypes to other complex traits, the generalisation of our conclusions needs to be confirmed by well-powered studies in the future. Despite these limitations, our study develops a powerful and flexible transcript-based sQTL method, generates the largest set of brain cis-sQTL summary data to date (with an online tool for data query), demonstrates an analysis paradigm to assess the relationship between two types of molecular QTLs, and provides multiple lines of evidence that most sQTLs are distinct from eQTLs.

Methods

Testing for heterogeneity between isoform-eQTL effects (THISTLE)

Let m be the number of transcript isoforms for a gene, $\mathbf{y}_j$ be a vector of mRNA abundances across n individuals for isoform j, and $\mathbf{x}$ be a vector of genotypes of a variant. If we define $\hat{\mathbf{b}} = \{\hat{b}_1, \dots \hat{b}_j, \dots, \hat{b}_m\}$ with $\hat{b}_j$ being the estimated isoform-eQTL effect for isoform j, we have $\hat{\mathbf{b}} \sim N(\mathbf{b}, \mathbf{S})$ with $\mathbf{S}$ being the variance-covariance matrix of $\hat{\mathbf{b}}$. The difference between $\hat{b}_j$ and $\hat{b}_k$ $(j \neq k)$ can be estimated as:

$$\hat{d}_{jk} = \hat{b}_j - \hat{b}_k$$

If we define $\hat{\mathbf{d}} = \{\hat{d}_{jk}\}_{j,k\in\{1,\dots,m\},j<k}$, we have $\hat{\mathbf{d}} \sim N(\mathbf{d}, \mathbf{V})$ with $\mathbf{V}$ being the variance-covariance matrix of $\hat{\mathbf{d}}$. The variance of $\hat{d}_{jk}$ over repeated experiments (i.e., a diagonal element of $\mathbf{V}$) can be written as

$$var(\hat{d}_{jk}) = var(\hat{b}_j - \hat{b}_k) = var(\hat{b}_j) + var(\hat{b}_k) - 2cov(\hat{b}_j, \hat{b}_k)$$

The covariance between $\hat{d}_{jk}$ and $\hat{d}_{gh}$ over repeated experiments (i.e., an off-diagonal element of **V**) can be written as

$$cov(\hat{d}_{jk}, \hat{d}_{gh}) = cov(\hat{b}_j - \hat{b}_k, \hat{b}_g - \hat{b}_h) = cov(\hat{b}_j, \hat{b}_g) - cov(\hat{b}_j, \hat{b}_h) - cov(\hat{b}_k, \hat{b}_g) + cov(\hat{b}_k, \hat{b}_h)$$

The covariance between $\hat{b}_j$ and $\hat{b}_k$ can be estimated as $cov(\hat{b}_j, \hat{b}_k) = \theta_{jk} S_j S_k$, where S_j^2 and S_k^2 are the variances of $\hat{b}_j$ and $\hat{b}_k$, respectively, and θ_{jk} is the correlation between $\hat{b}_j$ and $\hat{b}_k$. Since the isoform abundances are measured on the same set of individuals, $\hat{b}_j$ and $\hat{b}_k$ are likely to be correlated (i.e., $\theta_{jk} \neq 0$). We have shown in our previous studies[11,36] that $\theta_{jk} \approx r_p \rho$, where $\rho = \frac{n_s}{\sqrt{n_j n_k}}$ measures the sample overlap with n_j and n_k being the sample sizes of isoforms j and k, respectively, and n_s being the number of overlapping individuals between two isoforms, and r_p is the Pearson correlation of mRNA abundances between two isoforms in the overlapping sample. If the individual-level data are unavailable, θ_{jk} can be approximated by the Pearson correlation of the estimated isoform-eQTL effects between two isoforms across the "null" SNPs (e.g., $P_{\text{isoform-eQTL}} > 0.01$)[11,36]. Under the null hypothesis of no sQTL effect (i.e., $\mathbf{d} = \mathbf{0}$), there is no heterogeneity in isoform-eQTL effect between the isoforms. We have a vector of standard normal variables $\boldsymbol{z}_d = \{z_{d(jk)}\}_{j,k\in\{1,\ldots,m\},j<k}$ with $z_{d(jk)} = \hat{d}_{jk} / \sqrt{var(\hat{d}_{jk})}$ and $\boldsymbol{z}_d \sim N(\mathbf{0}, \mathbf{R})$, where **R** is the correlation matrix with $cor(z_{d(jk)}, z_{d(gh)}) = cov(\hat{d}_{jk}, \hat{d}_{gh}) / \sqrt{var(\hat{d}_{jk}) var(\hat{d}_{gh})}$. To test against the null hypothesis ($\mathbf{d} = \mathbf{0}$), we use an approximate multivariate approach, similar to the heterogeneity in dependent instruments (HEIDI) test[11], to construct a THISTLE test statistic $T_{THISTLE} = \boldsymbol{z}_d \mathbf{I} \boldsymbol{z}_d^T$, with **I** being an identify matrix. This distribution can be approximated by the Satterthwaite or Saddlepoint method[63,64]. We named the method THISTLE and has been implemented it in the OSCA[37] software package (https://yanglab.westlake.edu.cn/software/osca/#THISTLE).

Data used in this study

This study is approved by the Ethics Committee of Westlake University (approval number: 20200722YJ001). We used the genotype and RNA-seq data in dorsolateral prefrontal cortex (DLPFC) tissue from the PsychENCODE consortium. We processed the PsychENCODE data from six cohorts: UCLA-ASD, BrainGVEX, the Lieber Institute for Brain Development (LIBD), the CommonMind Consortium (CMC), the CMC's NIMH Human Brain Collection Core (HBCC), and Bipseq of Bipolar cohorts (Bipseq). Generation of the genotype and RNA-seq data and imputation of the genotype data have been detailed elsewhere[34,35]. RNA-seq data were filtered with a conventional QC process, e.g., retaining individuals with more than 10 million total reads and RNA integrity number (RIN) > 5.5. Genotyped and imputed SNP data were filtered with standard QC

criteria in each cohort separately using PLINK2[65], i.e., genotyping rate > 0.95, Hardy-Weinberg equilibrium test p-value $> 1 \times 10^{-6}$, MAF > 0.01, and imputation information score > 0.3. Individuals with duplicated sample ID across six cohorts, individuals with non-European ancestry as inferred from principal component analysis (**Supplementary Note; Supplementary Fig. 5**), or individuals with SNP-derived genetic relatedness > 0.05 were further removed. The resulting merged consensus data from six cohorts was filtered again according to the QC criteria described above. Finally, 1,073 unrelated individuals of European ancestry including 348 subjects with schizophrenia (SCZ), 164 subjects with bipolar disorder (BIP), 17 subjects with autism spectrum disorder (ASD), and 544 controls on 3,777,685 genotyped or imputed common SNPs were retained for further analysis (**Supplementary Fig. 6**).

We included in this study ten brain-related phenotypes, i.e., intelligence (IQ), educational attainment (EA), smoking initiation (SmkInit), schizophrenia (SCZ), Alzheimer's disease (AD), Parkinson's disease (PD), autism spectrum disorder (ASD), insomnia, bipolar disorder (BIP), and major depression (MD). GWAS summary statistics for IQ (n = 269,867)[43], EA (n = 766,345)[44], SmkInit (311,629 cases and 321,173 controls)[45], SCZ (40,675 cases and 64,643 controls)[46], AD (71,880 cases and 315,120 controls)[47], PD (33,674 cases and 449,056 controls)[52], ASD (18,381 cases and 27,969 controls)[48], insomnia (n = 386,533)[49], BIP (20,352 cases and 31,358 controls)[50], and MD (170,756 cases and 329,443 controls)[66] were from the latest published studies respectively (**Supplementary Table 1**).

Identification of cis-sQTLs using THISTLE and LeafCutter

To identify sQTLs using THISTLE, we first quantified the isoform abundance from short-read RNA-seq data. Detailed procedures of isoform quantification are available elsewhere[34,35]. In brief, RNA-seq reads of each of the 1,073 PsychENCODE samples were aligned to GRCh37 genome assembly by STAR v2.4.0f1[67], and isoform abundances, measured by transcripts per million (TPM), were quantified using the RSEM software package[68]. After filtering out isoforms with low expression (e.g., isoform-level TPM < 0.1 in more than 20% of the subjects) or genes with less than two annotated isoforms, a total of 77,511 isoforms of 13,366 genes were retained. VariancePartition[38] was then used to determine which known covariates (both biological and technical) should be used to adjust isoform expression using mixed-effects modeling. Probabilistic Estimation of Expression Residuals (PEER) method[39] was used to compute a set of latent covariates to capture hidden batch effects and other unknown technical and biological factors. We then performed isoform-eQTL analysis in OSCA[37] using the adjusted isoform expression data with the top 5 genetic principal components (PCs) and 50 PEER factors fitted as covariates, which was then followed by the THISTLE analysis to test for sQTLs. It is of note that

we used the same data for cis-eQTL analysis, which was essentially a meta-analysis of isoform-eQTL summary data across isoforms for each gene.

To identify sQTLs using LeafCutter[33], the alignments from STAR[67] were used as input into LeafCutter to identify intron clusters with at least one differentially excised intron by jointly modelling intron clusters using a Dirichlet-multinomial generalized linear model[33]. Default LeafCutter parameters were used to detect clusters, i.e., 50 split reads across all individuals were required to support each cluster, introns up to 500 kb were included, and introns used in less than 40% of individuals or with almost no variation were filtered out. The intron excision ratio (the ratio of intron defining reads to the total number of reads of an intron cluster) was standardized across individuals for each intron and quantile-normalized across introns. In total, 202,344 introns in 46,868 intronic excision clusters were identified with 46,199 clusters uniquely mapped to 16,518 genes and the remaining clusters either mapped to more than one gene or not located in the gene body. For covariate adjustment, we first identified known biological and technical covariates that were significantly associated with intron excision ratio by variancePartition[38] and then adjusted intron excision ratio by a mixed linear model, i.e., intron excision ratios ~ covariates + (1|Donor), where each intron was regressed on the covariates with donor (individual) fitted as random effects. PEER[39] was used to compute a set of latent covariates to capture unknown technical and biological factors. We then used FastQTL[40] to perform an association test for sQTL based on the following model: adjusted intron excision ratio ~ genotype + PC_1 + … + PC_5 + $PEER_1$ + … + $PEER_{15}$ (**Supplementary Fig. 8**), where PC_1 + … + PC_5 represent the top 5 genetic PCs and $PEER_1$ + … + $PEER_{15}$ represent 15 PEER factors.

Enrichment of the sQTLs and eQTLs for functional annotations

To test if the sQTLs and eQTLs are functionally enriched, we annotated the top cis-sQTL or cis-eQTL SNPs using the annotations from SnpEff[41] (e.g., splice region, intronic, and upstream) and 14 main functional annotations (e.g., promoter and enhancer) in 10 brain samples from the NIH Roadmap Epigenomics Mapping Consortium (REMC)[42]. More specifically, we annotated 9,106 unique top-associated cis-eQTL SNPs and 7,187 unique top-associated cis-sQTL SNPs in different functional annotations based on their physical positions and quantified the proportion of eQTLs or sQTLs in each functional category. To ameliorate ascertainment bias, we sampled at random the same number of cis-SNPs (i.e., SNPs included in the sQTL or eQTL analysis) as "control SNPs", with their MAF matched with the top sQTL (or eQTL) SNPs. This sampling procedure was repeated 1,000 times. We computed the fold enrichment in each functional annotation category as the ratio of the proportion of sQTL (or eQTL) SNPs in a functional category over the mean proportion of the control SNPs in the category across 1,000 replicates. The sampling variance of

the fold enrichment can be calculated approximately by the Delta method[69] (**Supplementary Note**).

Enrichment of the sQTLs and eQTLs for trait heritability

The stratified LD score regression (S-LDSC)[55,56] was used to quantify the enrichment of heritability attributable to the sQTLs and eQTLs (when fitted together with 53 other functional categories in the "baseline model") for the ten brain-related traits. Details of the "baseline model" that consists of 53 functional categories can be found elsewhere[55]. We created a binary annotation for sQTLs and eQTLs respectively. In brief, we assigned an annotation value of 1 for the most significant sQTL (or eQTL) SNP for each gene and a zero value for the remaining SNPs. This resulted in an sQTL annotation category with 6,547 SNPs and an eQTL annotation category with 8,192 SNPs with an overlap of 1,486 SNPs. LD scores of the SNPs were computed using SNP genotype data of the individuals of European ancestry from the 1000 Genomes Project (Phase 3)[70] with a window size of 1 cM. Heritability enrichment of a category was computed as the proportion of heritability explained by the category divided by the proportion of SNPs in the category. Considering that SNPs in or near genes explain disproportionately more trait variation than intergenic SNPs[57], we also computed the fold enrichment of heritability as the per-SNP heritability for the top cis-sQTL (or cis-eQTL) SNPs divided by a mean of a distribution generated by resampling MAF-matched cis-SNPs (i.e., the control SNPs mentioned above). Sampling variance of the fold enrichment of heritability can be calculated approximately by the Delta method[69] (**Supplementary Note**). Note that both the per-SNP heritability and annotation effect size (τ) reported by S-LDSC are used to quantify the relevance of a functional category to the trait variation[56], and the main difference between the two metrics lies in how the overlapping annotations are dealt with.

Author Contributions

JY conceived and supervised the study. TQ and JY designed the experiment and developed the method. TQ performed all the simulations and data analyses under the assistance or guidance from YW, JZ, and JY. TQ, FZ and SL developed the application tools. TQ and JY wrote the manuscript with the participation of all authors.

Acknowledgements

This research was partly supported by the Westlake Education Foundation, the Australian National Health and Medical Research Council (1107258 and 1113400) and the Australian Research Council (DP160101343 and FT180100186). We thank the Westlake University High-Performance Computing Center for assistance in computing. This study uses data from the

PsychENCODE consortium (Synapse accession: syn4921369). A full list of acknowledgments to this data set can be found in the Supplementary Note.

Data Availability

PsychENCODE data: https://www.synapse.org/#!Synapse:syn4921369/files/. Online tool for querying the sQTLs and eQTLs with FDR < 0.01: https://yanglab.westlake.edu.cn/qi-sqtl/. The full sQTL and eQTL summary data will be publicly available upon publication.

Code Availability

The computer code and documentation of THISTLE are available at https://yanglab.westlake.edu.cn/software/osca/#THISTLE.

URLs

OSCA: https://yanglab.westlake.edu.cn/software/osca.

LeafCutter: https://davidaknowles.github.io/leafcutter/index.html.

FastQTL: http://fastqtl.sourceforge.net/

S-LDSC: https://github.com/bulik/ldsc.

SMR: https://cnsgenomics.com/software/smr/#Overview.

COLOC: https://cran.r-project.org/web/packages/coloc/index.html.

MESC: https://github.com/douglasyao/mesc.

References

1 Visscher, P. M. *et al.* 10 years of GWAS discovery: biology, function, and translation. *The American Journal of Human Genetics* **101**, 5-22 (2017).

2 Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic acids research* **47**, D1005-D1012 (2018).

3 Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190-1195 (2012).

4 Ward, L. D. & Kellis, M. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology* **30**, 1095 (2012).

5 Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743-747 (2004).

6 Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv*, 447367 (2018).

7 GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318-1330 (2020).

8 Liu, B., Gloudemans, M. J., Rao, A. S., Ingelsson, E. & Montgomery, S. B. Abundant associations with gene expression complicate GWAS follow-up. *Nature genetics* **51**, 768-769 (2019).

9 Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS genetics* **10**, e1004383 (2014).

10 Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nature genetics* **48**, 245-252 (2016).

11 Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics* **48**, 481-487 (2016).

12 Hormozdiari, F. *et al.* Colocalization of GWAS and eQTL signals detects target genes. *The American Journal of Human Genetics* **99**, 1245-1260 (2016).

13 Barbeira, A. N. *et al.* Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. *Nature communications* **9**, 1-20 (2018).

14 Pavlides, J. M. W. *et al.* Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome medicine* **8**, 84 (2016).

15 Mancuso, N. *et al.* Integrating gene expression with summary association statistics to identify genes associated with 30 complex traits. *The American Journal of Human Genetics* **100**, 473-487 (2017).

16 Huckins, L. M. *et al.* Gene expression imputation across multiple brain regions provides insights into schizophrenia risk. *Nature genetics* **51**, 659-674 (2019).

17 Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nature communications* **10**, 1-12 (2019).

18 Yao, D. W., O'Connor, L. J., Price, A. L. & Gusev, A. Quantifying genetic effects on disease mediated by assayed gene expression levels. *Nature Genetics*, 1-8 (2020).

19 Walker, R. L. *et al.* Genetic Control of Expression and Splicing in Developing Human Brain Informs Disease Mechanisms. *Cell* **179**, 750-771. e722 (2019).

20 Li, Y. I. *et al.* RNA splicing is a primary link between genetic variation and disease. *Science* **352**, 600-604 (2016).

21 Hormozdiari, F. *et al.* Leveraging molecular quantitative trait loci to understand the genetic architecture of diseases and complex traits. *Nature genetics* **50**, 1041-1047 (2018).

22 Raj, T. *et al.* Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nature genetics* **50**, 1584 (2018).

23 Li, Y. I., Wong, G., Humphrey, J. & Raj, T. Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nature communications* **10**, 1-10 (2019).

24 Park, E., Pan, Z., Zhang, Z., Lin, L. & Xing, Y. The expanding landscape of alternative splicing variation in human populations. *The American Journal of Human Genetics* **102**, 11-26 (2018).

25 Katz, Y., Wang, E. T., Airoldi, E. M. & Burge, C. B. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature methods* **7**, 1009 (2010).

26 Trapnell, C. *et al.* Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature biotechnology* **31**, 46 (2013).

27 Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506-511 (2013).

28 Battle, A. *et al.* Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome research* **24**, 14-24 (2014).

29 Monlong, J., Calvo, M., Ferreira, P. G. & Guigó, R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. *Nature communications* **5**, 4698 (2014).

30 Shen, S. *et al.* rMATS: robust and flexible detection of differential alternative splicing from replicate RNA-Seq data. *Proceedings of the National Academy of Sciences* **111**, E5593-E5601 (2014).

31 Ongen, H. & Dermitzakis, E. T. Alternative splicing QTLs in European and African populations. *The American Journal of Human Genetics* **97**, 567-575 (2015).

32 Vaquero-Garcia, J. *et al.* A new view of transcriptome complexity and regulation through the lens of local splicing variations. *elife* **5**, e11752 (2016).

33 Li, Y. I. *et al.* Annotation-free quantification of RNA splicing using LeafCutter. *Nature genetics* **50**, 151 (2018).

34 Gandal, M. J. *et al.* Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. *Science* **362** (2018).

35 Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362** (2018).

36 Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nature communications* **9**, 2282 (2018).

37 Zhang, F. *et al.* OSCA: a tool for omic-data-based complex trait analysis. *Genome biology* **20**, 107 (2019).

38 Hoffman, G. E. & Schadt, E. E. variancePartition: interpreting drivers of variation in complex gene expression studies. *BMC bioinformatics* **17**, 483 (2016).

39 Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nature protocols* **7**, 500 (2012).

40 Ongen, H., Buil, A., Brown, A. A., Dermitzakis, E. T. & Delaneau, O. Fast and efficient QTL mapper for thousands of molecular phenotypes. *Bioinformatics* **32**, 1479-1485 (2015).

41 Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly* **6**, 80-92 (2012).

42 Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317 (2015).

43 Savage, J. E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nature genetics* **50**, 912 (2018).

44 Lee, J. J. *et al.* Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals. *Nature genetics* **50**, 1112 (2018).

45 Liu, M. *et al.* Association studies of up to 1.2 million individuals yield new insights into the genetic etiology of tobacco and alcohol use. *Nature genetics* **51**, 237 (2019).

46 Pardiñas, A. F. *et al.* Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection. *Nature genetics* **50**, 381 (2018).

47 Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. (2019).

48 Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nature genetics* **51**, 431 (2019).

49 Jansen, P. R. *et al.* Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nature genetics* **51**, 394 (2019).

50 Stahl, E. A. *et al.* Genome-wide association study identifies 30 loci associated with bipolar disorder. *Nature genetics* **51**, 793 (2019).

51 Howard, D. M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature neuroscience* **22**, 343 (2019).

52 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. *The Lancet Neurology* **18**, 1091-1102 (2019).

53 Yang, J. *et al.* Genomic inflation factors under polygenic inheritance. *European Journal of Human Genetics* **19**, 807-812, doi:10.1038/ejhg.2011.39 (2011).

54 Bulik-Sullivan, B. K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nature genetics* **47**, 291 (2015).

55 Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nature genetics* **47**, 1228 (2015).

56 Gazal, S. *et al.* Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. *Nature genetics* **49**, 1421 (2017).

57 Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nature genetics* **43**, 519 (2011).

58 Alinaghi, S. *et al.* Expression analysis and genotyping of DGKZ: a GWAS-derived risk gene for schizophrenia. *Molecular biology reports*, 1-7 (2019).

59 Pers, T. H. *et al.* Comprehensive analysis of schizophrenia-associated loci highlights ion channel pathways and biologically plausible candidate causal genes. *Human molecular genetics* **25**, 1247-1254 (2016).

60 Takata, A., Matsumoto, N. & Kato, T. Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. *Nature communications* **8**, 14519 (2017).

61 Wenger, A. M. *et al.* Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nature biotechnology*, 1-8 (2019).

62 Liu, X. *et al.* Functional architectures of local and distal regulation of gene expression in multiple human tissues. *The American Journal of Human Genetics* **100**, 605-616 (2017).

63 Kuonen, D. Miscellanea. Saddlepoint approximations for distributions of quadratic forms in normal variables. *Biometrika* **86**, 929-935 (1999).

64 Davies, R. B. Numerical inversion of a characteristic function. *Biometrika* **60**, 415-417 (1973).

65 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, s13742-13015-10047-13748 (2015).

66 Wray, N. R. *et al.* Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature genetics* **50**, 668 (2018).

67 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).

68 Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC bioinformatics* **12**, 323 (2011).

69 Lynch, M. & Walsh, B. *Genetics and analysis of quantitative traits*. Vol. 1 (Sinauer Sunderland, MA, 1998).

70 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (2012).

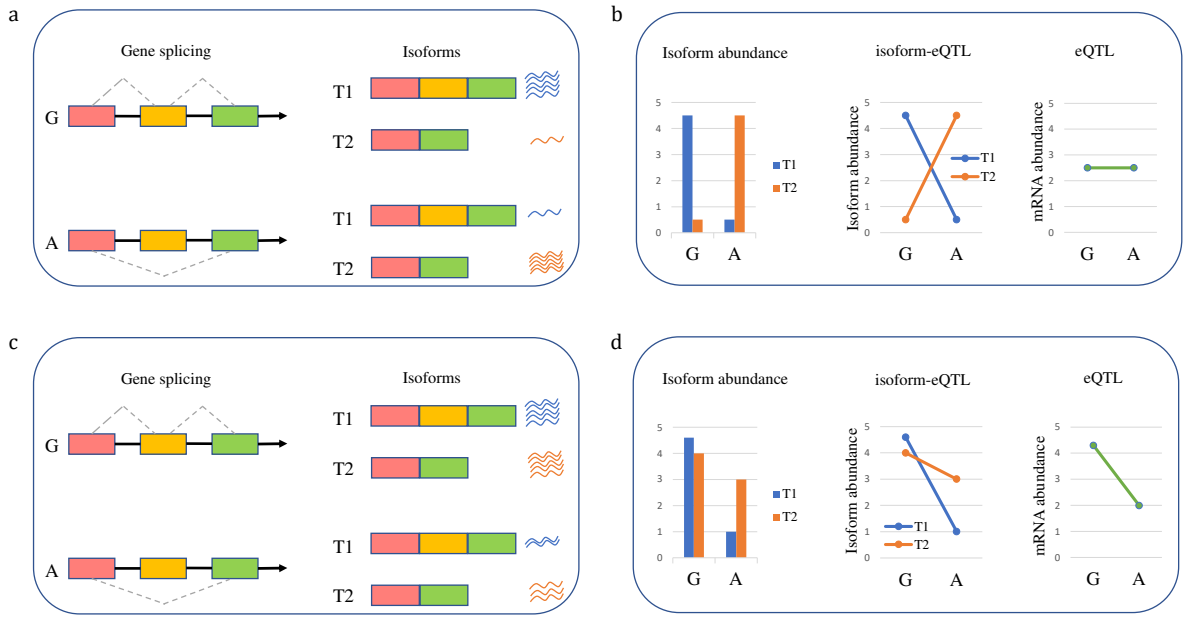

Figure 1. Schematic of THISTLE sQTL analysis. In this toy example, a genetic variant with two alleles, G and A, is associated with a splicing event (e.g., exon skipping) in a gene with two isoforms, T_1 and T_2. Panels **a**) and **b**): a schematic of THISTLE sQTL analysis in the absence of eQTL effect. In this scenario, individuals with G allele show higher mean transcription abundance for T_1 than T_2, and individuals with A allele show higher mean abundance for T_2 than T_1 (**a**), meaning that the genetic variant is associated with the difference in mRNA abundance between the isoforms. In other words, there is a difference in the isoform-eQTL effect between isoforms T_1 and T_2 (**b**). However, there is no difference in overall gene expression between individuals with different alleles, meaning that this genetic variant is not an eQTL. Panels **c**) and **d**): a schematic of THISTLE sQTL analysis in the presence of eQTL effect. In this scenario, individuals with G allele show similar transcription abundance between T_1 and T_2, and individuals with A allele show lower abundance for T_1 than T_2 (**c**). The isoform-eQTL effect for T_1 is different from that for T_2 albeit in the same direction (**d**). In this case, there is a difference in overall gene expression between alleles G and A, meaning that this genetic variant is also an eQTL.

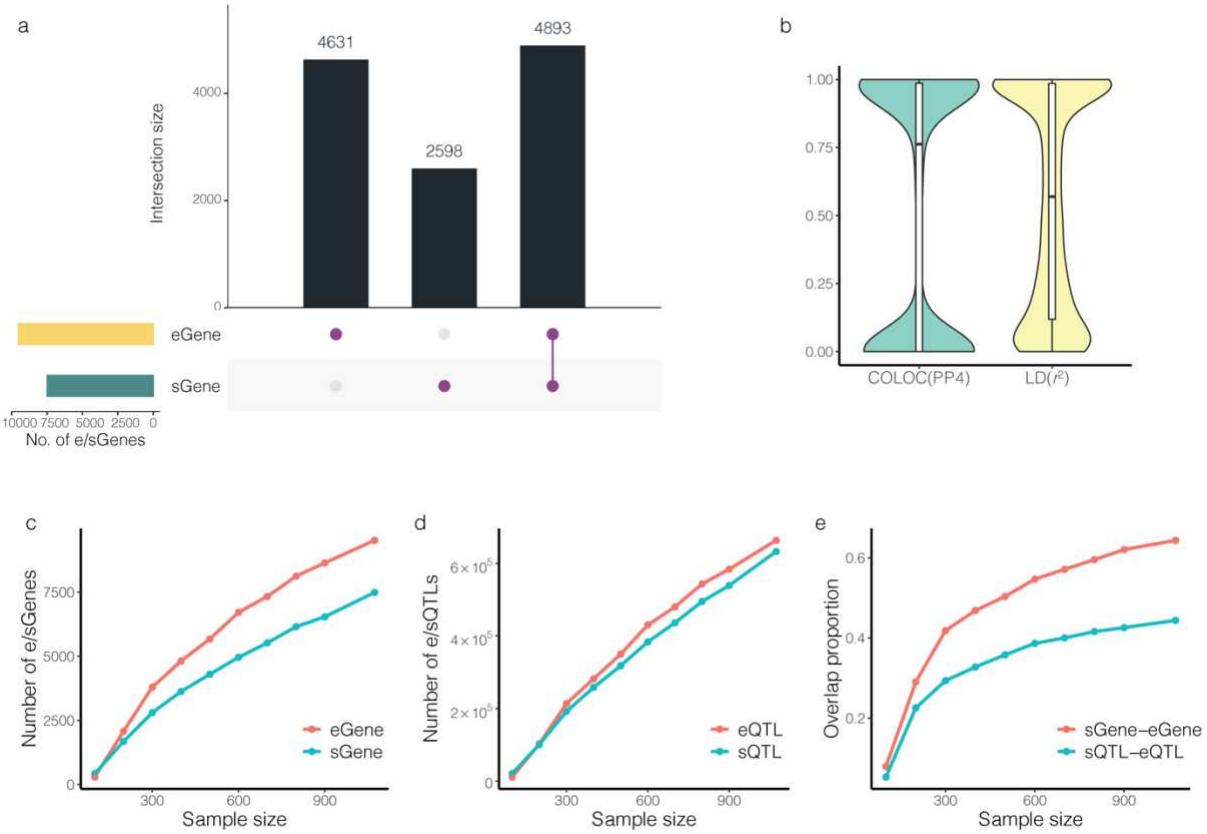

Figure 2. Relationship between sQTLs and eQTLs. **a**) Overlap between sGenes and eGenes in the PsychENCODE data. **b**) LD r^2 or COLOC PP4 between the top cis-sQTL SNP and the top cis-eQTL SNP for the 4,893 overlapping genes. **c**) The number of eGenes (or sGenes) discovered as a function of sample size. **d**) The number of eQTLs (or sQTLs) discovered as a function of sample size. **e**) The overlap between sGenes and eGenes (or between sQTLs and eQTLs) as a function of sample size. sQTL-eQTL overlap is defined as the proportion of sGenes for which the top sQTL is a significant eQTL SNP for the same gene.

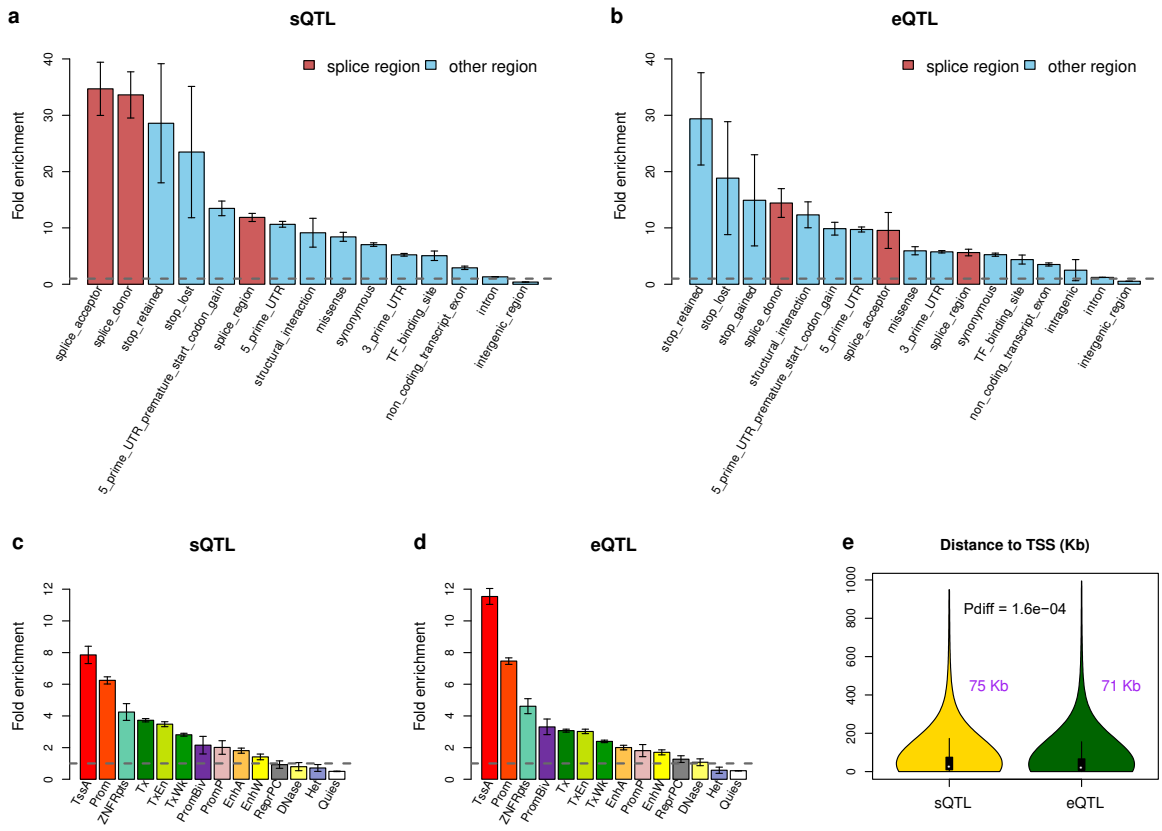

Figure 3. Enrichment of the top cis-sQTL or cis-eQTL SNPs for functional annotations from SnpEff (**a**, **b**) and REMC (**c**, **d**). Fold enrichment is computed by comparing the top cis-sQTL (or cis-eQTL) SNPs in a functional category with the control SNPs. Each error bar represents the 95% confidence interval around an estimate. The grey dashed line represents no enrichment. **e**) Distance of the top cis-sQTL (or cis-eQTL) SNP to TSS of the gene. The text in purple represents the median across genes, and P_{diff} is computed by a *t*-test for a mean difference between the two distributions.

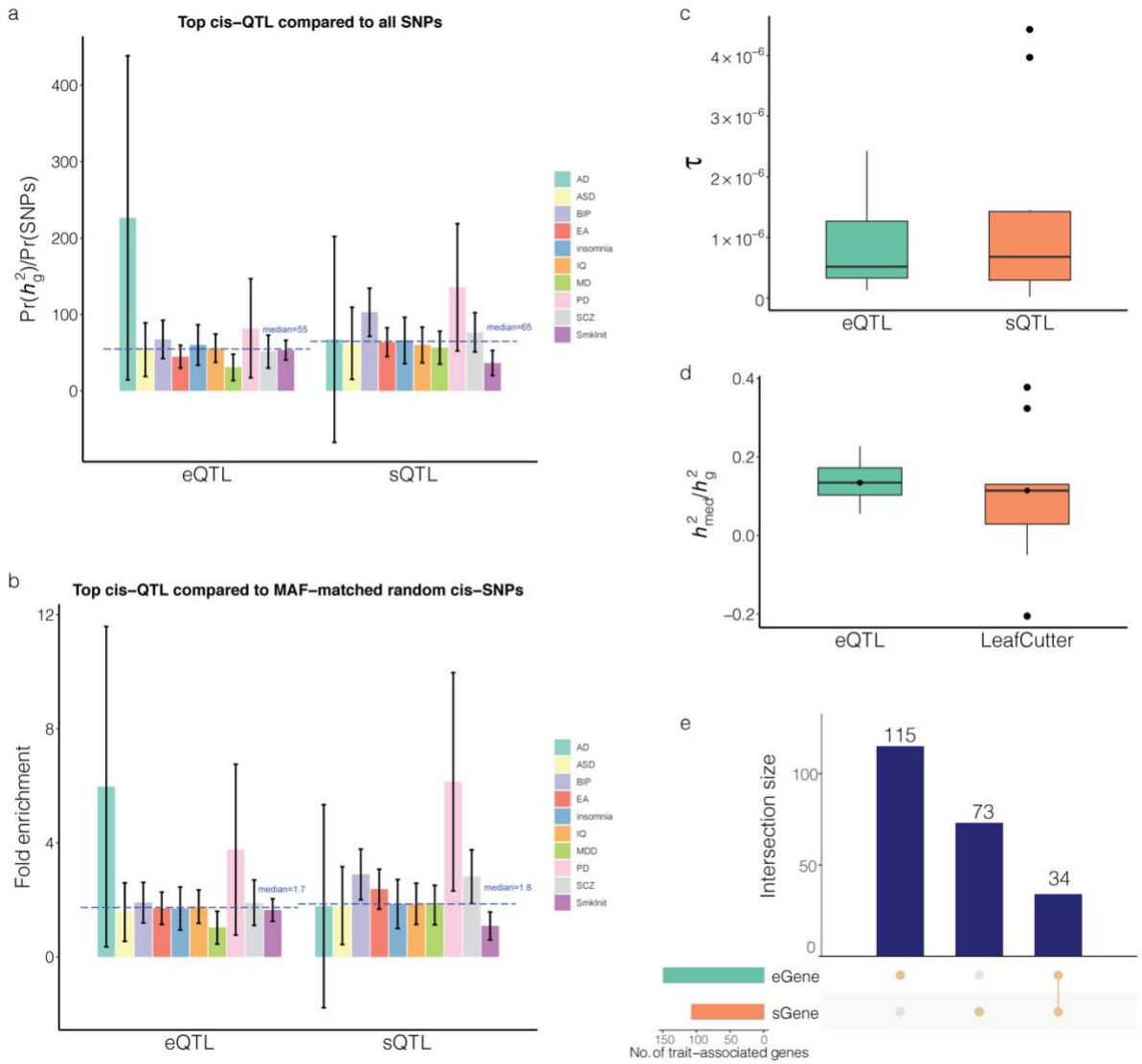

Figure 4. Enrichment of the top cis-sQTL or cis-eQTL SNPs for heritability of the ten brain-related traits. In panel **a**), heritability enrichment is defined as a ratio of the proportion of heritability explained by the SNPs in query (estimated by S-LDSC) to the proportion of the SNPs. In panel **b**), heritability enrichment is defined as a ratio of the proportion of heritability explained by the SNPs in query to that by the control SNPs. In panel **c**), annotation effect size (τ) is used to assess the contribution of the top cis-sQTL SNPs to heritability when fitted jointly with the top cis-eQTL SNPs. Panel **d**) shows the proportion of mediated heritability (h^2_{med}/h^2_g) by the cis-sQTL (or cis-eQTL) SNPs, estimated by MESC. In panels **a**) and **b**), each error bar represents the 95% confidence interval of an estimate, and the blue dashed line represents the median value across traits. Only the sQTLs discovered by LeafCutter were included in the MESC analysis. Panel **e**) shows the overlap of the trait-associated sGenes identified by SMR and COLOC with the trait-associated eGenes.

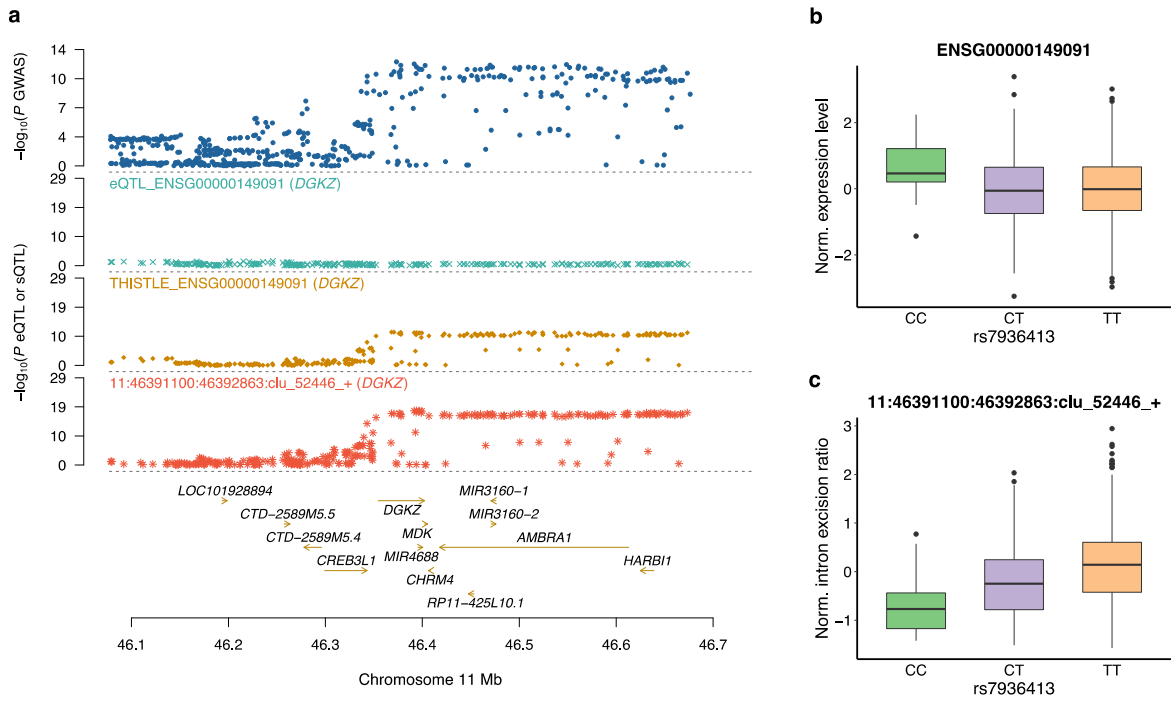

Figure 5. Association of *DGKZ* with SCZ through sQTLs but not eQTLs. **a**) The GWAS, sQTL, and eQTL p-values. The top plot shows −log10(p-values) of SNPs from the GWAS meta-analysis for SCZ. The second, third, and fourth plots show −log10(p-values) from the eQTL analysis, THISTLE sQTL analysis, and LeafCutter sQTL analysis (intron 11:46391100:46392863:clu_52446_+) for *DGKZ*, respectively. **b**) Association of rs7936413 (the top sQTL SNP) with the overall mRNA abundance of *DGKZ*. Each dot represents mRNA abundance of an individual. **c**) Association of rs7936413 with intron excision ratio of 11:46391100:46392863:clu_52446_+. Each dot represents intron excision ratio of an individual.

Figures

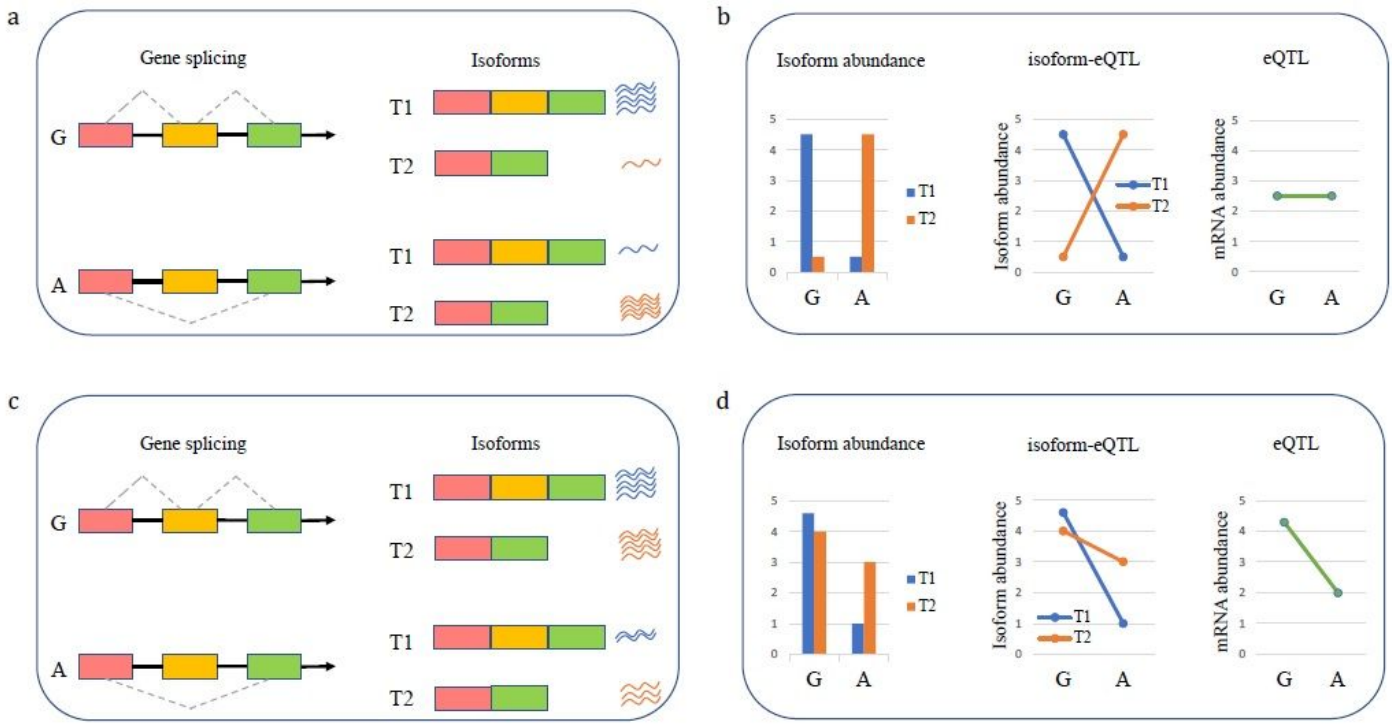

Figure 1

Schematic of THISTLE sQTL analysis. In this toy example, a genetic variant with two alleles, G and A, is associated with a splicing event (e.g., exon skipping) in a gene with two isoforms, T1 and T2. Panels a) and b): a schematic of THISTLE sQTL analysis in the absence of eQTL effect. In this scenario, individuals with G allele show higher mean transcription abundance for T1 than T2, and individuals with A allele show higher mean abundance for T2 than T1 (a), meaning that the genetic variant is associated with the difference in mRNA abundance between the isoforms. In other words, there is a difference in the isoform-eQTL effect between isoforms T1 and T2 (b). However, there is no difference in overall gene expression between individuals with different alleles, meaning that this genetic variant is not an eQTL. Panels c) and d): a schematic of THISTLE sQTL analysis in the presence of eQTL effect. In this scenario, individuals with G allele show similar transcription abundance between T1 and T2, and individuals with A allele show lower abundance for T1 than T2 (c). The isoform-eQTL effect for T1 is different from that for T2 albeit in the same direction (d). In this case, there is a difference in overall gene expression between alleles G and A, meaning that this genetic variant is also an eQTL.

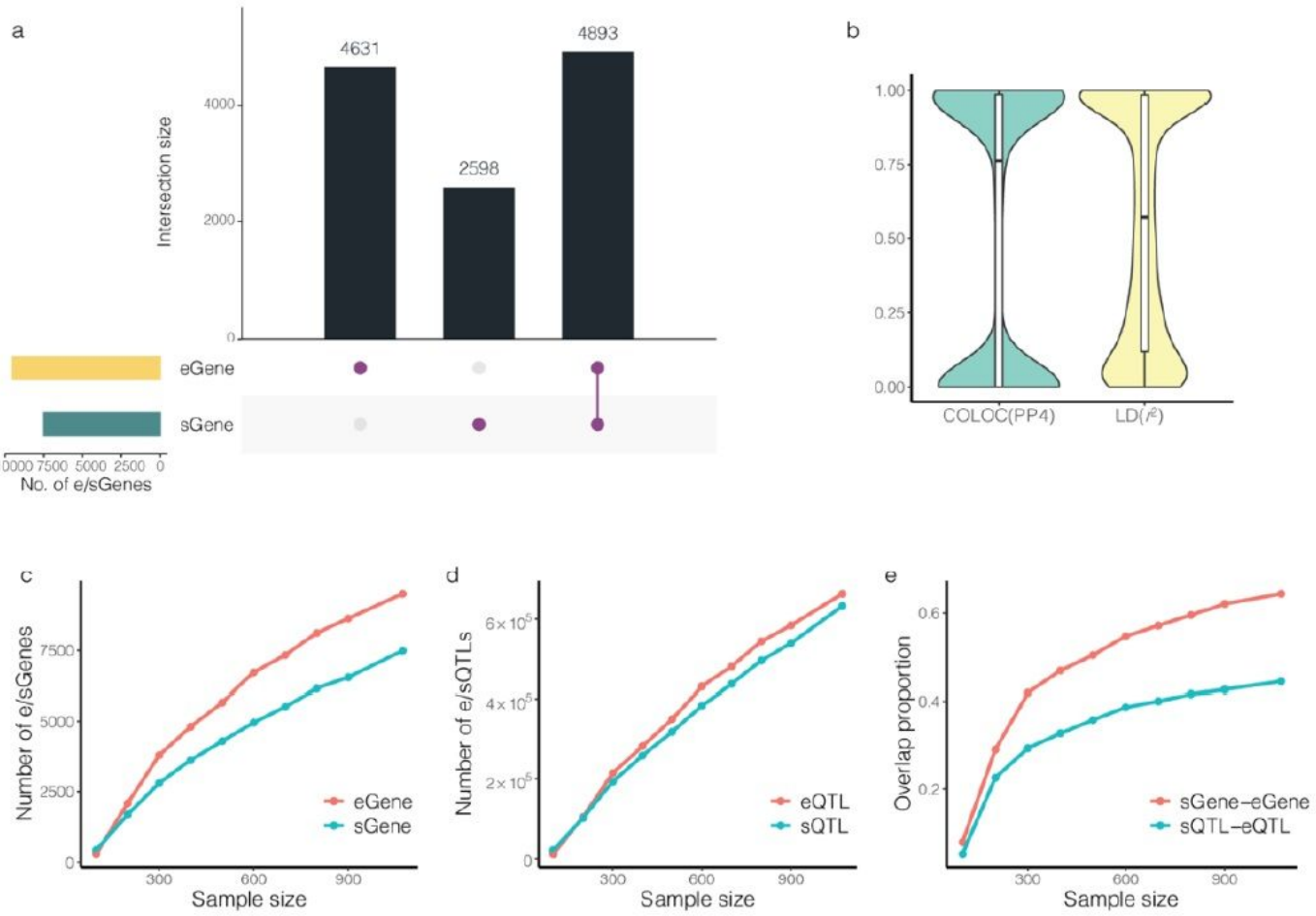

Figure 2

Relationship between sQTLs and eQTLs. a) Overlap between sGenes and eGenes in the PsychENCODE data. b) LD r2 or COLOC PP4 between the top cis-sQTL SNP and the top cis-eQTL SNP for the 4,893 overlapping genes. c) The number of eGenes (or sGenes) discovered as a function of sample size. d) The number of eQTLs (or sQTLs) discovered as a function of sample size. e) The overlap between sGenes and eGenes (or between sQTLs and eQTLs) as a function of sample size. sQTL-eQTL overlap is defined as the proportion of sGenes for which the top sQTL is a significant eQTL SNP for the same gene.

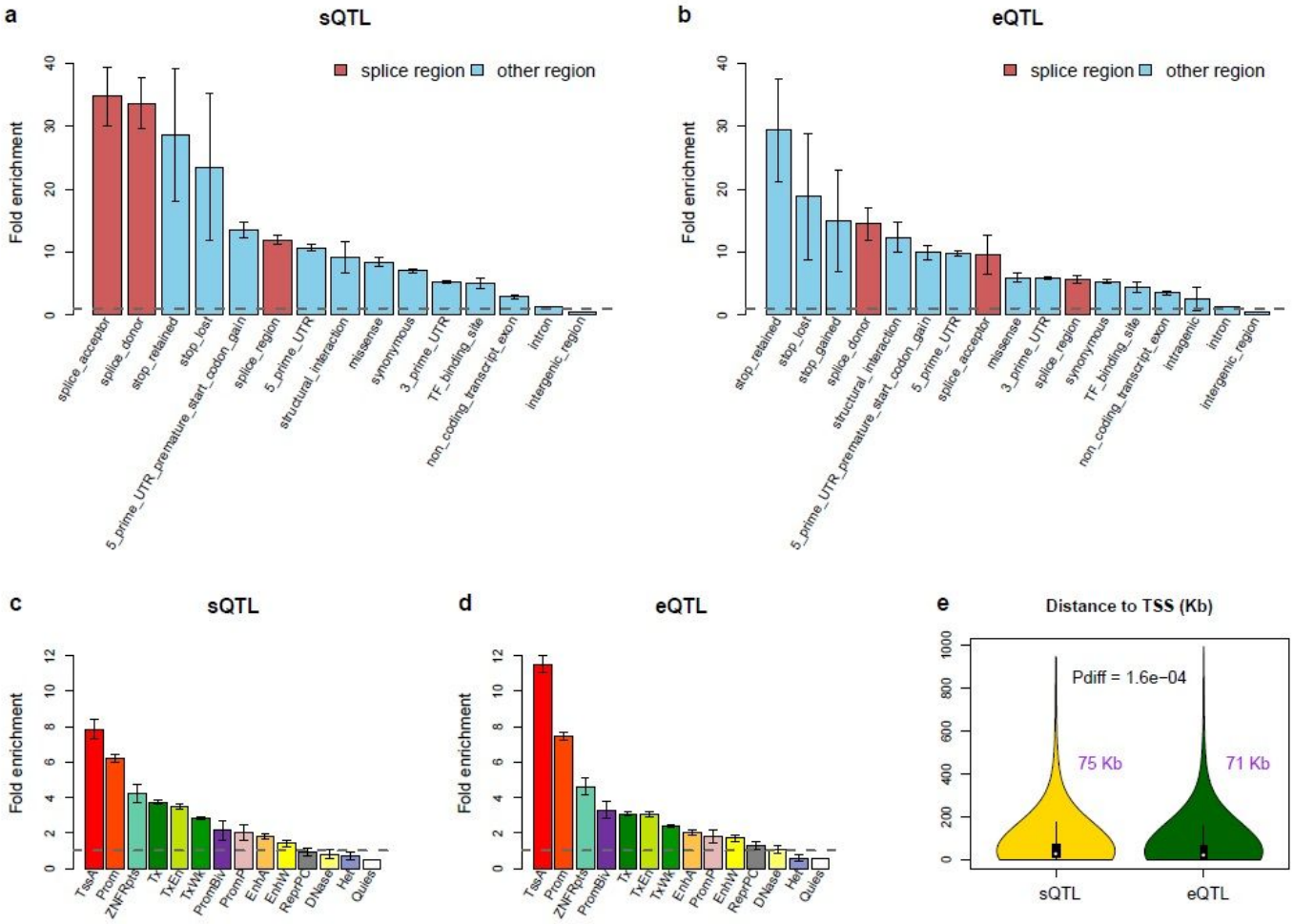

Figure 3

Enrichment of the top cis-sQTL or cis-eQTL SNPs for functional annotations from SnpEff (a, b) and REMC (c, d). Fold enrichment is computed by comparing the top cis-sQTL (or cis-eQTL) SNPs in a functional category with the control SNPs. Each error bar represents the 95% confidence interval around an estimate. The grey dashed line represents no enrichment. e) Distance of the top cis-sQTL (or cis-eQTL) SNP to TSS of the gene. The text in purple represents the median across genes, and Pdiff is computed by a t-test for a mean difference between the two distributions.

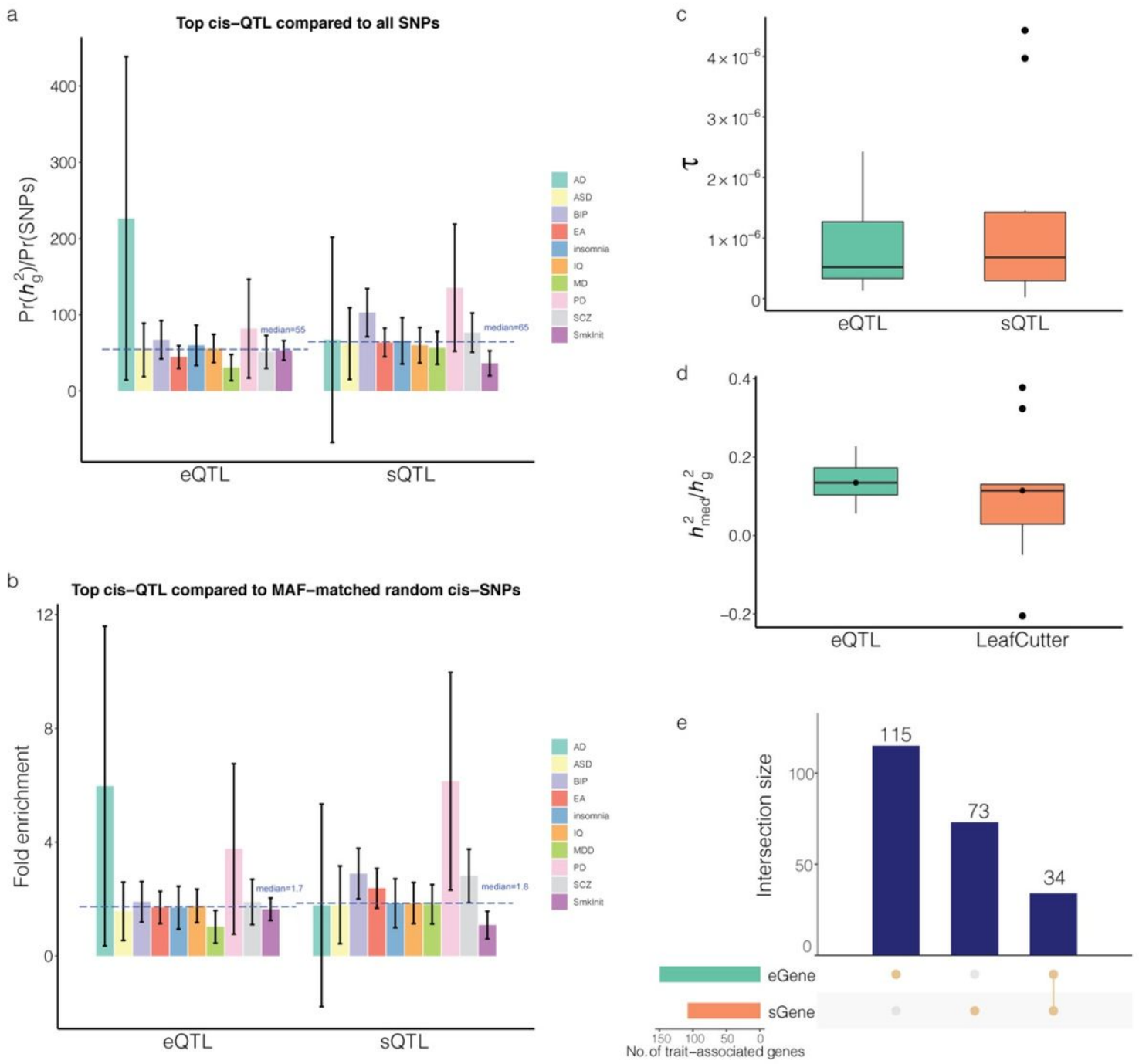

Figure 4

Enrichment of the top cis-sQTL or cis-eQTL SNPs for heritability of the ten brain-related traits. In panel a), heritability enrichment is defined as a ratio of the proportion of heritability explained by the SNPs in query (estimated by S-LDSC) to the proportion of the SNPs. In panel b), heritability enrichment is defined as a ratio of the proportion of heritability explained by the SNPs in query to that by the control SNPs. In panel c), annotation effect size (□) is used to assess the contribution of the top cis-sQTL SNPs to heritability when fitted jointly with the top cis-eQTL SNPs. Panel d) shows the proportion of mediated heritability (□□□□2/□□2) by the cis-sQTL (or cis-eQTL) SNPs, estimated by MESC. In panels a) and b), each error bar represents the 95% confidence interval of an estimate, and the blue dashed line represents the median

value across traits. Only the sQTLs discovered by LeafCutter were included in the MESC analysis. Panel e) shows the overlap of the trait-associated sGenes identified by SMR and COLOC with the trait-associated eGenes.

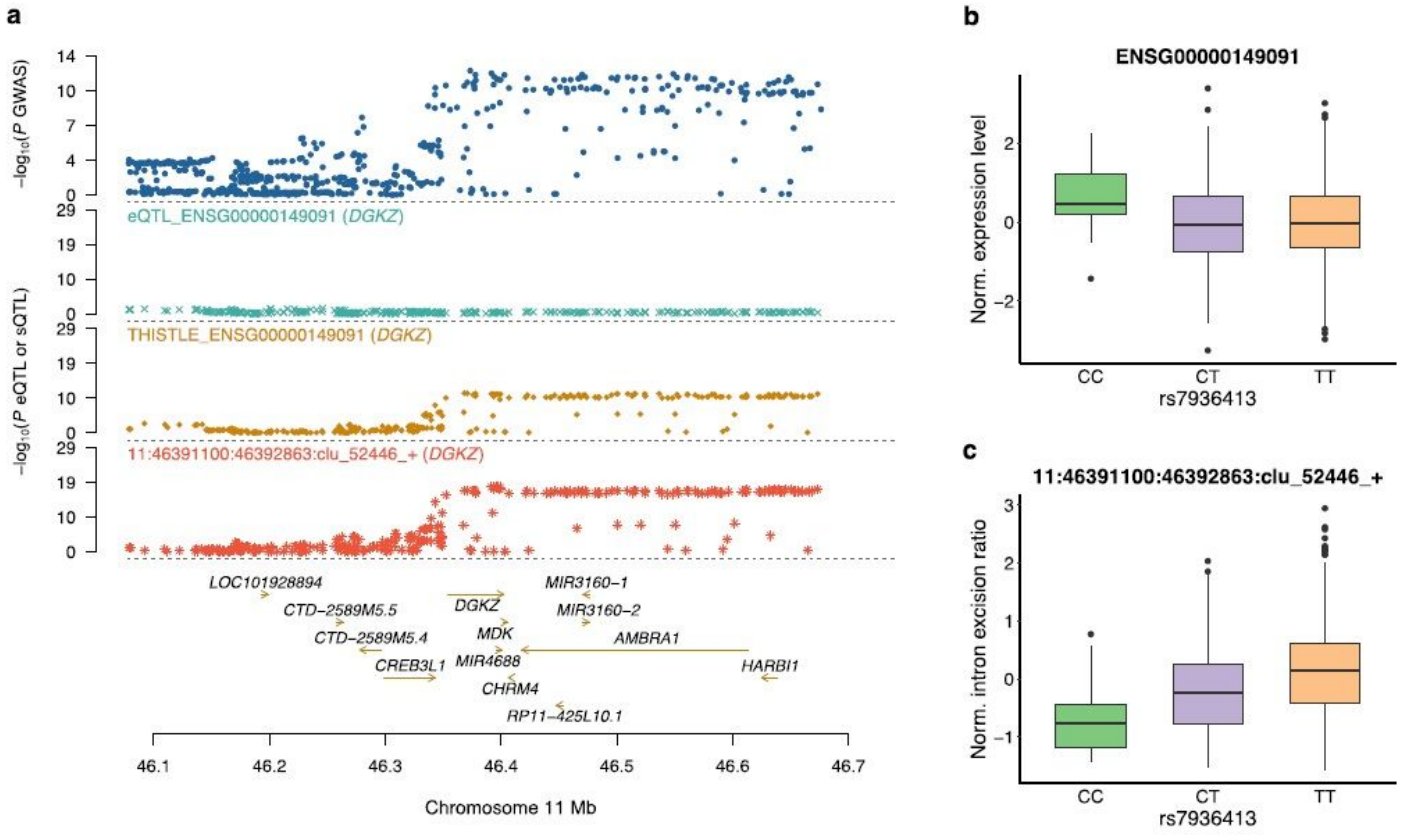

Figure 5

Association of DGKZ with SCZ through sQTLs but not eQTLs. a) The GWAS, sQTL, and eQTL p-values. The top plot shows −log10(p-values) of SNPs from the GWAS meta-analysis for SCZ. The second, third, and fourth plots show −log10(p-values) from the eQTL analysis, THISTLE sQTL analysis, and LeafCutter sQTL analysis (intron 11:46391100:46392863:clu_52446_+) for DGKZ, respectively. b) Association of rs7936413 (the top sQTL SNP) with the overall mRNA abundance of DGKZ. Each dot represents mRNA abundance of an individual. c) Association of rs7936413 with intron excision ratio of 11:46391100:46392863:clu_52446_+. Each dot represents intron excision ratio of an individual.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- sQTLSoM25Jan2021.pdf