Supplementary Information

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], 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

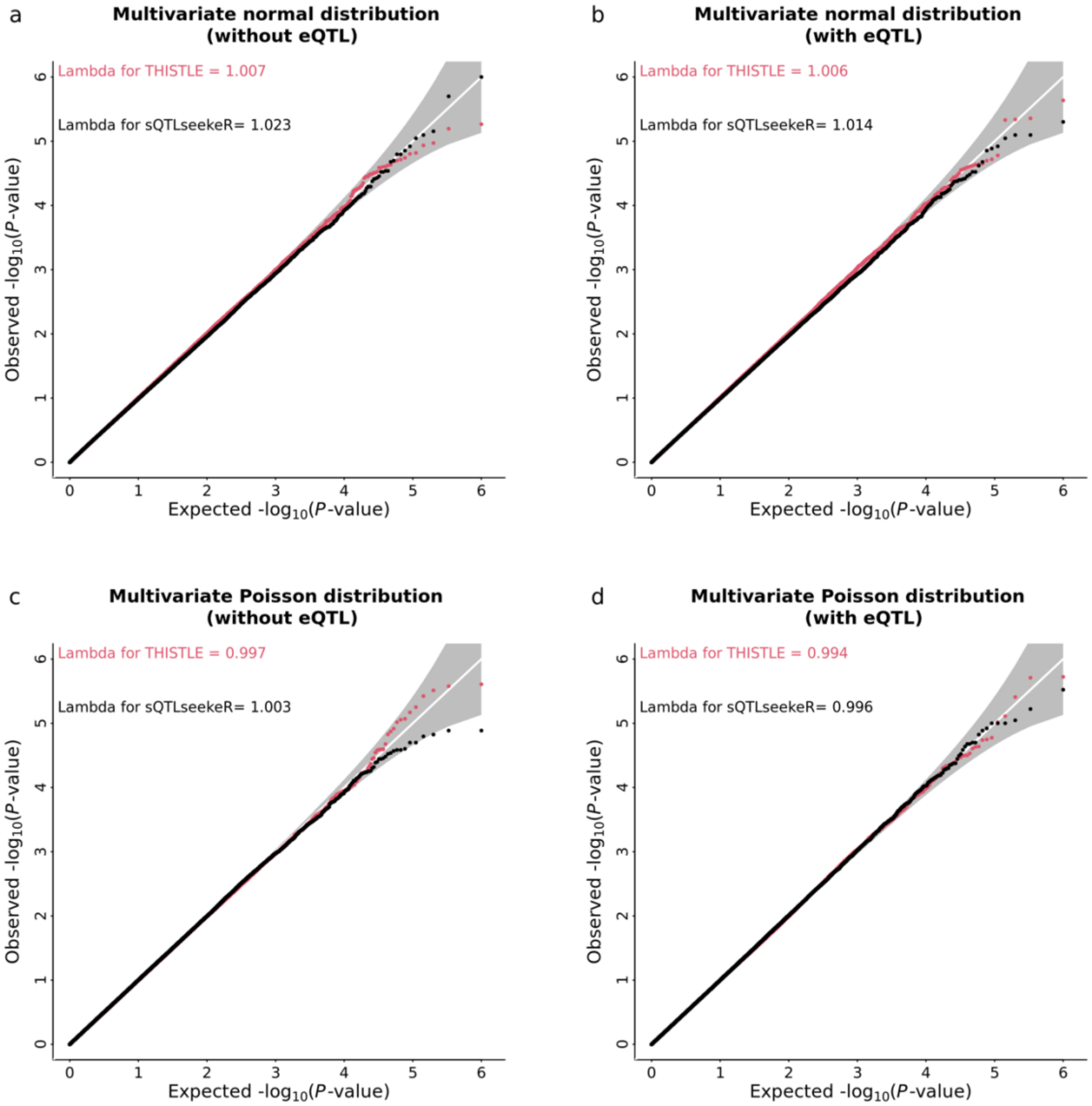
[3]Institute for Molecular Bioscience, The University of Queensland, Brisbane, Queensland 4072, Australia

[4]Neuroscience Research Australia, Sydney, New South Wales, 2031

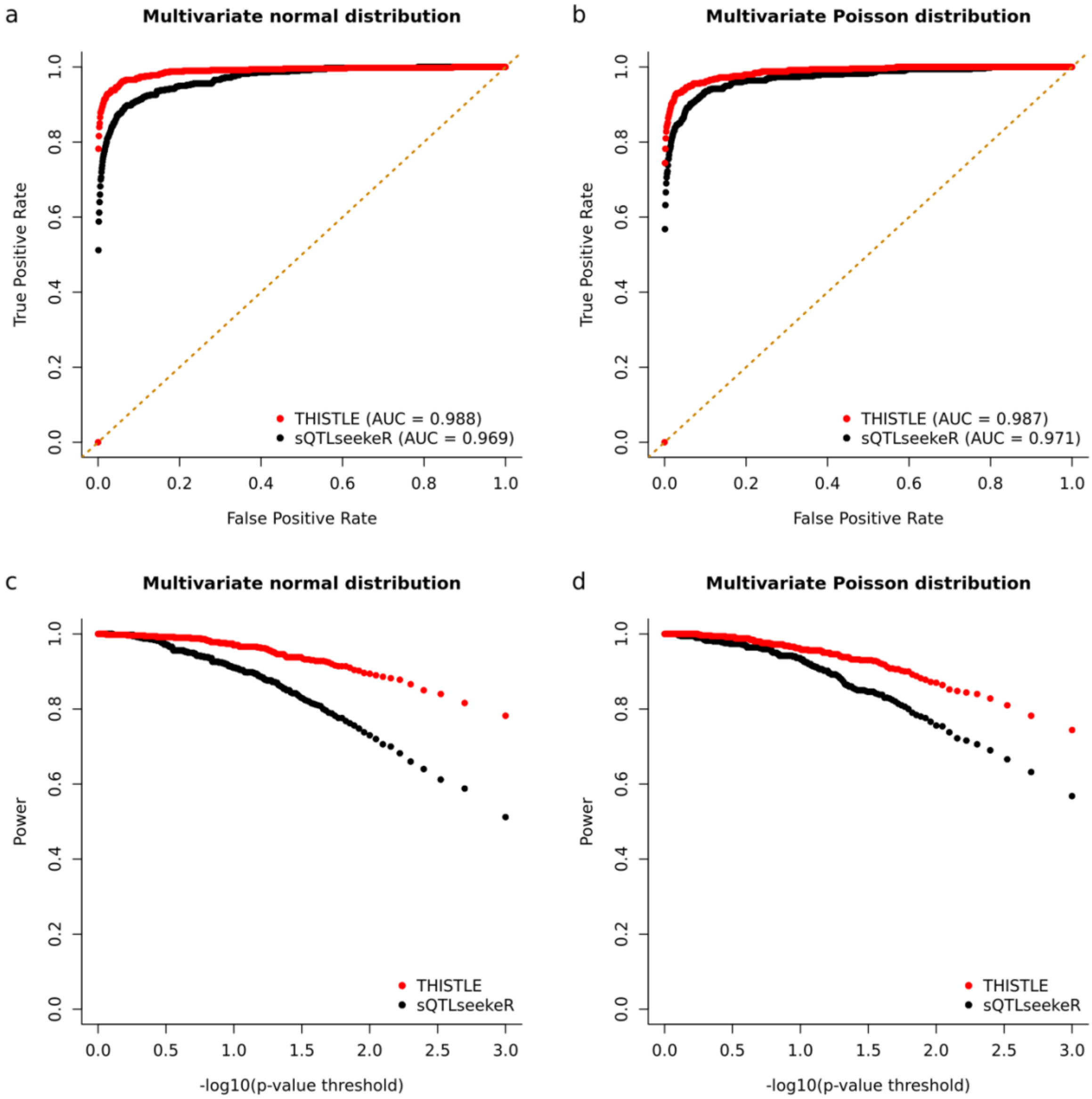
[5]Clinical Genetics and Genomics, NSW Health Pathology Randwick, Sydney, New South Wales, 2031

[*]Correspondence: Jian Yang (jian.yang@westlake.edu.cn)

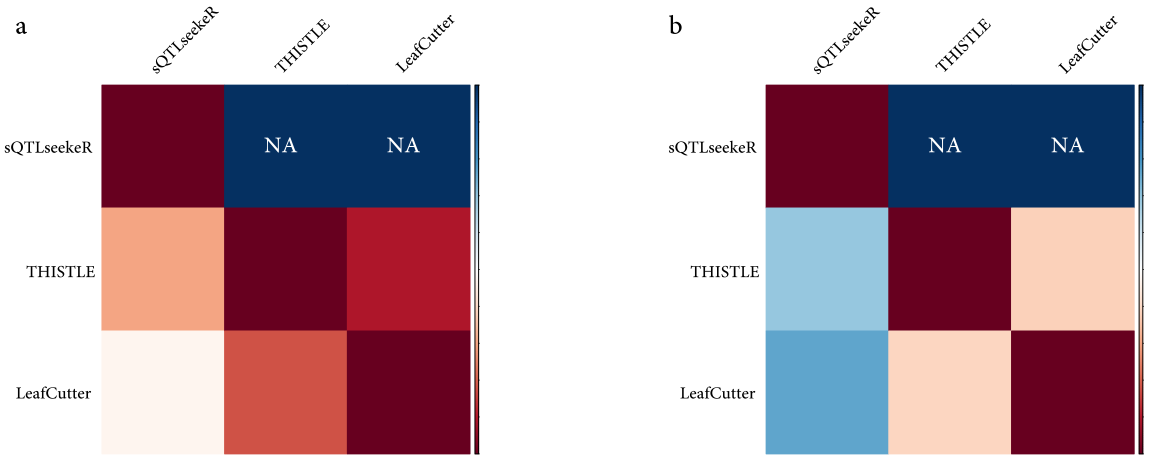
Supplementary Figures



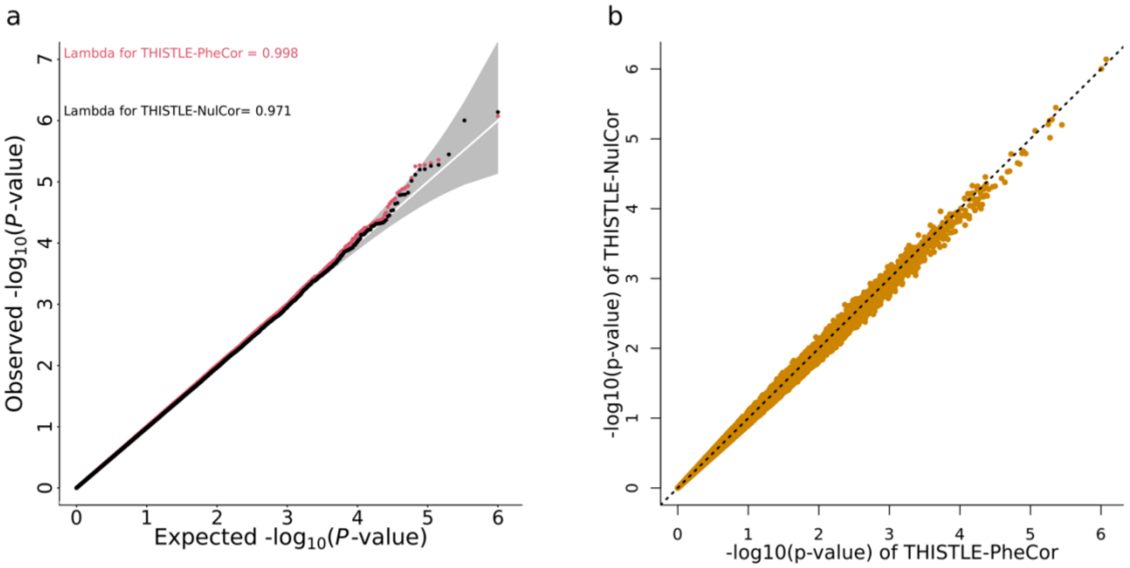
Supplementary Figure 1. Quantile-quantile (QQ) plots for sQTL analysis under the null of no sQTL effect using THISTLE (red) and sQTLseekeR (black). Panels **a**) and **b**) show the QQ plots for sQTL analysis with the transcription abundance simulated from a multivariate normal distribution (with or without eQTL effect). Panels **c**) and **d**) show the QQ plots for sQTL analysis with the transcription abundance simulated from a multivariate Poisson distribution.



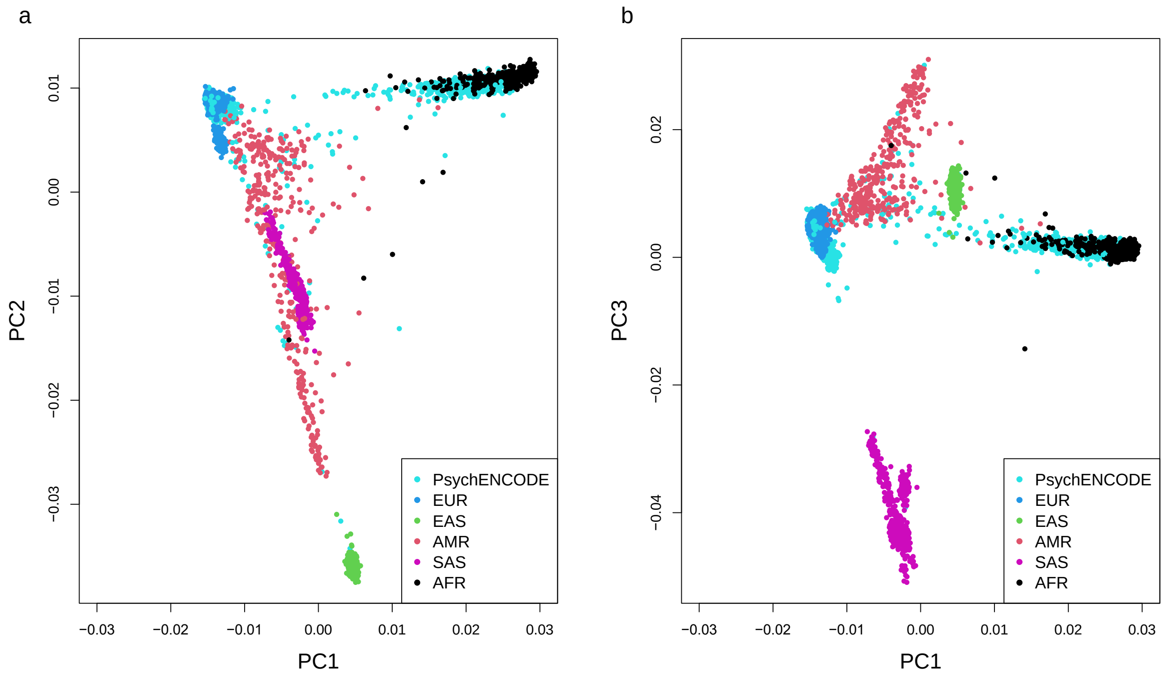
Supplementary Figure 2. The aera under the receiver operating characteristic curve (AUC) and statistical power for THISTLE (red) and sQTLseekeR (black) in simulations. Transcription abundance was simulated from either a multivariate normal distribution (**a** & **c**) or a multivariate Poisson distribution (**b** & **d**).



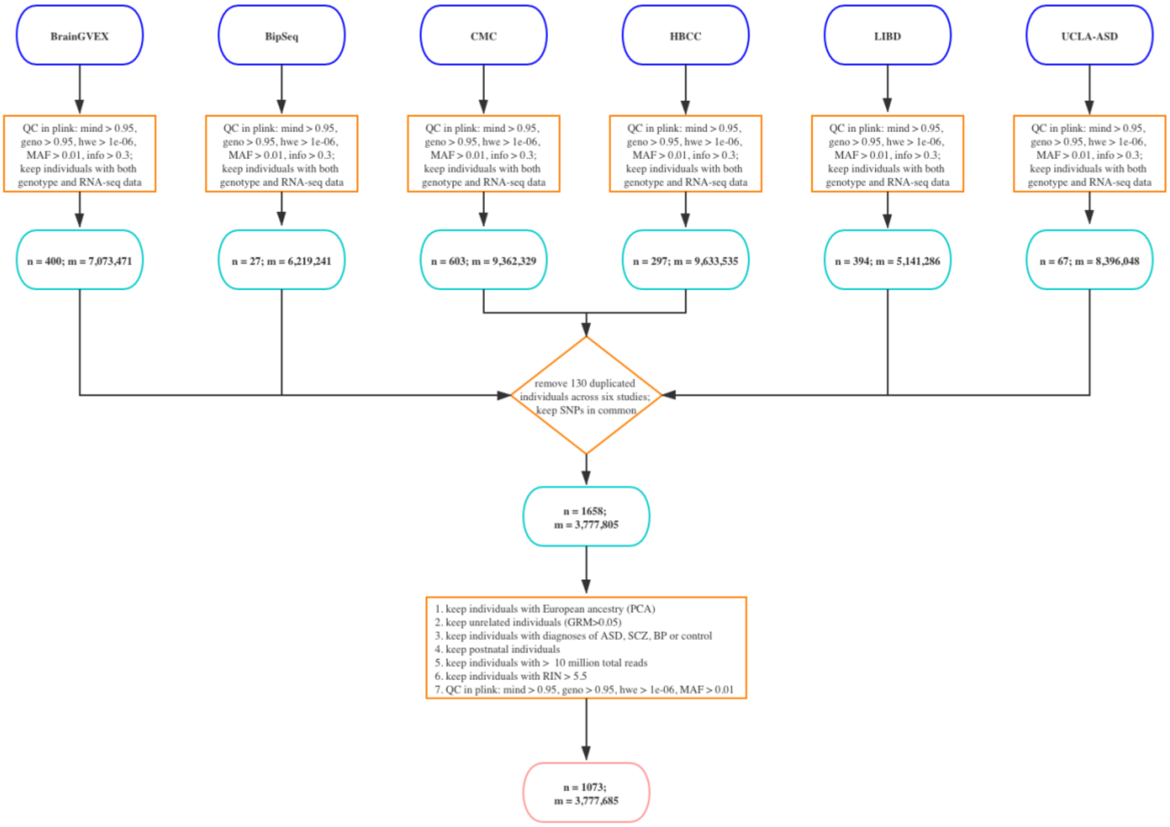
Supplementary Figure 3. Replication of sQTLs between sQTLseekeR, THISTLE, and LeafCutter. Each row represents a method from which the top cis-sQTL SNPs ($P < 5 \times 10^{-8}$) were identified (one SNP per gene), and each column represents a method in which the SNPs were replicated. Note that it was unfeasible to select sQTLs from sQTLseekeR for replication because all the sQTLseekeR p-values were capped at 1×10^{-6}. Panels **a**) and **b**) show the replication results at $P_{\mathrm{sQTL}} < 0.05$ and $P_{\mathrm{sQTL}} < 0.05/m$ (where m is the number of SNPs replicated for each method), respectively.



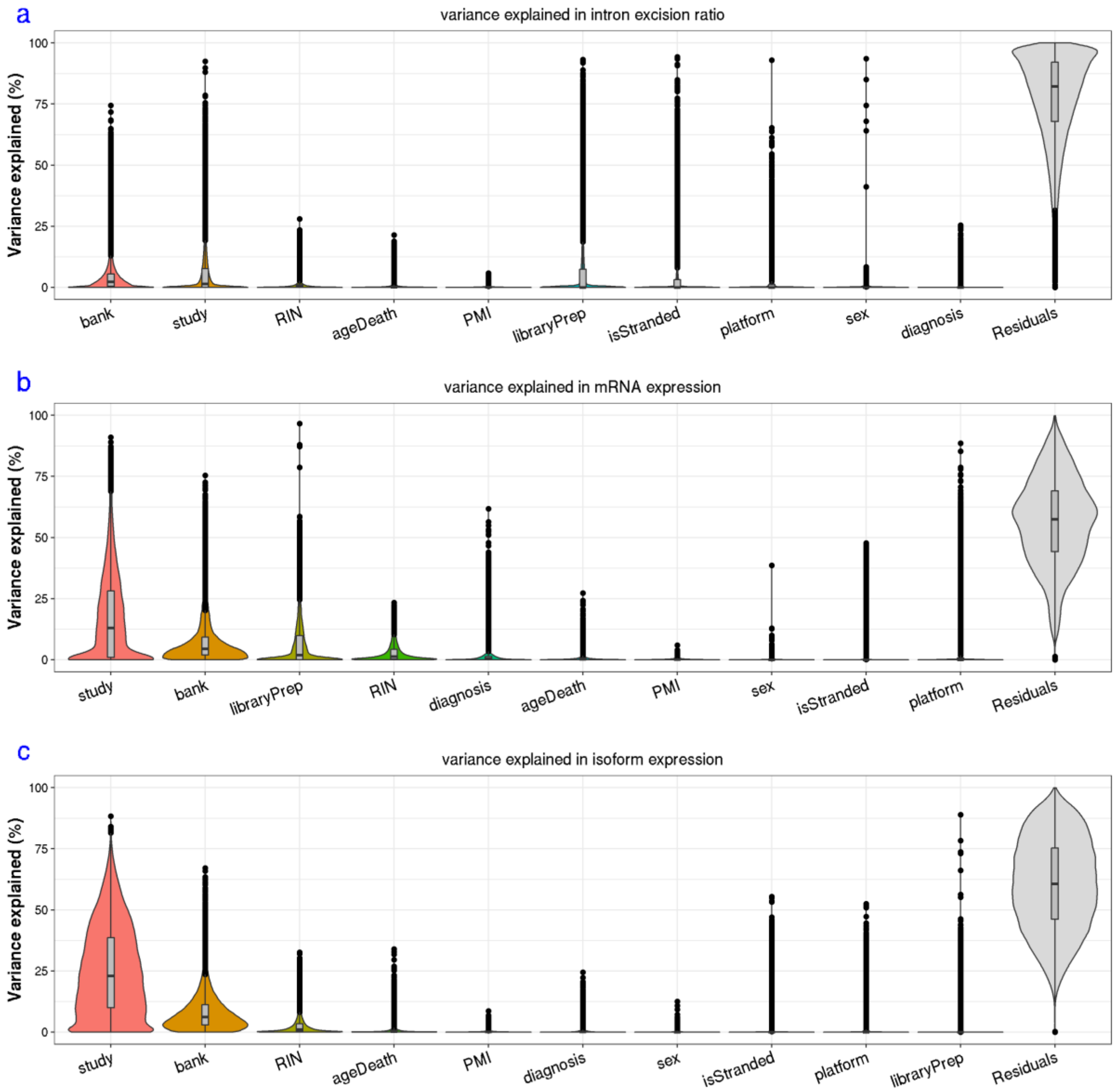
Supplementary Figure 4. Comparison between individual-level data- and summary-level data-based THISTLE. The only difference between the individual-level data- and the summary-level data-based approaches is how the sampling correlation (θ_{jk}) in estimated SNP effect between two isoforms is obtained. a) QQ plot under the null for THISTLE-PheCor (θ_{jk} estimated from observed phenotypes; coloured in red) and THISTLE-NulCor (θ_{jk} estimated from SNPs with $P_{isoform\text{-}eQTL}$ > 0.01 using summary data; coloured in black). b) Comparison of association statistics between THISTLE-PheCor and THISTLE-NulCor. The black dashed line is the diagonal line.



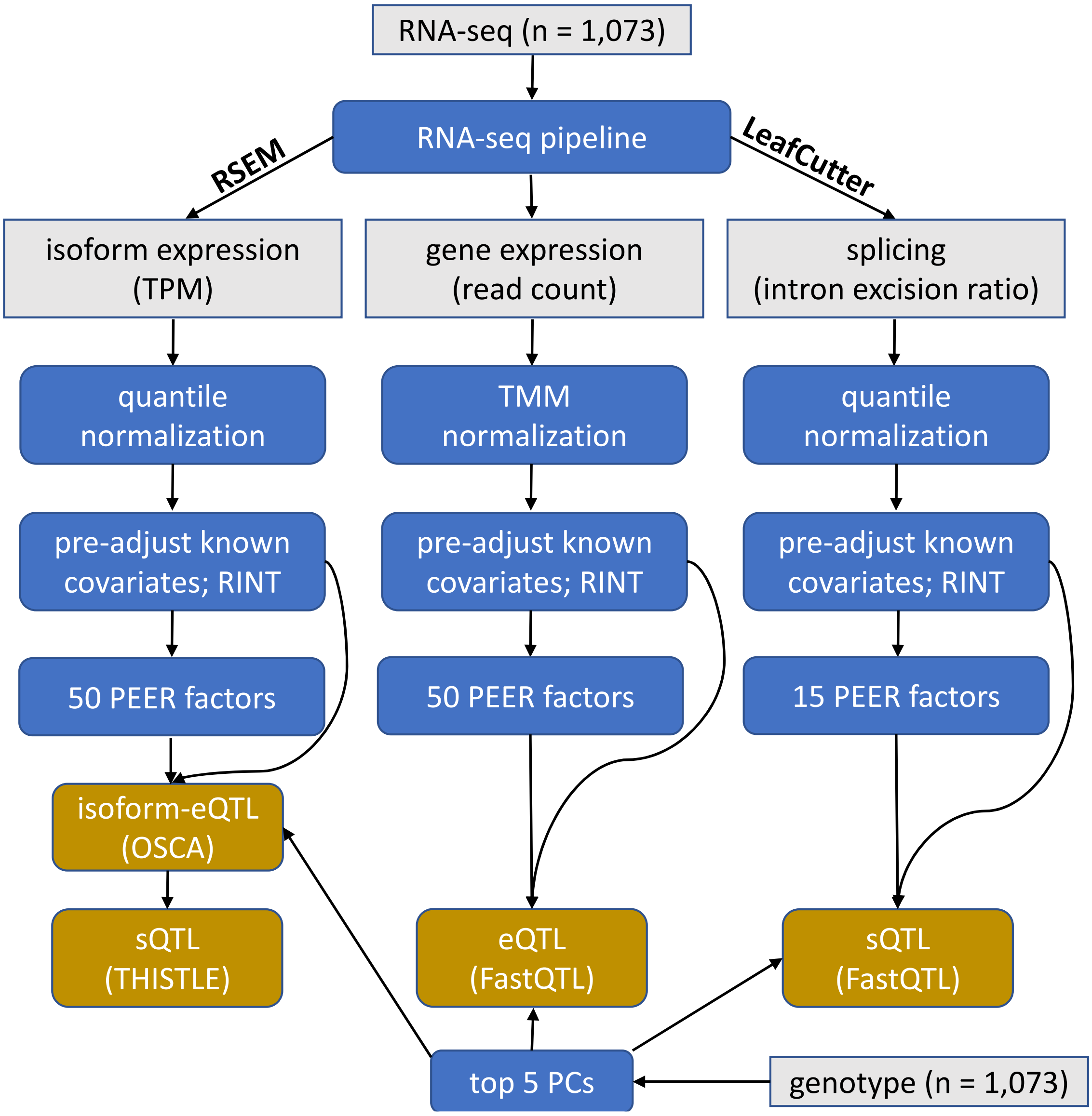
Supplementary Figure 5. Principal component analysis (PCA). The 1000 Genomes Project (1000GP)[1] cohort (*n* = 2,504), comprising whole-genome sequence data from individuals of European (EUR), East Asian (EAS), Admixed American (AMR), South Asian (SAS), and African (AFR) ancestries, was used as a reference panel to demonstrate the population structure in the PsychENCODE cohort (n = 1,658). PCA was performed on a combined genotype data set of the PsychENCODE and 1000GP (593,365 SNPs on 4,162 individuals in total).


BrainGVEX
BipSeq
CMC
HBCC
LIBD
UCLA-ASD
QC in plink: mind > 0.95, geno > 0.95, hwe > 1e-06, MAF > 0.01, info > 0.3; keep individuals with both genotype and RNA-seq data
QC in plink: mind > 0.95, geno > 0.95, hwe > 1e-06, MAF > 0.01, info > 0.3; keep individuals with both genotype and RNA-seq data
QC in plink: mind > 0.95, geno > 0.95, hwe > 1e-06, MAF > 0.01, info > 0.3; keep individuals with both genotype and RNA-seq data
QC in plink: mind > 0.95, geno > 0.95, hwe > 1e-06, MAF > 0.01, info > 0.3; keep individuals with both genotype and RNA-seq data
QC in plink: mind > 0.95, geno > 0.95, hwe > 1e-06, MAF > 0.01, info > 0.3; keep individuals with both genotype and RNA-seq data
QC in plink: mind > 0.95, geno > 0.95, hwe > 1e-06, MAF > 0.01, info > 0.3; keep individuals with both genotype and RNA-seq data
n = 400; m = 7,073,471
n = 27; m = 6,219,241
n = 603; m = 9,362,329
n = 297; m = 9,633,535
n = 394; m = 5,141,286
n = 67; m = 8,396,048
remove 130 duplicated individuals across six studies; keep SNPs in common
n = 1658; m = 3,777,805
1. keep individuals with European ancestry (PCA)
2. keep unrelated individuals (GRM>0.05)
3. keep individuals with diagnoses of ASD, SCZ, BP or control
4. keep postnatal individuals
5. keep individuals with > 10 million total reads
6. keep individuals with RIN > 5.5
7. QC in plink: mind > 0.95, geno > 0.95, hwe > 1e-06, MAF > 0.01
n = 1073; m = 3,777,685

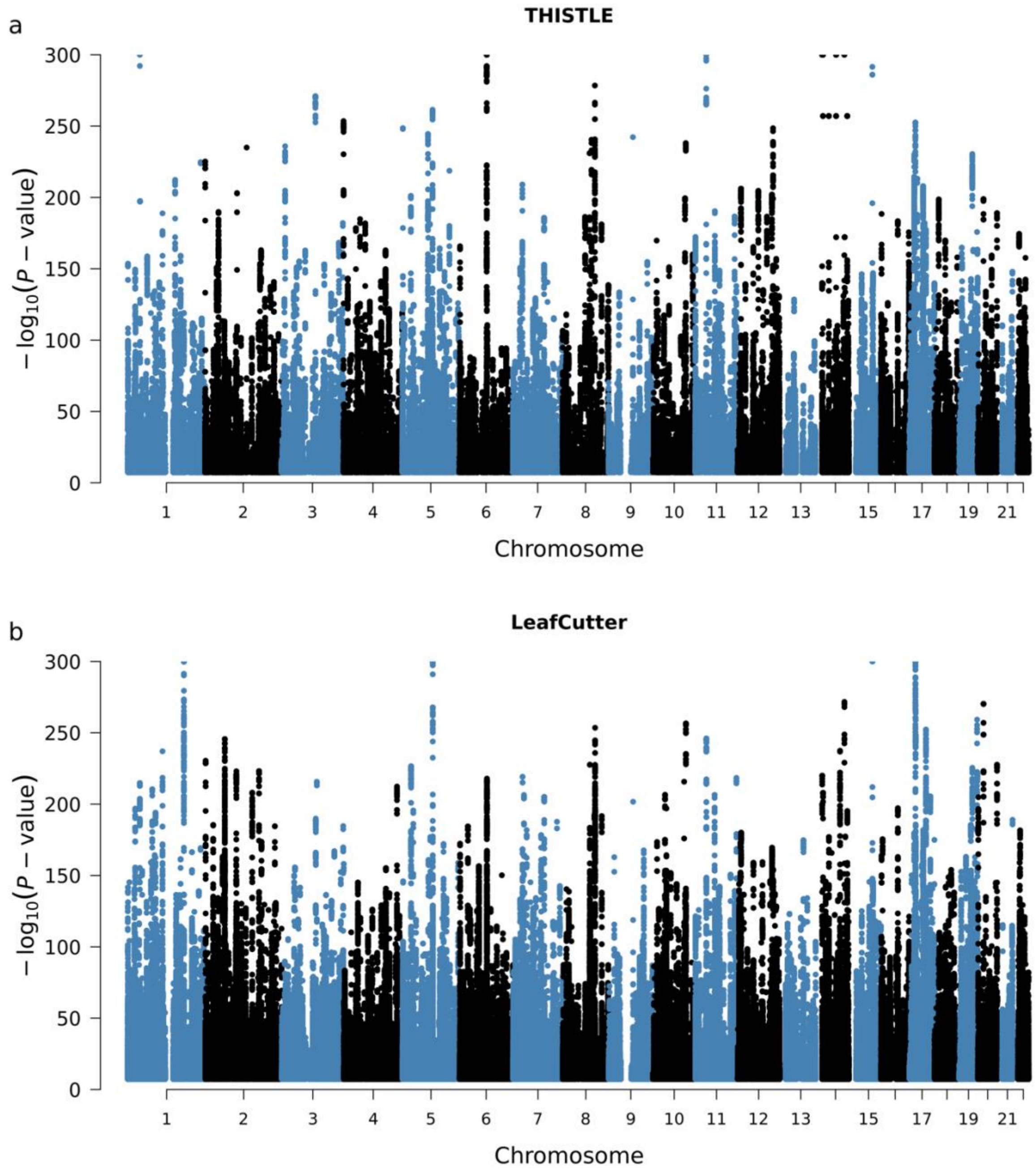
Supplementary Figure 6. Quality control of the genotype and RNA-seq data in each of the 6 cohorts of the PsychENCODE.



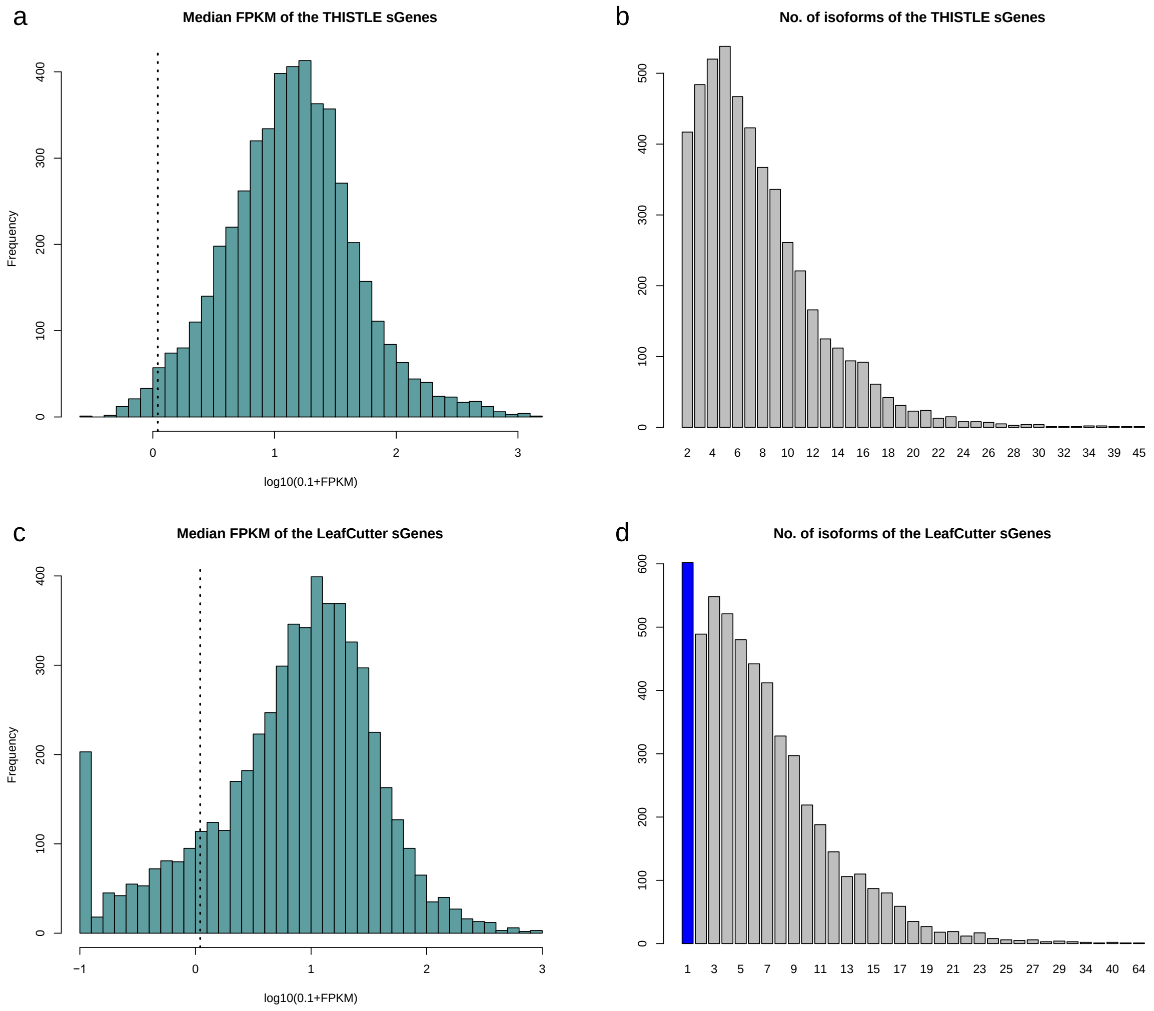
Supplementary Figure 7. Violin and box plots of the proportion of variation in intronic excision ratio (**a**), mRNA abundance (**b**), or isoform abundance (**c**) explained by the biological and technical factors using variancePartition[2]. The factors include study (UCLA-ASD, BrainGVEX, LIBD, CMC, HBCC, BipSeq), isStranded (either paired-end stranded or single-end unstranded libraries), sequencing platform, libraryPrep, RIN (RNA integrity number), ageDeath (age of death), PMI (Post-Mortem Interval), bank (brain bank: MSSM, Mount Sinai brain bank; Pitt, University of Pittsburgh brain bank; Penn, University of Pennsylvania brain bank), sex, and diagnosis (either SCZ, BIP, ASD, or control). Each dot represents an intron (**a**), a gene (**b**), or an isoform (**c**).



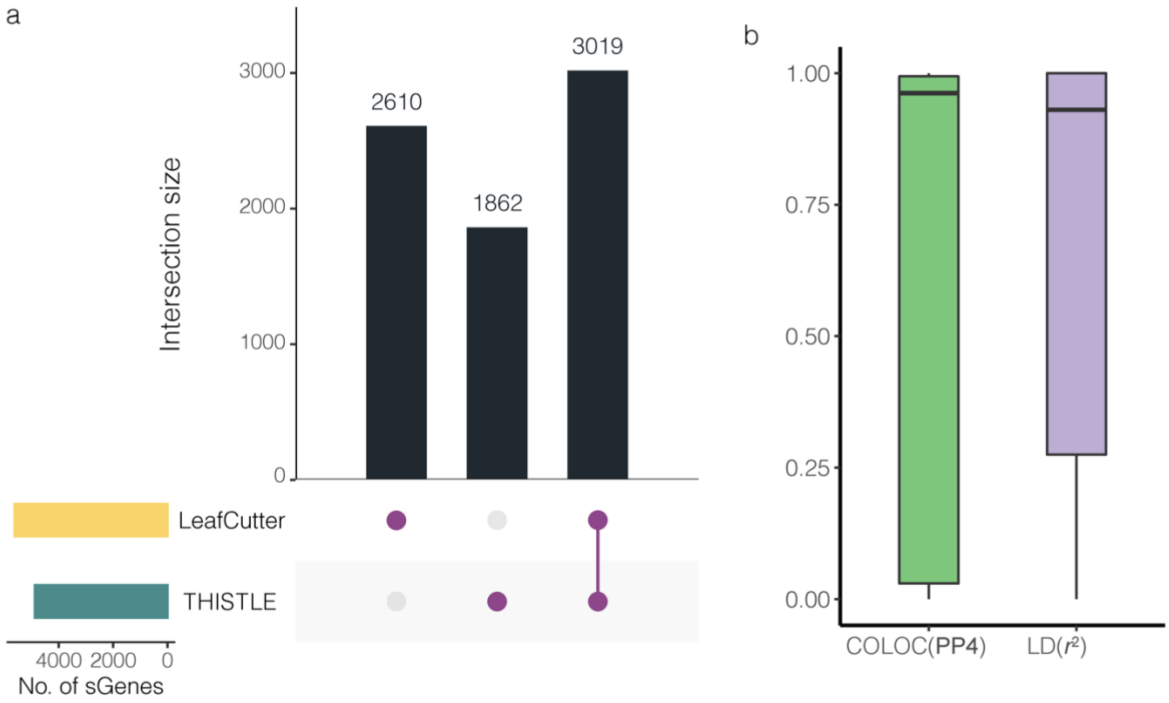
Supplementary Figure 8. Workflow of the eQTL and sQTL analyses in this study.



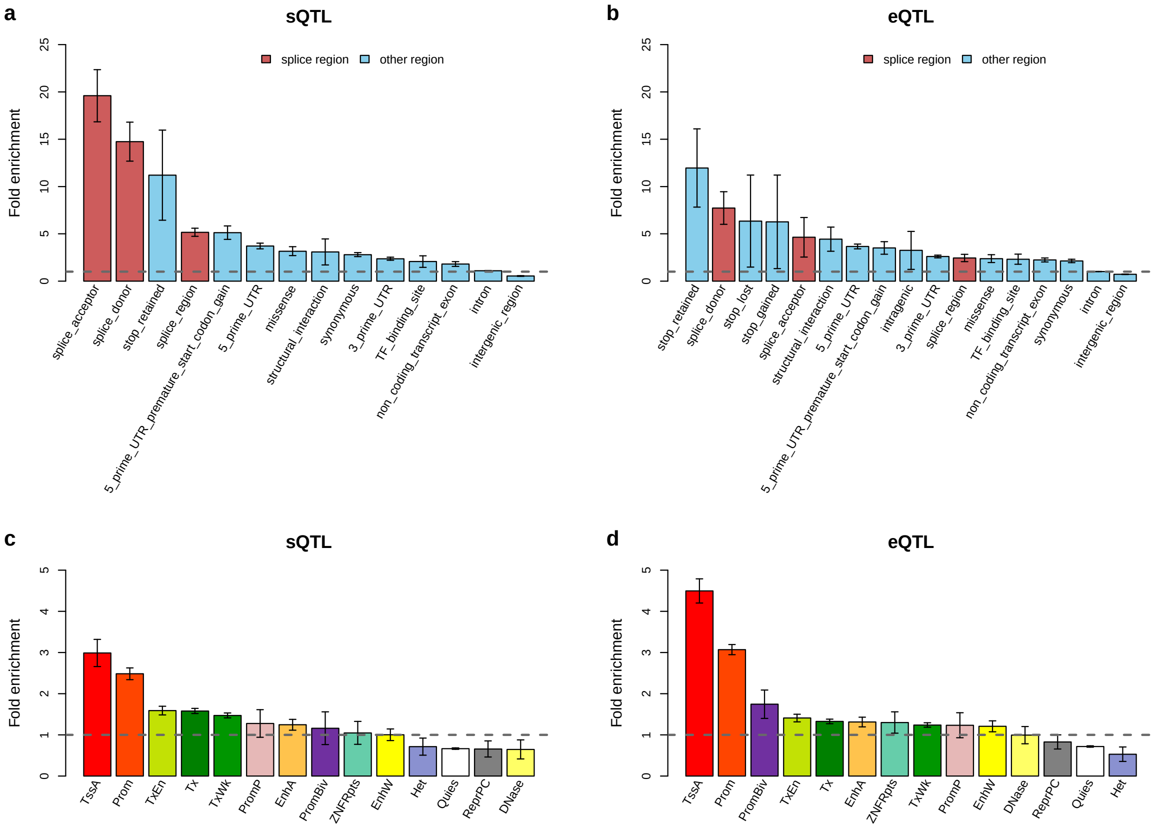
Supplementary Figure 9. Manhattan plots of cis-sQTLs identified by THISTLE (**a**) and LeafCutter (**b**) in the PsychENCODE data.



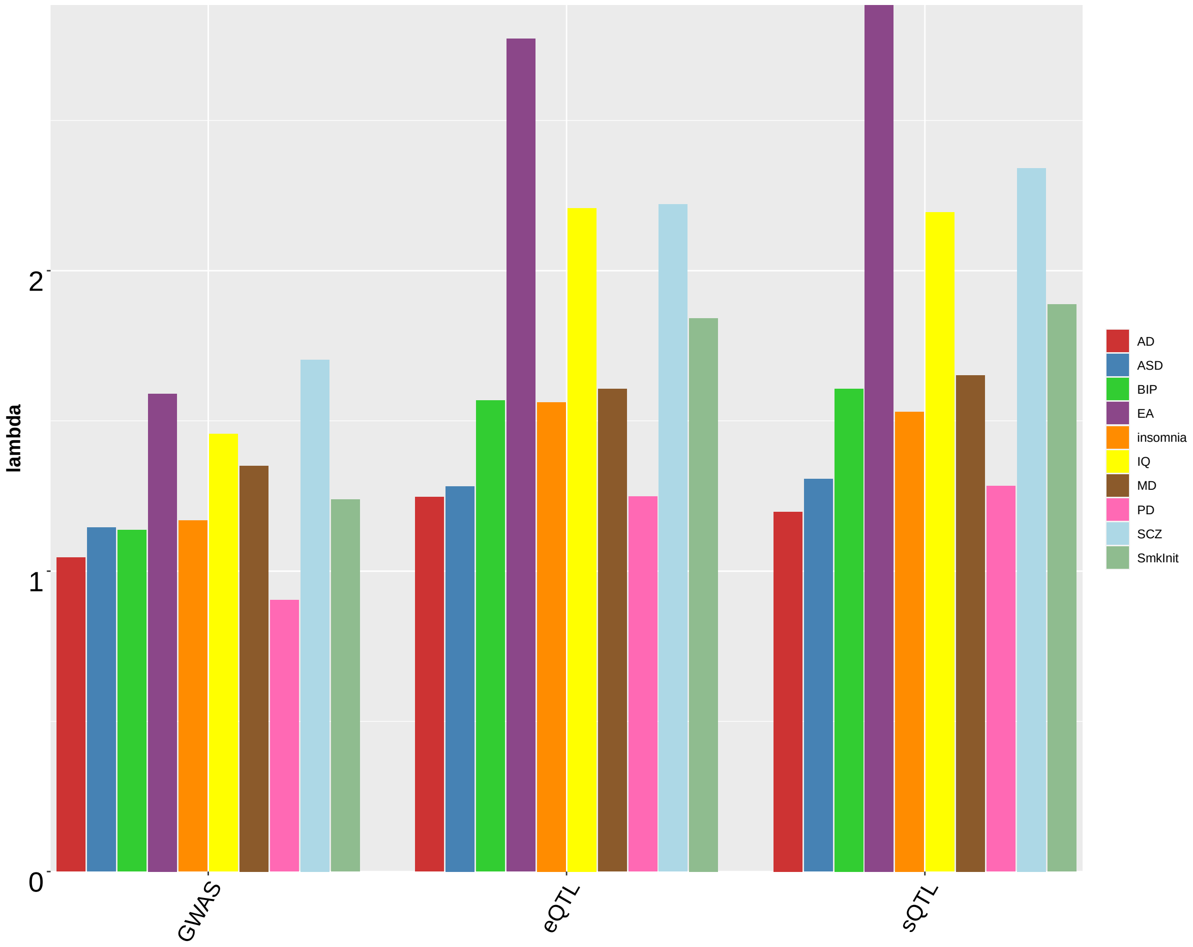
Supplementary Figure 10. FPKM and number of isoforms of the sGenes identified by THISTLE (**a** & **b**) and LeafCutter (**c** & **d**). FPKM: Fragments Per Kilobase of transcript per Million mapped reads. The black dashed lines in panels **a**) and **c**) represent median FPKM of 1. The blue bar in panel **d**) represents isoform number of 1.



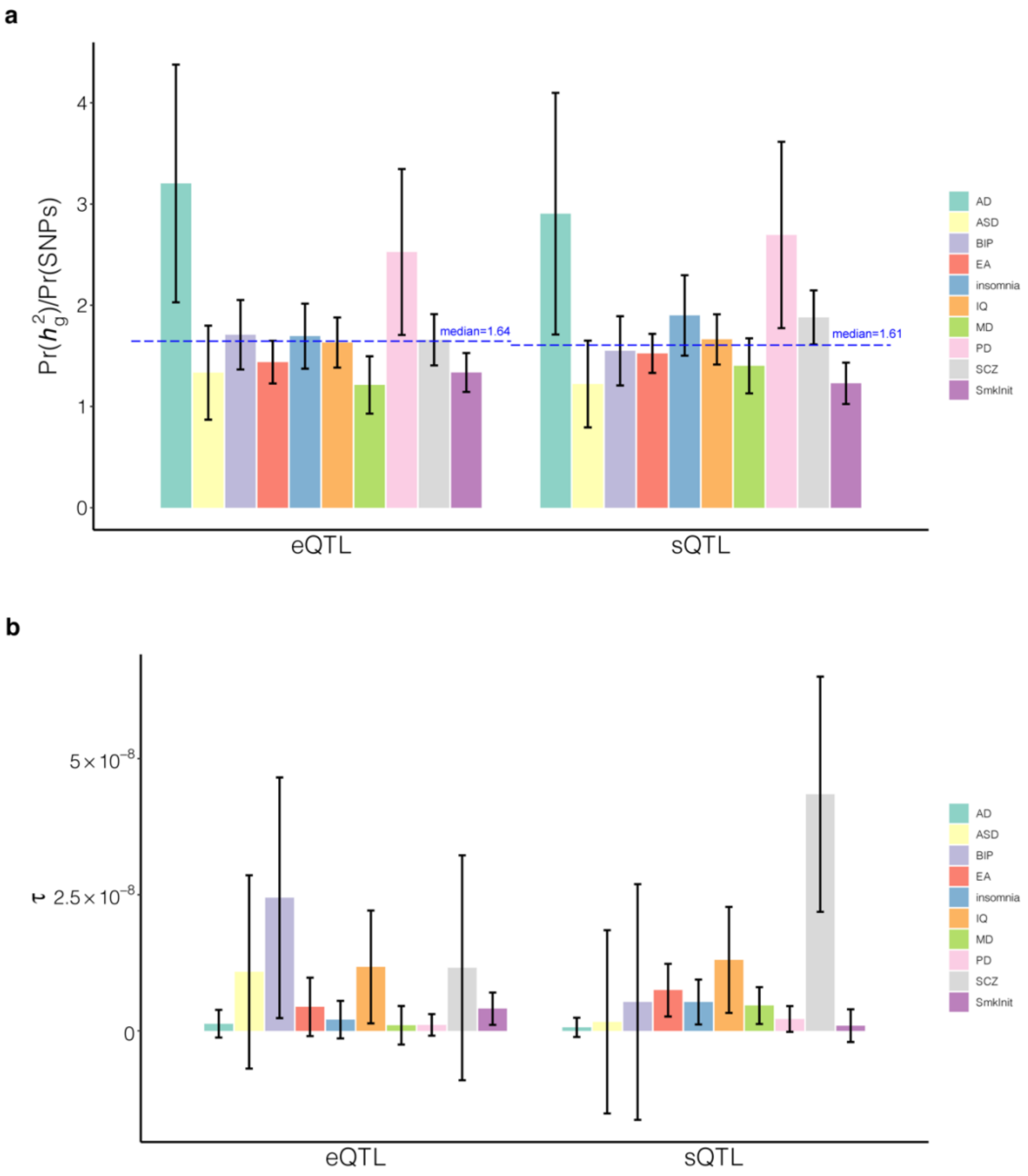
Supplementary Figure 11. Overlap of sQTLs between THISTLE and LeafCutter. **a**) Overlap of sGenes between THISTLE and LeafCutter. **b**) LD r^2 or COLOC PP4 of the top cis-sQTL SNPs between LeafCutter and THISTLE for 3,019 overlapping genes.



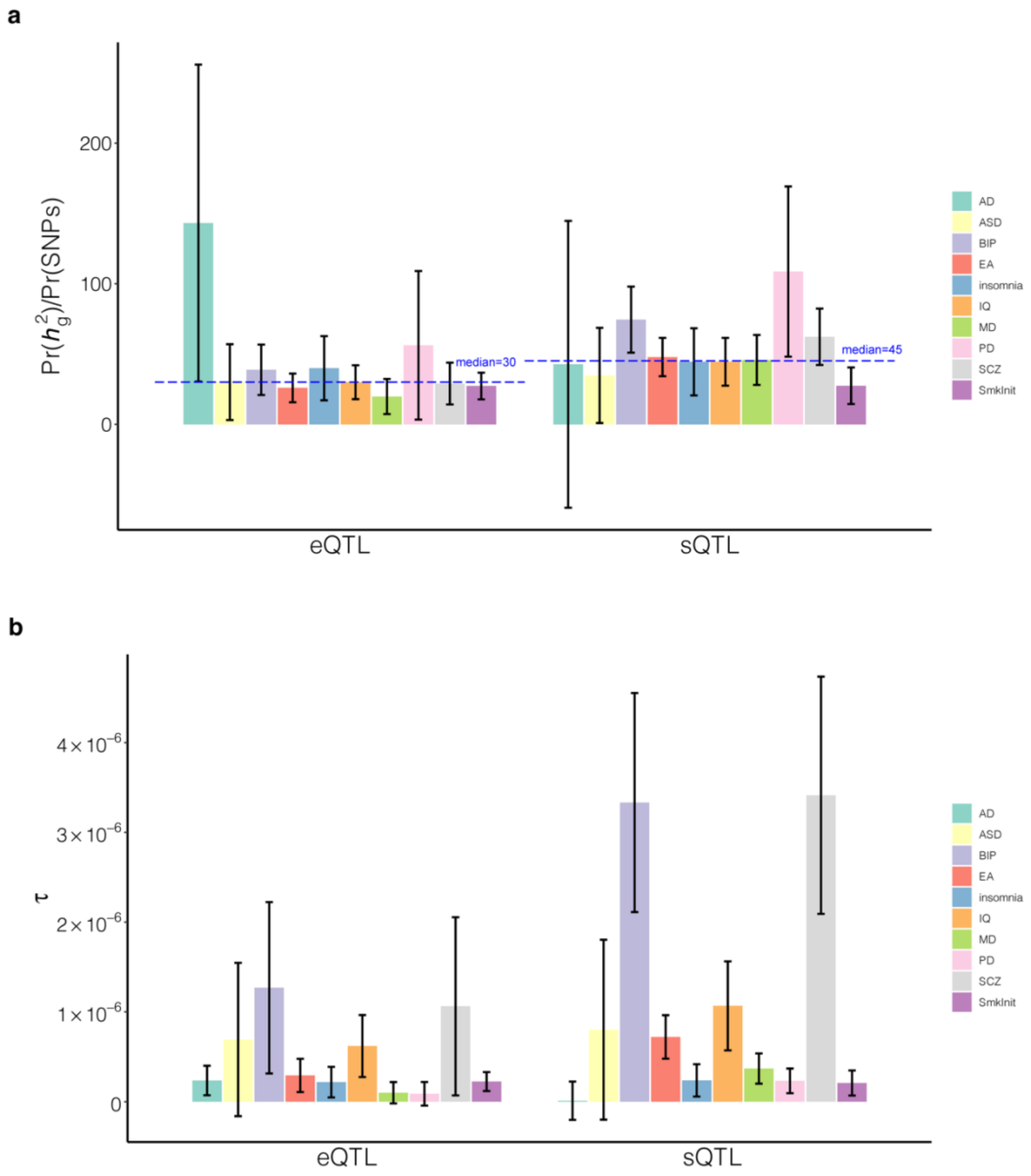
Supplementary Figure 12. Enrichment of the top cis-sQTL or cis-eQTL SNPs for functional categories from SnpEff (**a**, **b**) and REMC (**c**, **d**). Fold enrichment is computed by comparing the top cis-sQTL (or cis-eQTL) SNPs in a functional category with the control SNPs with MAF and TSS matched. Each error bar represents the 95% confidence interval around an estimate. The grey dashed line represents no enrichment.



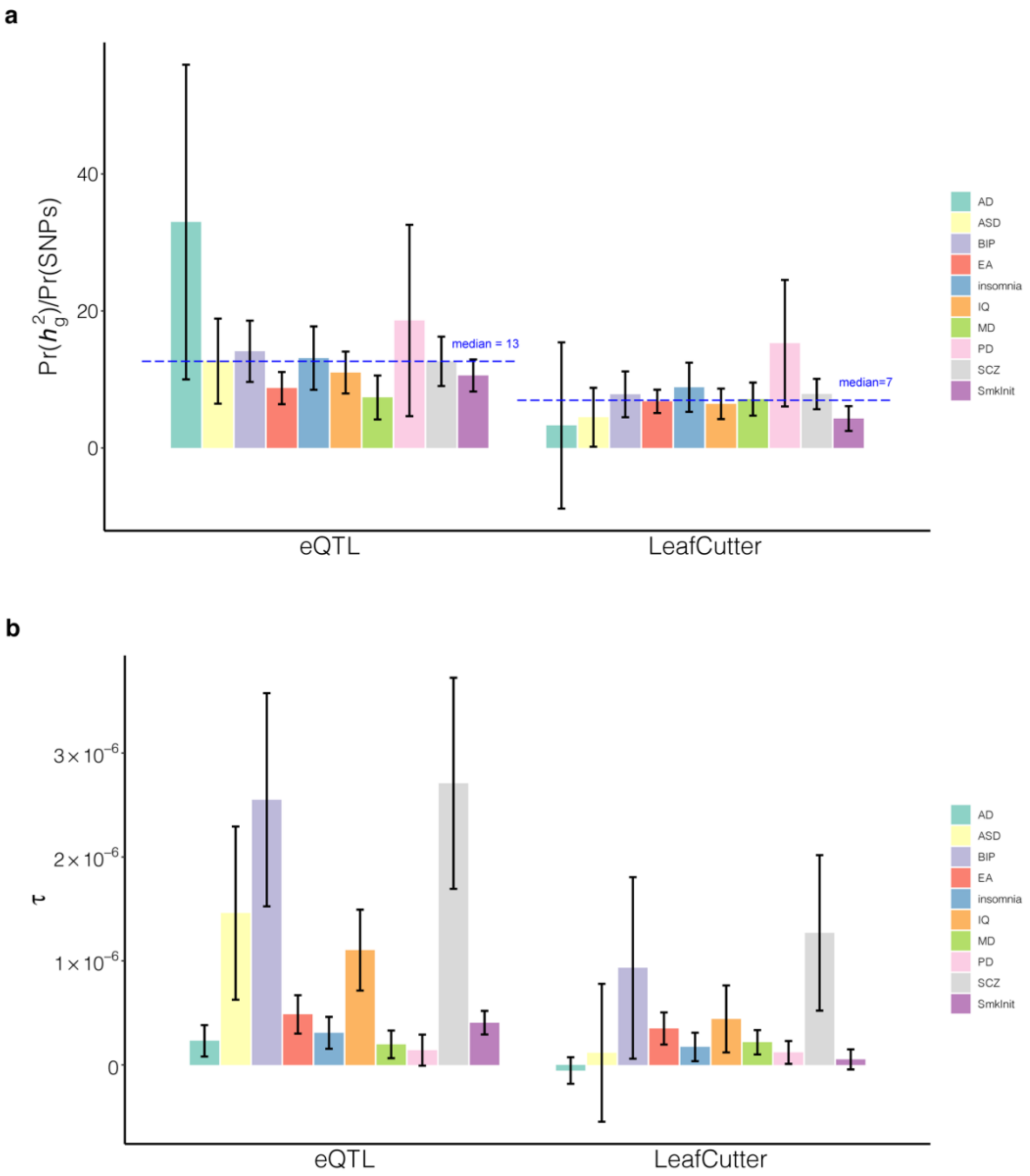
Supplementary Figure 13. Inflation of GWAS test-statistics for the top cis-sQTL SNPs, the top cis-eQTL SNPs, and all SNPs for the ten traits.



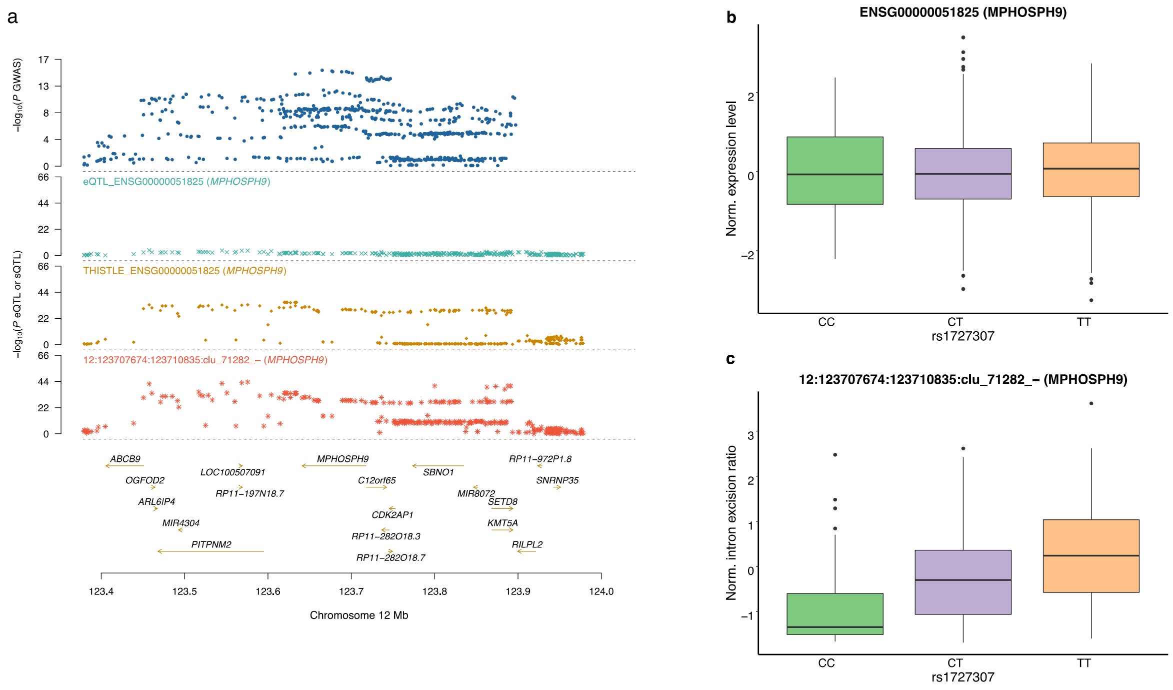
Supplementary Figure 14. Enrichment of all the significant cis-sQTL or cis-eQTL SNPs for heritability of the ten brain-related traits. In panel **a**), heritability enrichment is defined as a ratio of the proportion of heritability explained by the SNPs in query to the proportion of the SNPs, $\mathrm{Pr}(h_g^2)$ /Pr $(SNPs)$. The blue dashed line represents the median value across traits. In panel **b**), annotation effect size (τ) is used to assess the contribution of all the significant cis-sQTL (or cis-eQTL) SNPs to heritability when fitted jointly with all the significant cis-eQTL (or cis-sQTL) SNPs. Each error bar represents the 95% confidence interval of an estimate.



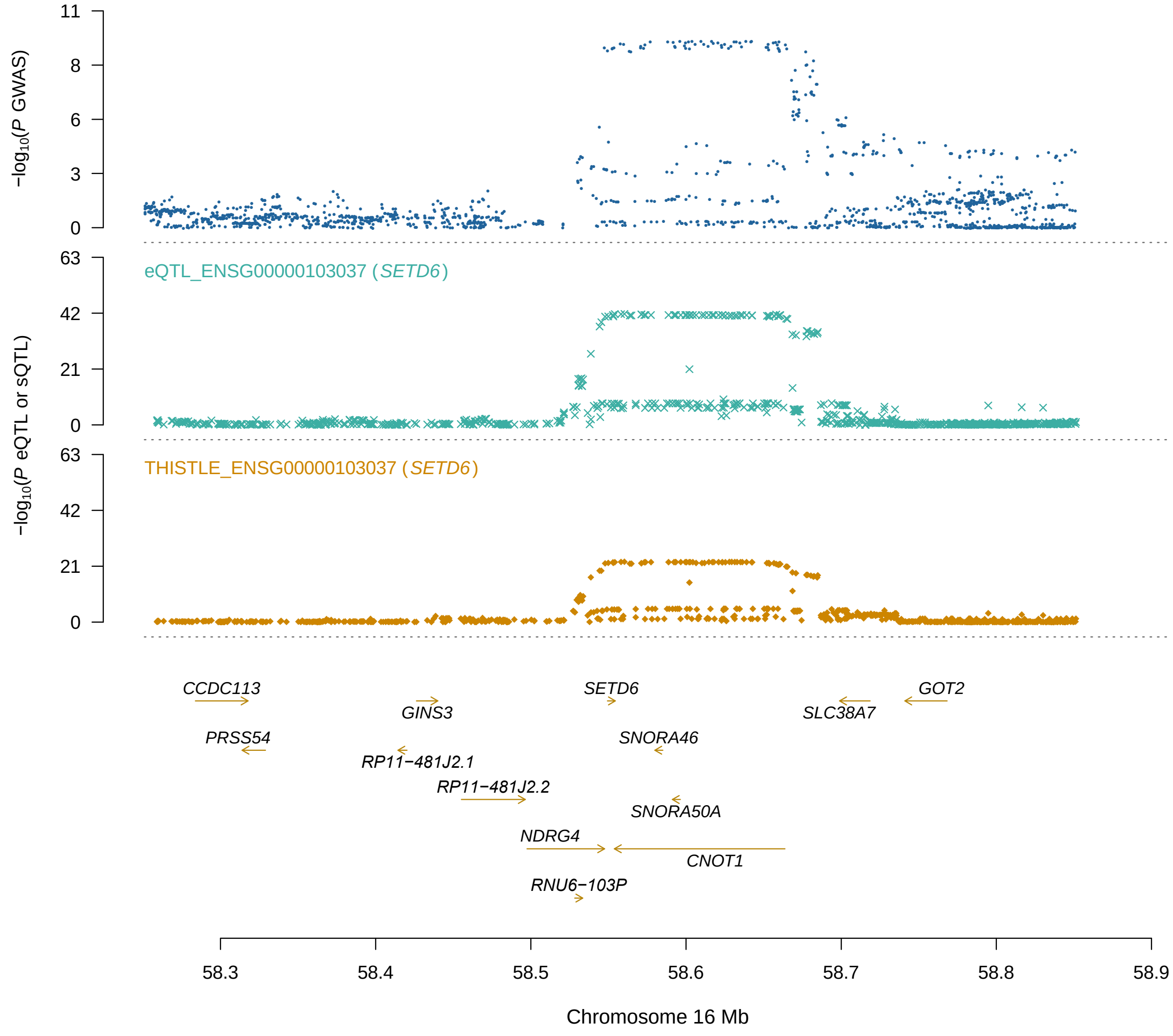
Supplementary Figure 15. Enrichment of all the clumped cis-sQTL or cis-eQTL SNPs for heritability of the ten brain-related traits. We performed clumping analysis for cis-sQTL (or cis-eQTL) SNPs using an LD r^2 threshold of 0.10, a window size of 2 Mb, and p-value thresholds of 5×10^{-8} and 1×10^{-3}. In panel **a**), heritability enrichment is defined as a ratio of the proportion of heritability explained by the SNPs in query to the proportion of the SNPs, $\Pr(h_g^2)\,/\Pr\,(SNPs)$. The blue dashed line represents the median value across traits. In panel **b**), annotation effect size (τ) is used to assess the contribution of all the clumped cis-sQTL (or cis-eQTL) SNPs to heritability when fitted jointly with all the clumped cis-eQTL (or cis-sQTL) SNPs. Each error bar represents the 95% confidence interval of an estimate.



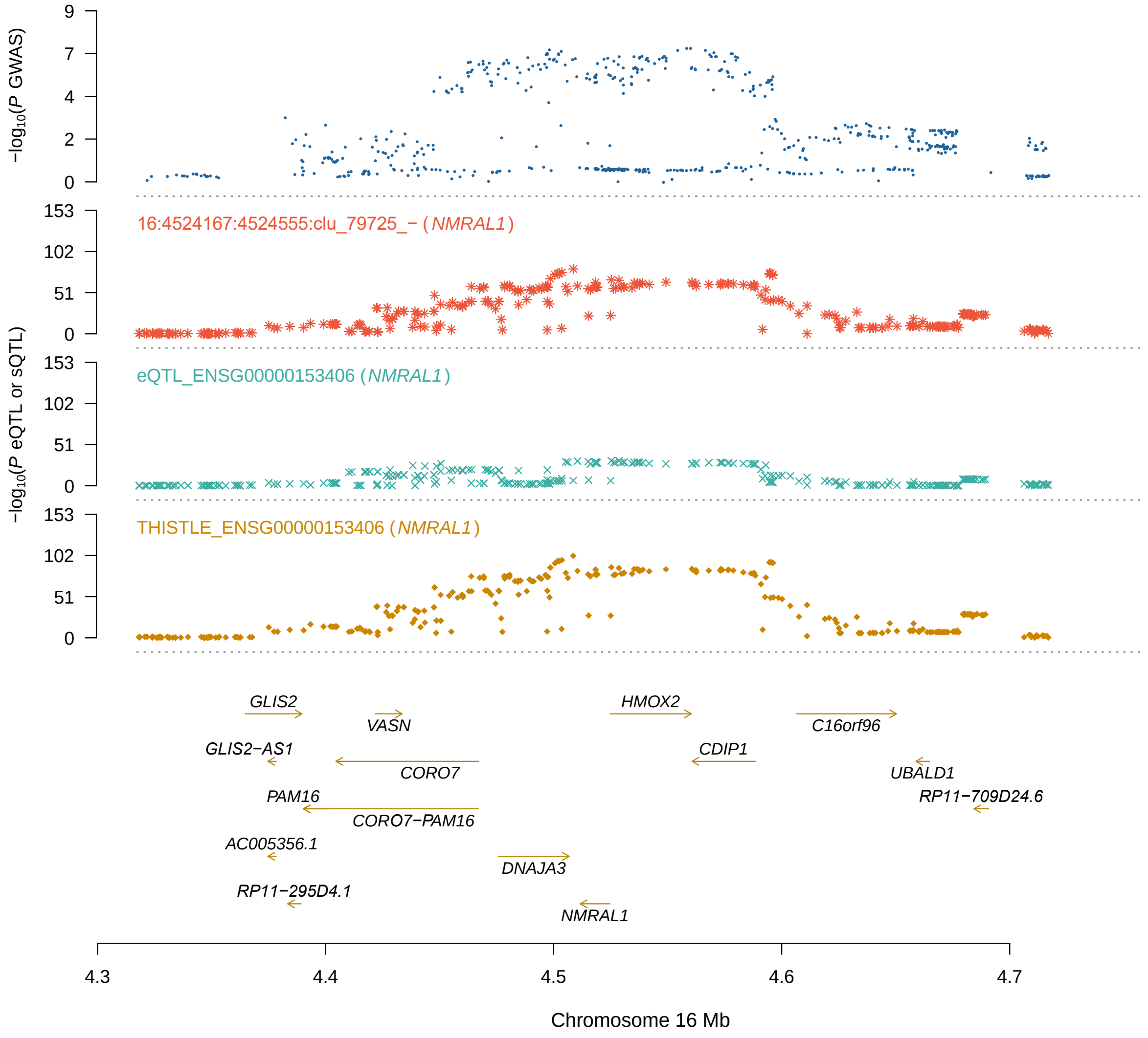
Supplementary Figure 16. Enrichment of all the fine-mapped cis-sQTL or cis-eQTL SNPs for heritability of the ten brain-related traits. We fine-mapped each cis-sQTL (or cis-eQTL) region and computed the causal posterior probability (CPP) of each SNP in the region using SuSiE[3]. We assigned the maximum CCP across all genes (or introns) to an SNP as its annotation value and a zero value to an SNP that does not belong to any 95% credible set. In panel **a**), heritability enrichment is defined as a ratio of the proportion of heritability explained by the SNPs in query to the proportion of the SNPs, $\Pr\left(h_g^2\right)/\Pr\,(SNPs)$. The blue dashed line represents the median value across traits. In panel **b**), annotation effect size (τ) is used to assess the contribution of all the fine-mapped cis-sQTL (or cis-eQTL) SNPs to heritability when fitted jointly with all the fine-mapped cis-eQTL (or cis-sQTL) SNPs. Each error bar represents the 95% confidence interval of an estimate.



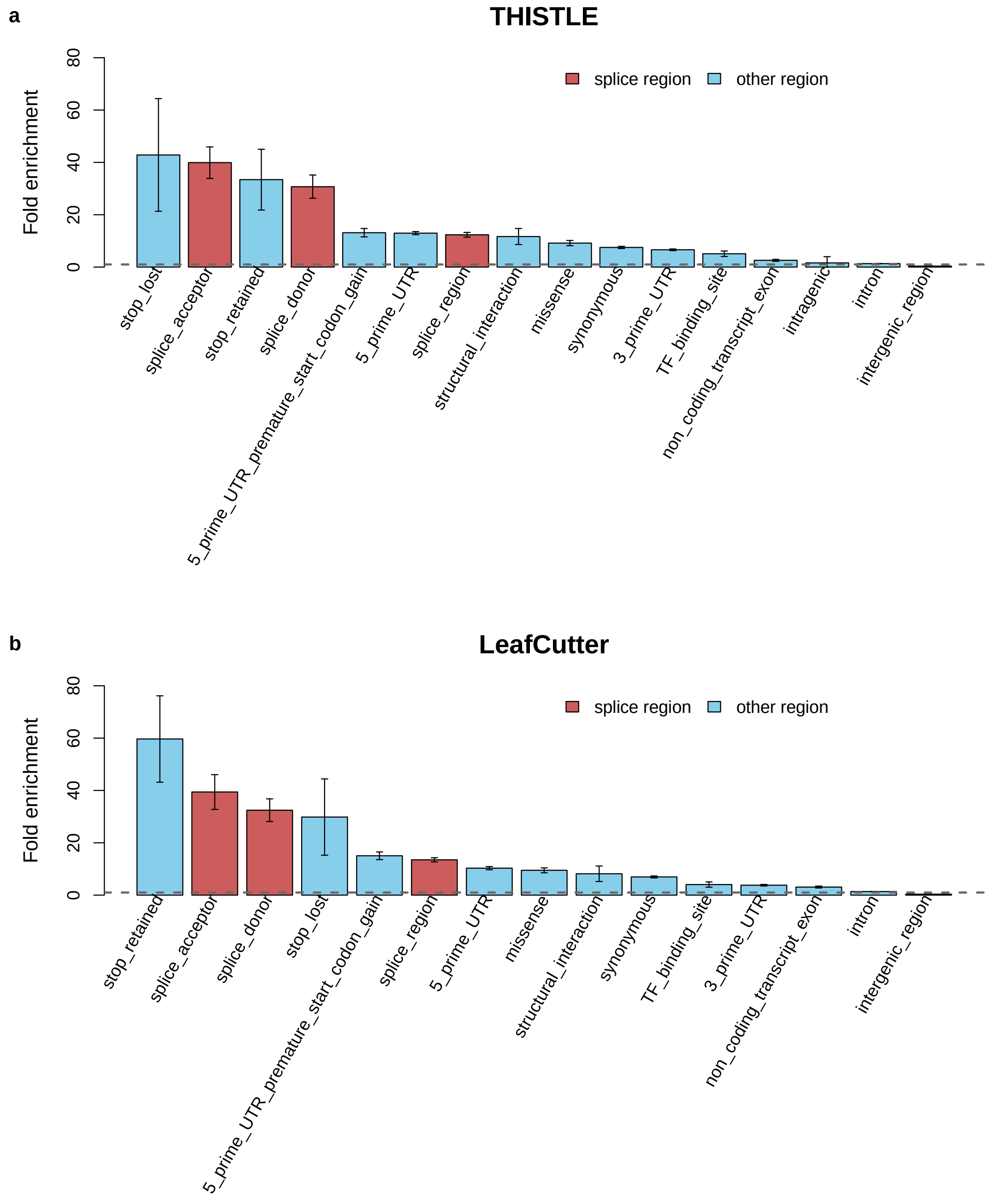
Supplementary Figure 17. Association of *MPHOSPH9* with schizophrenia (SCZ). The association of *MPHOSPH9* with SCZ was identified through the sQTLs. The top sQTL, rs1727307, of an intronic excision event 12:123707674:123710835:clu_71282_- is located in an intron retention region. **a**) The GWAS, sQTL, 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 for *MPHOSPH9*, THISTLE sQTL analysis for *MPHOSPH9*, and LeafCutter sQTL analysis for intron 12:123707674:123710835:clu_71282_- of *MPHOSPH9*, respectively. **b**) Association of rs1727307 with the overall mRNA abundance of *MPHOSPH9*. Each dot represents mRNA abundance of an individual. **c**) Association of rs1727307 with intron excision ratio of 12:123707674:123710835:clu_71282_-. Each dot represents intron excision ratio of an individual.



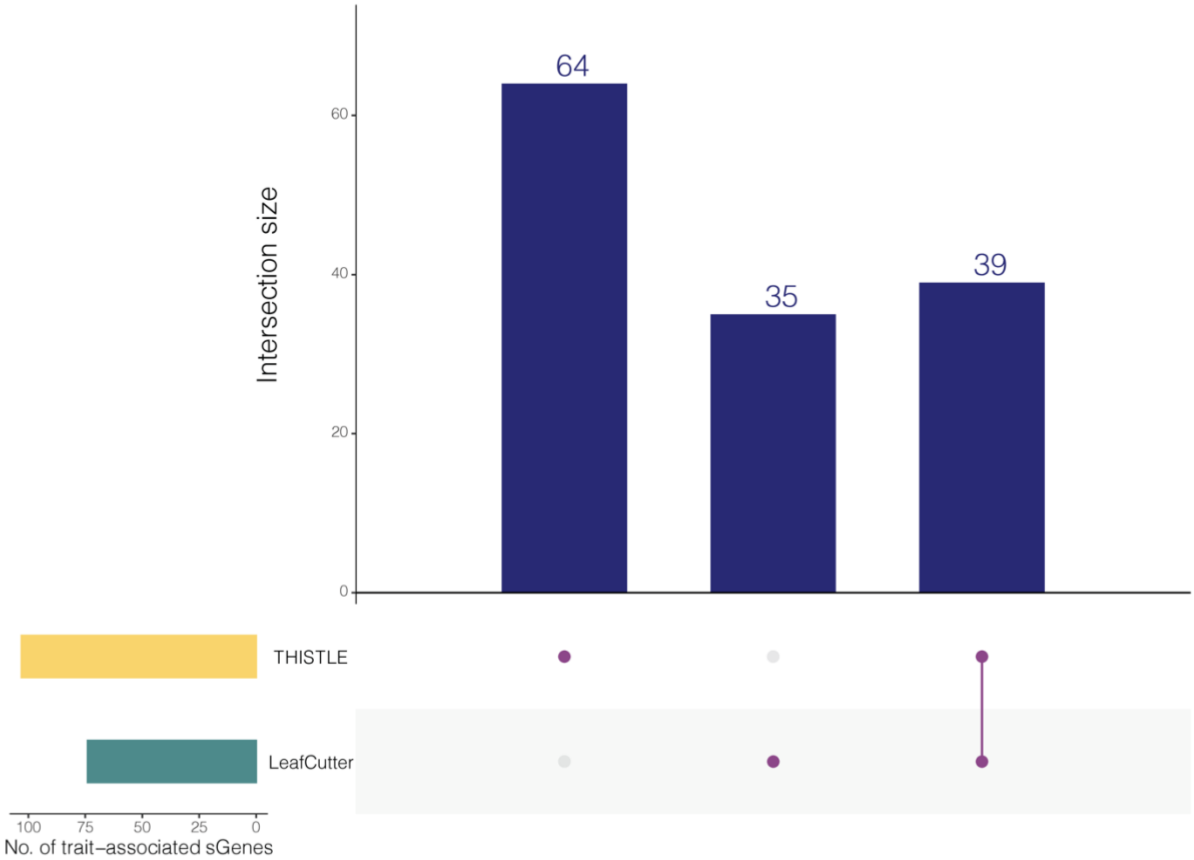
Supplementary Figure 18. Association of *SETD6* with SCZ identified through both sQTLs and eQTLs. The top plot shows −log10(p-values) of SNPs from the GWAS meta-analysis for SCZ. The second and third plots show −log10(p-values) from the eQTL analysis and THISTLE sQTL analysis, respectively.



Supplementary Figure 19. Association of *NMRAL1* with SCZ through two distinct genetic regulatory mechanisms. 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 LeafCutter sQTL analysis for intron 16:4524167:4524555:clu_79725_- of *NMRAL1*, eQTL analysis for *NMRAL1*, and THISTLE sQTL analysis for *NMRAL1*, respectively.



Supplementary Figure 20. Enrichment of the top cis-sQTL SNPs identified from **a**) THISTLE and **b**) LeafCutter in functional categories. The genetic variants are annotated by SnpEff. Fold enrichment is computed by comparing the top cis-sQTL 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.



Supplementary Figure 21. Overlap between the trait-associated sGenes identified using the THISTLE and LeafCutter sQTL summary data.

Supplementary Tables

Supplementary Table 1 GWAS summary data

Phenotype	n	n_{case}	n_{control}	No. of SNPs
autism spectrum disorder (ASD)	46,350	18,381	27,969	8,487,894
bipolar disorder (BIP)	51,710	20,352	31,358	13,413,244
schizophrenia (SCZ)	105,318	40,675	64,643	5,426,250
intelligence (IQ)	269,867	/	/	9,295,118
insomnia	386,533	/	/	10,862,567
Alzheimer's disease (AD)	387,000	71,880	315,120	13,283,327
Parkinson's disease (PD)	482,730	33,674	449,056	17,481,233
major depression (MD)	500,199	170,756	329,443	8,483,302
smoking initiation (SmkInt)	632,802	311,629	321,173	11,802,366
educational attainment (EA)	766,345	/	/	10,101,242

n: sample size; n_{case}: number of cases; n_{control}: number of controls.

Supplementary Table 2 Number of genes associated with the ten brain-related phenotypes

Trait	LeafCutter			THISTLE			eQTL		
	No. of tested introns (genes)	**No. of sig. introns (genes)**	**No. of coloc genes**	**No. of tested genes**	**No. of sig. genes**	**No. of coloc genes**	**No. of tested genes**	**No. of sig. genes**	**No. of coloc genes**
IQ	18,444 (5561)	80 (43)	35 (22)	4823	65	24	9407	89	27
EA	18,547 (5585)	189 (84)	85 (37)	4838	107	47	9432	163	64
SmkInt	18,445 (5564)	28 (8)	5 (3)	4827	17	11	9406	23	15
SCZ	16,869 (5149)	67 (34)	30 (16)	4486	47	21	8718	60	31
AD	18,503 (5574)	6 (3)	0 (0)	4830	9	2	9424	11	3
ASD	18,341 (5531)	0 (0)	0 (0)	4800	6	0	9346	0	0
insomnia	18,384 (5548)	5 (2)	0 (0)	4813	1	1	9383	3	1
BIP	18,551 (5588)	7 (4)	6 (4)	4837	5	4	9433	11	7
MD	18,445 (5554)	13 (4)	3 (1)	4813	8	6	9399	10	6
PD	18,561 (5589)	42 (14)	10 (3)	4838	18	0	9437	33	6
Total	/	337 (147)	164 (74)	/	220	103	/	317	149

Genes (introns) associated with the brain-related traits were identified by an analysis (SMR + COLOC) that integrates the LeafCutter sQTLs, THISTLE sQTLs, and eQTLs into GWAS. The ten brain-related traits are intelligence (IQ), educational attainment (EA), smoking initiation (SmkInt), schizophrenia (SCZ), Alzheimer's disease (AD), autism spectrum disorder (ASD), insomnia, bipolar disorder (BIP), major depression (MD), and Parkinson's disease (PD). **No. of tested introns** or **No. of tested genes**: number of introns or genes included in the SMR analysis. **No. of sig. introns (genes)**: number of introns associated with a phenotype at a genome-wide significance level in the SMR analysis with the number of unique genes shown in the parentheses. **No. of sig. genes**: number of genes associated with a phenotype at a genome-wide significance level in the SMR analysis. **No. of coloc genes**: number of genes passed the SMR test and showed a COLOC PP4 value of > 0.8.

Supplementary Note

Simulated data

We calibrated THISTLE and compared it with sQTLseekeR using a set of simulated data. We simulated genotype data of 1,000 unlinked SNPs in 500 individuals using a binomial distribution, with minor allele frequencies (MAFs) of the SNPs ranging from 0.01 to 0.5. Assuming a gene with three transcript isoforms, we randomly sampled a SNP as the causal variant with its effect on three isoforms denoted by $\mathbf{b} = \{b_1, b_2, b_3\}$. We generated $\mathbf{b}$ under four different scenarios: 1) $\mathbf{b} = \{0, 0, 0\}$, where the causal variant is neither an sQTL nor an eQTL, 2) $\mathbf{b} = \{2, 2, 2\}$, where the causal variant is an eQTL but not an sQTL, 3) $\mathbf{b} = \{0, 2, -2\}$, where the causal variant is an sQTL but not an eQTL, and 4) $\mathbf{b} = \{1, 2, 3\}$, where the causal variant is both an sQTL and an eQTL. We generated the transcript abundance as $y_{ij} = x_j b_i + e_{ij}$ where y_{ij} is the transcript abundance of isoform i in individual j, x_j is the genotype of the causal variant of individual j, b_i is the effect size of the causal variant on isoform i, and e_{ij} is the residual with its variance denoted by $var(e_j)$. Considering that in reality RNA-seq read counts usually follow a Poisson distribution and that expression levels of different isoforms of a gene are often correlated, we generated residuals of the three isoforms of each individual (denoted by $\mathbf{e}_j = \{e_{1j}, e_{2j}, e_{3j}\}$) from a multivariate Poisson distribution with mean $\mathbf{0}$ and variance-covariance matrix $\mathbf{S}$. The kl-th element of $\mathbf{S}$ is $S_{kl} = r_e\sqrt{var(e_{kj})var(e_{lj})}$, where r_e is the residual correlation and $var(e_{kj}) = 2p(1-p)b_k^2(\frac{1}{q_k^2} - 1)$ with p being the MAF of the causal variant and q_k^2 being the proportion of variance in transcript abundance of isoform k explained by the causal variant (which was set 10% for all the three isoforms). We also generated residuals from a multivariate normal distribution for comparison, i.e., $\mathbf{e}_j \sim MVN(\mathbf{0}, \mathbf{S})$. We repeated each simulation scenario with 500 replicates.

sQTL analysis using sQTLseekeR

We also compared THISTLE with sQTLseekeR[4] in the analysis of the PsychENCODE data[5,6]. Following the analysis pipeline provided by the authors of sQTLseekeR[4], we included in the analysis genes with splicing dispersion > 0.01 and more than 25 splicing patterns. For each gene, only the individuals with TPM > 0.1 were included in the sQTL test. To be consistent with the THISTLE analysis, we included in the sQTLseekeR analysis only the SNPs located in the gene in query or within 2 Mb upstream or downstream of the gene. In total, 13,361 genes and 1,341,182 SNPs were included in the sQTLseekeR analysis, and a false discover rate (FDR) of 0.01 was used to correct for multiple testing.

Principal component analysis

To identify individuals of European ancestry, we performed a principal component analysis (PCA) in a combined genotype data set of the PsychENCODE (*n* = 1,658) and the 1000 Genomes Project[7] (1000GP; *n* = 2,504). The 1000GP cohort comprises whole-genome sequence data from individuals of European (EUR), East Asian (EAS), Admixed American (AMR), South Asian (SAS), and African (AFR) ancestries. Only the autosomal SNPs with missingness rate <5% and MAF >1% and individuals with missingness rate <5% (593,365 SNPs in common with HapMap3[8] on 4,162 individuals in total) were included in the PCA. After the PCA, we removed the PsychENCODE individuals whose principal component 1 (PC_1) or PC_3 deviated more than 6 standard deviations from the mean of the corresponding PC of the 1000GP individuals of European ancestry. Finally, a total of 1,073 individuals of European ancestry were retained for further analysis.

Sampling variance of the estimated fold enrichment

Let x represent the estimated per-SNP heritability for the SNPs in query, and $\boldsymbol{y} = \{y_1, y_2, \dots, y_j, \dots, y_m\}$ with y_j being the corresponding estimate for the control SNPs in the j_{th} replicate (noted that the control SNPs are randomly sampled with MAF and genomic location matched with the SNPs in query). The fold enrichment is calculated as $x/\bar{y}$, where $\bar{y}$ is the mean across all the elements of $\boldsymbol{y}$. The variance of $x/\bar{y}$ can be computed approximately by the Delta method[9],

$$var\left(\frac{x}{\bar{y}}\right) \approx \left(\frac{x}{\bar{y}}\right)^2 [\frac{var(x)}{x^2} + \frac{var(\bar{y})}{\bar{y}^2} - \frac{2cov(\mathrm{x}, \bar{y})}{x\bar{y}}]$$

If we assume the covariance between x and $\bar{y}$ is 0 and the variance of x is $var(x) \approx \widehat{var}(y)$, where $\widehat{var}(y)$ is the observed variance of y across m replicates, the variance of fold enrichment can be approximated by

$$var\left(\frac{x}{\bar{y}}\right) \approx \left(\frac{x}{\bar{y}}\right)^2 [\frac{var(x)}{x^2} + \frac{var(\bar{y})}{\bar{y}^2}] \approx \left(\frac{x}{\bar{y}}\right)^2 [\frac{\widehat{var}(y)}{x^2} + \frac{\widehat{var}(y)}{m\bar{y}^2}]$$

Acknowledgments

Data were generated as part of the PsychENCODE Consortium supported by: U01MH103339, U01MH103365, U01MH103392, U01MH103340, U01MH103346, R01MH105472, R01MH094714, R01MH105898, R21MH102791, R21MH105881, R21MH103877, and P50MH106934 awarded to: Schahram Akbarian (Icahn School of Medicine at Mount Sinai), Gregory Crawford (Duke), Stella Dracheva (Icahn School of Medicine at Mount Sinai), Peggy Farnham (USC), Mark Gerstein (Yale), Daniel Geschwind (UCLA), Thomas M. Hyde (LIBD), Andrew Jaffe (LIBD), James A. Knowles (USC), Chunyu Liu (UIC), Dalila Pinto (Icahn School of Medicine at Mount Sinai), Nenad Sestan (Yale), Pamela Sklar (Icahn School of Medicine at Mount

Sinai), Matthew State (UCSF), Patrick Sullivan (UNC), Flora Vaccarino (Yale), Sherman Weissman (Yale), Kevin White (UChicago) and Peter Zandi (JHU).

Supplementary references

1 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56-65 (2012).

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

3 Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **82**, 1273-1300 (2020).

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

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

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

7 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56 (2012).

8 Consortium, I. H. Integrating common and rare genetic variation in diverse human populations. *Nature* **467**, 52 (2010).

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