Phenomics (2023) 3:217–227
https://doi.org/10.1007/s43657-023-00100-6

BRIEF COMMUNICATION

Hidden Genetic Regulation of Human Complex Traits via Brain Isoforms

Lu Pan[1,2] · Chenqing Zheng[1] · Zhijian Yang[1] · Yudi Pawitan[2] · Trung Nghia Vu[2] · Xia Shen[1,2,3,4,5]

Received: 17 August 2022 / Revised: 15 February 2023 / Accepted: 17 February 2023 / Published online: 20 March 2023

Abstract

Alternative splicing exists in most multi-exonic genes, and exploring these complex alternative splicing events and their resultant isoform expressions is essential. However, it has become conventional that RNA sequencing results have often been summarized into gene-level expression counts mainly due to the multiple ambiguous mapping of reads at highly similar regions. Transcript-level quantification and interpretation are often overlooked, and biological interpretations are often deduced based on combined transcript information at the gene level. Here, for the most variable tissue of alternative splicing, the brain, we estimate isoform expressions in 1,191 samples collected by the Genotype-Tissue Expression (GTEx) Consortium using a powerful method that we previously developed. We perform genome-wide association scans on the isoform ratios per gene and identify isoform-ratio quantitative trait loci (irQTL), which could not be detected by studying gene-level expressions alone. By analyzing the genetic architecture of the irQTL, we show that isoform ratios regulate educational attainment via multiple tissues including the frontal cortex (BA9), cortex, cervical spinal cord, and hippocampus. These tissues are also associated with different neuro-related traits, including Alzheimer's or dementia, mood swings, sleep duration, alcohol intake, intelligence, anxiety or depression, etc. Mendelian randomization (MR) analysis revealed 1,139 pairs of isoforms and neuro-related traits with plausible causal relationships, showing much stronger causal effects than on general diseases measured in the UK Biobank (UKB). Our results highlight essential transcript-level biomarkers in the human brain for neuro-related complex traits and diseases, which could be missed by merely investigating overall gene expressions.

Keywords Alternative splicing · Isoform-ratio quantitative trait loci (irQTL) · Expression quantitative trait loci (eQTL) · Genome-wide Association Studies · Neuro-related human complex traits

Abbreviations

cis-irQTL	Cis-isoform ratio quantitative trait loci
CRP	Cluster response profile
eQTL	Expression quantitative trait loci
FDR	False discovery rate
GENCODE	The GENCODE project for the identification and classification of all gene features in the human and mouse genomes with high accuracy
GTEx	Genotype-tissue expression
GWAS	Genome-wide association studies
ICD	International classification of diseases
IVW	Inverse-variance weighted
irQTL	Isoform-ratio quantitative trait loci
LD	Linkage disequilibrium
MAF	Minor allele frequency
MMAB	Metabolism Of Cobalamin Associated B
mQTL	Metabolites quantitative trait loci
MR	Mendelian randomization
PCs	Principal components
PCA	Principal component analysis

Lu Pan, Chenqing Zheng, and Zhijian Yang contributed equally.

✉ Xia Shen
shenx@fudan.edu.cn

1 Biostatistics Group, School of Life Sciences, Sun Yat-Sen University, Guangzhou 510006, China

2 Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm 17177, Sweden

3 State Key Laboratory of Genetic Engineering, School of Life Sciences, Fudan University, Shanghai 200433, China

4 Center for Intelligent Medicine Research, Greater Bay Area Institute of Precision Medicine (Guangzhou), Fudan University, Guangzhou 511458, China

5 Centre for Global Health Research, Usher Institute, University of Edinburgh, Edinburgh EH8 9AG, UK

PLINK	PLINK is a whole genome association analysis toolset
PSI	Percent spliced index
pQTL	Protein quantitative trait loci
S-LDSC	Stratified linkage disequilibrium score regression
sQTL	Splicing quantitative trait loci
THISTLE	Testing for heterogeneity between isoform-eQTL effects
TPM	Transcripts per million
XAEM	X-matrix alternating expectation–maximization
UKB	UK Biobank

Introduction

Alternative splicing is an essential mechanism in diversifying the genetic and proteomic landscapes in eukaryotes (Thakur et al. 2019). It is a crucial process by which proteins with different functions are produced from a single gene through patterned joining and excising of the introns and exons within the gene (Nature Portfolio 2022). Alternative splicing occurs in around 95% of multi-exonic genes in humans (Nilsen and Graveley 2010; Pan et al. 2008), and is subjected to tissue-specific regulations (Zaghlool et al. 2014). It is a process that is most complex in the nervous system (Yeo et al. 2004). In the central nervous system, our brain has the highest level of alternative splicing out of all tissues in the human body (Porter et al. 2018; Xu et al. 2002).

Genetic regions that control and regulate these alternative splicing events, often known as splicing quantitative trait loci (sQTL), have been successively discovered in large-scale genome-wide association studies (GWAS) (Ardlie et al. 2015; Battle et al. 2014; Park et al. 2018), and recent studies have shown sQTL landscapes in human brain sub-tissues (Takata et al. 2017; Walker et al. 2019; Zhang et al. 2020). These studies used percent spliced index (PSI, ratios between exon-included and exon-excluded reads) values (Schafer et al. 2015) calculated directly from RNA-sequencing data to obtain splicing scores representing splicing patterns in genes prior to sQTL discoveries. Expressed isoform ratios per gene were also considered as phenotypes in mapping sQTL (Lappalainen et al. 2013). However, directly mapping isoform level quantitative trait loci (QTL) has always been challenging due to the difficulty in quantifying isoform expressions using short RNA-sequencing reads.

Here, we aim to discover the QTL regulating expressed isoform ratios per gene (i.e., irQTL) across brain tissues, where the isoform expressions were quantified using our previously developed isoform expression estimation method, X-matrix alternating expectation–maximization (XAEM) (Deng et al. 2019), which outperforms the other state-of-the-art methods in isoform estimation. We quantify isoform ratios in multi-isoform genes in 1,191 samples from 13 brain tissues and carry out GWAS analysis of the isoform ratios with the genotyping data to acquire cis-irQTL. These irQTL regulate the relative proportions across the isoforms per gene instead of the overall gene expression. We show that genes with such irQTL genetic basis in the brain contribute significantly to neuro-related phenotypes.

Materials and Methods

Samples Origin and Data Acquired

RNA Sequencing and Whole-Genome Sequencing Data

RNA sequencing data used in this study were obtained from the genotype-tissue expression (GTEx) project (Ardlie et al. 2015) portal (https://www.gtexportal.org, version phs000424.v7.p2.c1). We considered 13 brain tissues, consisting of 1,236 RNA sequencing samples from 172 individuals from the GTEx project (Fig. 2a). As an individual might die from different causes, the tissue(s) sampled from the individual was from diseased-free sampling sites. GTEx donors are aged between 21 and 70 with the following criteria exclusion criteria: individuals with human immunodeficiency virus (HIV) infection or high-risk behaviors, viral hepatitis, metastatic cancer, chemotherapy or radiation therapy for any condition within the past two years, and whole-blood transfusion in the past 48 h or body mass index > 35 or < 18.5 (Ardlie et al. 2015).

Whole-genome sequencing (WGS) data for these individuals were obtained from the GTEx portal under version phs000424.v7.p1. There are 6,496,708 markers called from WGS, including 5,987,177 SNPs and 509,531 InDels. The final sample size used for QTL analysis is 1,191 samples, for which both RNA-sequencing and whole-genome sequencing data are available.

GTEx cis-eQTL Data

Cis-eQTL are genomic loci near the corresponding coding genes that explain a fraction of the genetic variance of the gene expression phenotypes (Glass et al. 2013). GTEx has summarized a list of cis-eQTL for each tissue type in its data portal. In this study, only cis-eQTL from the brain tissues were considered.

RNA-Sequencing Mapping and Quantification

We acquired demultiplexed raw RNA sequencing FASTQ files from the GTEx portal and used the fast mapping and

isoform quantification tool XAEM (Deng et al. 2019), which has higher accuracy than popular methods such as Salmon (Patro et al. 2017) and Kallisto (Bray et al. 2016), to process the data for RNA sequencing alignment and isoform expression quantification. Mapping was performed using XAEM V0.1.0 with reference to the human reference genome hg19/GRCh37 (UCSC hg19 annotation). Isoform quantification was done subsequently using the default setting in XAEM to produce isoform counts, normalized to transcripts per million (TPM) values, for each sub-brain tissue. Mapping and quantification were carried out separately for each brain tissue.

Quality Control Measures

Quality control was carried out for isoform expression data after quantification using XAEM software. We considered in total 24,629 genes with 46,710 isoforms based on the human reference genome hg19/GRCh37. To identify the genetic regulation of isoform expression missed by eQTL analysis, we considered only multi-isoform genes, which led to a total of 9,401 genes with 31,482 isoforms after filtering. Of these 9,401 genes, 8,382 of them are protein-coding genes, and the rest includes lncRNA, antisense-RNA, pseudo-genes, etc. (Supplementary Table 3). Individuals with half or more than half of their genes having zero counts were removed for subsequent analyses, which reduced the original 1,671 viable samples in GTEx to the 1,236 retained samples we started with. For the principal component analysis (PCA) and QTL mapping, only variants with minor allele frequency (MAF) > 0.05 were considered, which resulted in 6,164,423–6,316,616 variants across the brain tissues.

irQTL Mapping

In this study, we identify irQTL for 13 brain tissues. For genes with multiple isoforms, isoform ratios were defined as TPM values of isoforms divided by their respective gene-level TPM. PCA was carried out on the genomic kinship matrix constructed via the whole-genome sequencing genotype data using PLINK (Purcell et al. 2007), to obtain a set of genomic principal components (PCs) to be used as covariates in the subsequent association scan. Age, sex, and the first three PCs were then used as covariates, whose effects were taken away from the isoform ratios using linear regression. The resulting values were used as covariates-corrected isoform ratios for downstream analyses. We performed cis-regulatory region association analysis by regressing the isoform ratio phenotypes on the genotype data using RegScan V0.5, a GWAS analysis tool for linear regression analysis with continuous traits maximally fast on large data sets with many phenotypes (Planell et al. 2021).

Locus Definition

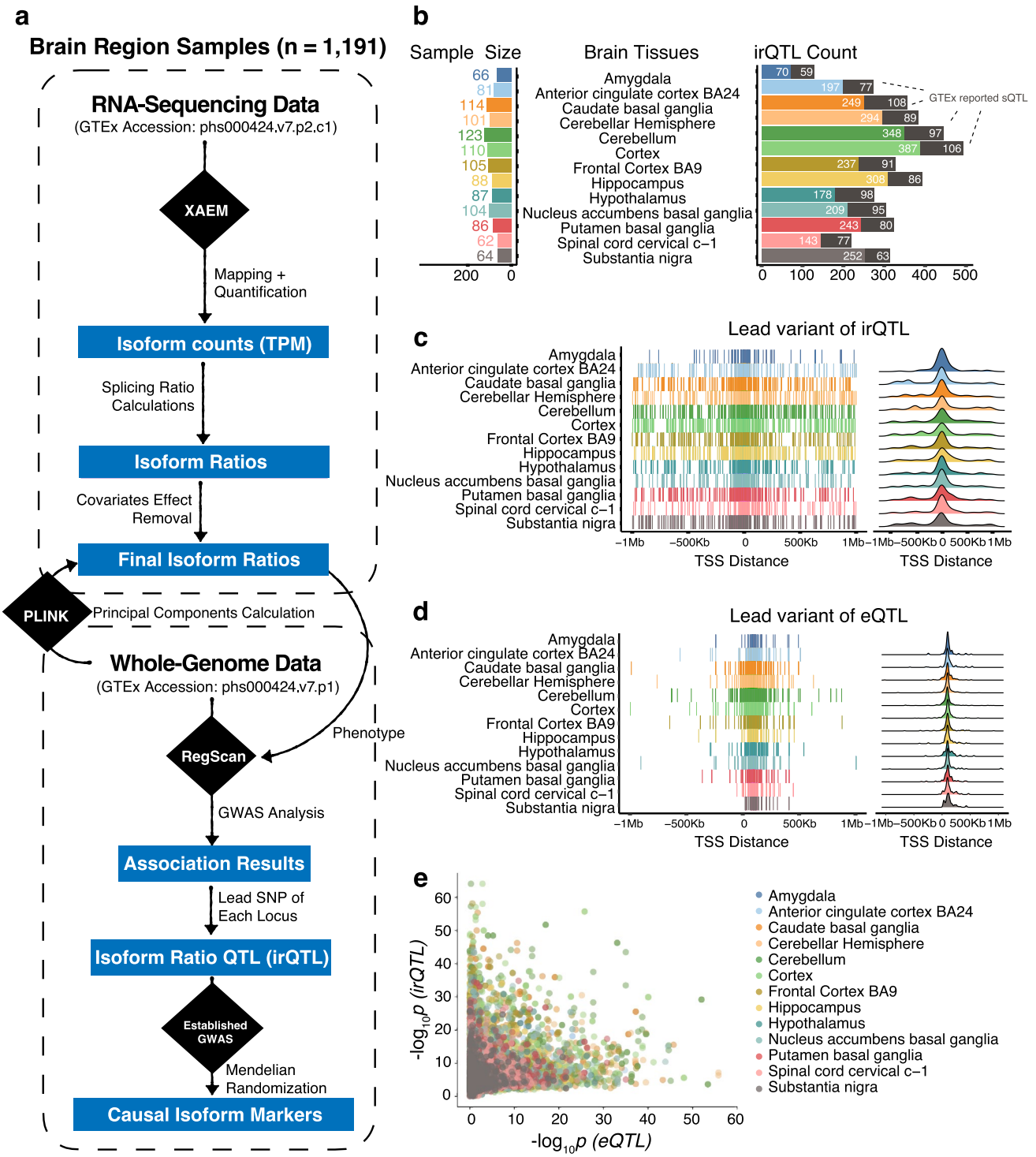
Each cis-regulatory locus, as well as the irQTL region, was defined as the ± 1 Mb region around the corresponding gene. The SNP with the lowest p-value within each locus was selected as the lead variant, and the associations having $p < 5 \times 10^{-8}$ were retained for the subsequent analyses. These significant irQTL from each brain tissue were then compared with the cis-eQTL by the GTEx Consortium of the corresponding gene. We focused on the irQTL with eQTL p-values greater than 0.05 as the final set of irQTL.

Stratified Linkage Disequilibrium Score Regression (S-LDSC)

We used S-LDSC (Bulik-Sullivan et al. 2015) to test whether the annotated genic regions are enriched for heritability of a certain trait, where the GWAS summary statistics were available via linkage disequilibrium (LD) hub (LD-Hub) (Zheng et al. 2016). The summary statistics were harmonized by the munge_sumstats.py procedure of the LDSC software. LD scores of HapMap3 SNPs (Altshuler et al. 2010) (major histocompatibility complex region excluded) for the annotated genes in each brain tissue were pre-computed using a 1-cM window (default). The heritability enrichment in each tissue was evaluated by an enrichment score, defined as the proportion of heritability captured divided by the proportion of annotated SNPs. The LDSC-v1.2 baseline annotations were fitted as covariates as LDSC suggested (Bulik-Sullivan et al. 2015; Gazal et al. 2017), which controls the residual variance in the chi-squared statistics and thus produces a more robust estimation of heritability enrichment on our desired annotation. For each tissue, we ran a separate model to test the heritability enrichment at the tissue-specific irQTL. This fits our hypothesis testing purpose, meanwhile avoiding potential multi-collinearity due to the similarity across brain tissues.

Mendelian Randomization (MR) Analysis

Prior to the analysis, we extracted 152 neuro-related traits from LD-Hub and 200 UK Biobank (UKB) diseases with international classification of diseases (ICD) codings from Neale's lab GWAS results. From the original 152 neuro-related traits, we removed the duplicated traits from different sources and highly correlated traits e.g., defined by alternative phenotype codings. This resulted in 114 neuro-related phenotypes for subsequent analysis. We



conducted an MR analysis between the isoform ratios and 114 neuro-related traits and the 200 UKB diseases using the standard inverse-variance weighted (IVW) method for all the cis-irQTL and the MR Egger regression (Bowden et al. 2017) for the cis-irQTL with at least three independent instruments after LD-clumping ($r^2 < 0.001$). Here, the cis-irQTL were used as genetic instruments, and the coding alleles were matched between the exposure isoforms and the outcome phenotypes before estimating potential causal effects.

◂**Fig. 1** irQTL workflow and summary statistics. **a** RNA sequencing data of 1,191 samples from 13 brain regions were obtained from the GTEx Consortium. Alignment and isoform quantification were analyzed using the XAEM software based on TPM for each sample. For multi-isoform genes, the isoform ratio for each isoform was calculated as the TPM value for each isoform divided by the overall corresponding gene expression TPM value. PLINK was used for calculating the genomic kinship matrix and three PCs. Fixed effects of age, sex, and the first three PCs were removed from the isoform ratios phenotypes. The phenotypes were inverse-Gaussian transformed prior to GWAS analysis using RegScan. The lead variants of the mapped irQTL were passed onto subsequent analysis, including causal inference referencing the PhenoScanncer database. **b** The sample size in irQTL mapping and the corresponding detected irQTL count for each brain tissue, where the number that overlaps with GTEx-reported sQTL ($p < 5 \times 10^{-8}$ is marked. **c** Lead variants locations of the mapped irQTL ($p < 5 \times 10^{-8}$) with respect to their distance to the transcription start sites (TSS). **d** Lead variants locations of the mapped eQTL ($p < 5 \times 10^{-8}$) with respect to their distance to the transcription start sites (TSS). **e** $-\log_{10}p$ values comparison between irQTL and eQTL, where the 4,241 irQTL signals ($p_{irQTL} < 5 \times 10^{-8}$, and $p_{eQTL} > 0.05$) were annotated to test for heritability enrichment

Results

We aim to identify irQTL in 13 sub-brain tissues, where RNA and DNA sequencing data are both available in 1,191 GTEx consortium samples. The overall workflow is illustrated in Fig. 1a. After RNA sequencing reads mapping and quantification, isoform counts were estimated using XAEM for each sub-brain tissue. For the multi-isoform genes, the isoform ratios were defined as the isoform TPM values divided by their corresponding gene-level TPM values. Cis-regulatory region association analysis was performed on each isoform ratio phenotype, where fixed effects including sex, age, and the first three genomic principal components (PCs) were corrected for the phenotype. The corrected phenotypic values were also used in downstream analyses.

The association analysis identified 7,099 cis-irQTL in the brain ($p < 5 \times 10^{-8}$, equivalent to an estimated false discovery rate (FDR) of 9.6×10^{-5} to 5.1×10^{-4} across different tissues), where 4,241 of those did not show any effect as gene expression quantitative trait loci (eQTL) on general gene expression levels ($p_{eQTL} > 0.05$) (Fig. 1b,e). For these irQTL, the corresponding genes have relatively stable expression levels in the brain tissues, but the proportions of their isoforms are genetically regulated by the irQTL. Taking the frontal cortex as an example, 493 irQTL were detected for 265 genes whose overall expressions do not show eQTL effects (Fig. 1b, Supplementary Table 1). We cross-referenced these detected irQTL in the latest sQTL results using sQTLseekeR (Monlong et al. 2014) by the GTEx Consortium. Based on the same significance threshold ($p < 5 \times 10^{-8}$), GTEx reported 1,126 irQTL out of the total 4,241 irQTL as sQTL (Fig. 1b, Supplementary Table 3). We also cross-referenced the detected irQTL in the sQTL reported by the recent testing for heterogeneity between isoform-eQTL effects (THISTLE) method. As THISTLE utilizes a heterogeneity test per gene for sQTL discovery, we compared the sGene (genes with significant splicing QTL) counts between the irQTL and THISTLE sQTL results. Overall, the discovered irQTL mapped to 874 sGenes, where 96 overlap with THISTLE sGenes.

The majority of the detected irQTL lead variants are centered at the transcription start sites (TSS) (Fig. 1c), which is a feature also seen in eQTL (Fig. 1d) and even protein QTL (pQTL) in the human plasma (Sun et al. 2018). Nevertheless, we found that the lead variants of cis-eQTL were generally more condensed around the TSS of the corresponding genes than those of cis-irQTL. This could be caused by two reasons: (1) different isoforms of the same gene had different regulatory elements for their transcription; (2) the isoform expressions were estimated and thus had lower statistical power compared to the corresponding overall gene expressions given the same sample size, as unlike for gene expression where one could count the sequencing reads for quantification, the shared reads between isoforms provide incomplete information for isoform expression.

In each brain tissue, we annotated the genomic regions for the corresponding irQTL and applied S-LDSC (Bulik-Sullivan et al. 2015) to estimate and test for heritability enrichment of complex traits. We considered the 114 neuro-related traits (Supplementary Table 4) whose GWAS summary statistics are available through LD-Hub (Zheng et al. 2016). Across the 13 (tissues) × 114 (traits) = 1,482 enrichment tests, the distribution of the S-LDSC reported p-values significantly deviated from the null (Supplementary Fig. 5). With FDR of less than 0.05, three brain tissues were found to be significantly associated with 13 neuro-related traits via the genetic regulation of isoform proportions per gene instead of gene expression levels (Fig. 2a), as the genome annotation of the detected irQTL does not carry any nominal eQTL effect. Such heritability enrichment on irQTL genes was also significantly higher than that on the other coding genes (Fig. 2b). The frontal cortex (BA9) was associated with Alzheimer's or dementia, mood swings, nervous feelings, sensitivity or hurt feelings, sleep duration, alcohol intake, and contraceptive pill intake; the cortex was found to be associated with educational attainment, alcohol intake, intelligence, and knee pain; the cervical spinal cord was found to be connected to anxiety or depression.

We subsequently extracted the established genotype–phenotype association records of the corresponding irQTL LD-clumped ($r^2 < 0.001$) significant variants from the same set of 114 neuro-related traits. We conducted inverse-variance weighted (IVW) MR analysis for all the isoform-trait pairs and an MR Egger regression (Bowden et al. 2017) for the irQTL with at least three instrumental variants after LD-clumping (Supplementary Table 5). This procedure revealed

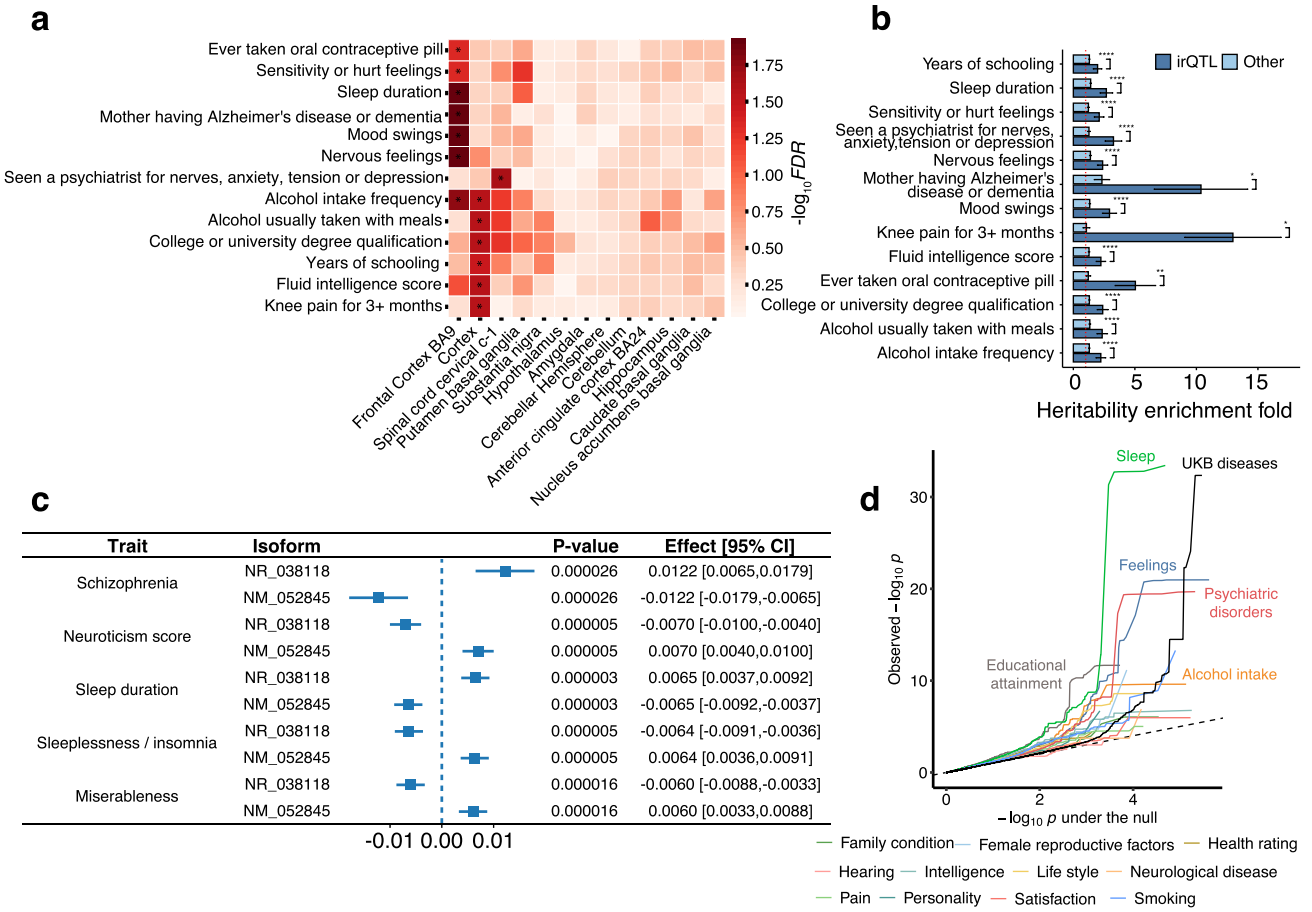


Fig. 2 Genetic effects on complex traits mediated through irQTL. **a** Complex traits heritability enrichment signals at the irQTL regions, where the significant enrichment (FDR < 5%) results were highlighted. **b** Comparison of the heritability enrichment at irQTL genes and other genes for the significant traits in (**a**). **c** Example of the MR results for gene *MMAB*, whose two isoforms have genetically regulated ratios in different tissues, showing plausible causal effects on multiple neuro-related phenotypes. **d** Quantile–quantile plot for the significance of all the MR causal effects tests in 114 neuro-related traits (in categories) and 200 UKB diseases. **e** Genotype–phenotype map of the significant irQTL of *MMAB*, as instruments for the MR discoveries for neuro-related phenotypes in (**c**). In each case, the normalized isoform ratios of the two isoforms of the same gene sum up to similar values across different genotypes; thus the irQTL could not be mapped as significant eQTL

1,092 isoform-trait pairs with plausible causal relationships (FDR less than 0.05, Supplementary Table 5). For comparison, we also conducted the same analysis for 200 UKB diseases with ICD codes whose GWAS summary statistics were publicly available from Neale's lab. This procedure revealed 250 isoform-disease pairs with plausible causal relationships (FDR less than 0.05, Supplementary Fig. 3, Supplementary Table 6). As most of the 200 UKB diseases are not neuro-related, we found, as expected, generally stronger signals of MR discoveries for the neuro-related traits than for the UKB diseases; for instance, the causal inference discoveries were enriched for the traits categories including educational attainment, sleep, psychiatric disorders, feelings, and alcohol intake (Fig. 2d). Taking the gene *MMAB* with the most MR discoveries as an example, its irQTL could be mapped in multiple tissues, leading to the downstream causal inference of its isoforms on multiple neuro-related phenotypes, such as sleep duration, insomnia, neuroticism, miserableness, and schizophrenia (results in the amygdala are illustrated in Fig. 2c). Reverse causal inference analysis did not reveal statistically significant effects of the complex traits on the isoform expressions (Supplementary Fig. 4). Vitamin B12 is involved in the production of sleep-regulating neurotransmitter melatonin (Hashimoto et al. 1996; Mayer et al. 1996), and the protein product of *MMAB* was reported to catalyze the conversion of vitamin B12 into its final product adenosylcobalamin (Safran et al. 2021). For the two isoforms of *MMAB* quantified via the XAEM algorithm, the genotype–phenotype maps of its irQTL in different tissues illustrated that these isoform-level mediators could not be detected in the standard eQTL analysis (Fig. 2e), as not the gene expression themselves but rather the relative proportions between the isoforms regulated the downstream phenotypes.

Discussion

We have conducted a series of investigations for brain irQTL, i.e., the cis-regulatory loci in the brain tissues that control the relative isoform proportions per gene instead of the expression levels. Besides identifying hundreds of irQTL that could not be detected as eQTL, we found that genes with such irQTL regulatory property harbor enriched heritability for human complex traits, especially here for neuro- or nerve-related phenotypes. We also inferred that genetically regulated isoform distributions have downstream effects on the phenotypes via MR. Our analysis highlights the importance of quantifying and studying isoform expressions rather than general gene expressions. Some genetically regulated functional transcripts may only be detected when the isoforms are adequately quantified.

We used our previously developed XAEM algorithm to estimate isoform expression in the GTEx brain tissue samples. Although the isoform expression level could not be directly obtained from RNA sequencing reads, the XAEM algorithm allows powerful quantification of isoform expression for multi-isoform genes. The estimated isoform expressions allowed us to demonstrate the regulation of gene expressions that could not be well characterized without dissecting into isoforms. For isoform expression estimation, transcriptome annotation reference is a factor to be considered. The more comprehensive references, such as Ensembl (Cunningham et al. 2021) and GENCODE (Frankish et al. 2020), contain much more transcripts, and many of them are not curated isoforms. Many exons in these references for a gene are only a few bases different from others, making some isoforms very similar to each other. Too many of too similar isoforms in the cluster response profile (CRP; the X matrix) of XAEM would worsen the estimation. This is true for any isoform quantification algorithm that relies on an isoform annotation reference, as it is difficult to have sequencing reads that well distinguish very similar isoforms. For normal analysis such as this study, it is better to use curated isoforms in RefSeq (O'Leary et al. 2015) (default CRPs inbuilt in XAEM), so that the results are more reliable.

The mapped irQTL could not be identified as a typical eQTL since the genetic effects on different isoforms per gene had different signs. Thus, although the genetic effects on different isoform expressions for the same gene were all strong, the gene's overall expression could still be consistent across individuals with different genotypes (lacking genetic variance). This phenomenon can be generalized to other composite phenotypes too. In general, we would need to go deeper into the specific genetically regulated phenotype (in this case, isoform expressions) instead of only studying the composite phenotype (in this case, overall gene expressions).

We decided only to use age, sex, and three PCs as covariates mainly due to two perspectives: (1) we aimed to control the number of covariates in such small-sample association analysis to save degrees of freedom, as long as the inflation factor can be well controlled, and we found that the current setting was sufficient (inflation factor $\lambda = 1.030$ at the median and $\lambda = 1.018$ at the 25% quantile, Supplementary Fig. 1); (2) although the RNA sequencing data here are from multiple tissues, they are all from the brain, and the technical and biological conditions are less heterogeneous comparing to experiments in other tissues used in GTEx. Considering these, we did not consider more PCs or other covariates. In general, in small-sample genetic association analysis, the trade-off between degrees of freedom and power is a concern.

It is essential to clarify the difference in mapping ordinary sQTL and irQTL. First, sQTL mapping tools such as sQTLseekeR (Monlong et al. 2014) (used by the GTEx sQTL analysis), LeaftCutter (Li et al. 2018) (annotation-free), and the recently developed THISTLE method (Qi

et al. 2022) (annotation-based isoform-level genetic effects heterogeneity test) aim to detect genetic regulation of "alternative splicing events", namely, whether alternative splicing happens more for a certain genotype. Studying the isoform expressions as phenotypes themselves was not straightforward, and the main reason behind this is the great challenge in estimating isoform expressions using short-read RNA-sequencing data. Our initial idea of this work is to emphasize that isoform expression itself can be treated as an analyzable phenotype, as long as the estimation accuracy is sufficient. XAEM is a tool that substantially improves isoform expression estimation and thus fits the purpose. Although the estimation is not perfect, we showed that a number of novel isoform-level expression QTL could be mapped, which were missed in eQTL mapping (which neglects alternative splicing) and standard GTEx sQTL analysis (which does not have comparable isoform expression estimation). We would like to note that the comparison between irQTL and sQTL is subject to current statistical power. A slight difference in the genetic effects of irQTL for the isoforms of the same gene would be detected as sQTL when the power grows; nevertheless, as long as the isoform expressions can be well quantified (e.g., more commonly in the future with long-read sequencing techniques), studying isoform expressions as phenotypes would directly give us information about sQTL.

Different types of molecular QTL were studied in the human brain. Besides sQTL (Qi et al. 2022; Takata et al. 2017; Zhang et al. 2020) and eQTL (O'Brien et al. 2018), methylation QTL (mQTL) (Gibbs et al. 2010; Ng et al. 2017) were also investigated to integrate epigenetic biology with gene expressions. Some of these molecular QTL were found to target biomarkers for neuropsychiatric disorders, and these efforts essentially have constructed a roadmap for genetically regulated molecular mechanisms in the human brain. We also focused on the brain tissues as the brain has the richest alternative splicing events and particular functions, allowing us to subsequently link to particular phenotypes strongly related to the brain functions. More could be done by assessing all available tissue samples from GTEx Consortium; nevertheless, it would require substantially more computational resources. We also expect the heterogeneity test method THISTLE to gain further power when incorporating the XAEM algorithm in future studies. As the most collected tissue, whole-blood RNA sequencing data are available in multiple human cohorts. We foresee a consortium-based investigation of irQTL in larger consortia and potentially provide a comprehensive assessment of irQTL associated with various human complex traits and diseases.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s43657-023-00100-6.

Acknowledgements We thank the GTEx Consortium and the dbGaP repository for providing access to the data analyzed. We thank the Swedish National Infrastructure for Computing (SNIC) for providing the computational resources for processing and analyzing the data. We thank Dr. Jie Zheng for helping us access the summary-level data in LD-Hub. We thank Benjamin Neale's lab for releasing the publicly available summary association statistics of UK Biobank phenotypes.

Authors' Contributions XS initiated and coordinated the study. LP, CZ, and ZY performed data analysis. YP, TNV, and XS supervised the study. LP, CZ, ZY, and XS wrote the manuscript. All authors approved the submitted version of the manuscript.

Funding XS was in receipt of a National Natural Science Foundation of China (NSFC) grant (No. 12171495), a Natural Science Foundation of Guangdong Province grant (No. 2021A1515010866), a National Key Research and Development Program grant (No. 2022YFF1202105), and Swedish Research Council (Vetenskapsrådet) grants (No. 2017-02543 & No. 2022-01309). LP was supported by the Swedish Research Council grant (No. 2017-02543) to XS. The Swedish National Infrastructure for Computing (SNIC) utilized was partially funded by the Swedish Research Council through grant agreement No. 2018-05973.

Code Availability Codes to reproduce the results are deposited in Github (https://github.com/eudoraleer/Code_for_Brain_Isoform).

Data Availability RNA-Sequencing data and WGS data from GTEx Consortium were obtained from the dbGaP repository under accession phs000424.v7.p2.c1 and phs000424.v7.p1 respectively. GWAS summary statistics for complex traits and diseases were obtained from the LD-Hub (https://ldsc.broadinstitute.org/) and Neale's lab UKB round2 GWAS summary-level data (http://www.nealelab.is/uk-biobank).

Declarations

Conflict of interests The authors declare no competing financial interest.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication All authors have approved the publication of this work.

References

Altshuler DM, Gibbs RA, Peltonen L, Altshuler DM, Gibbs RA, Peltonen L, Dermitzakis E, Schaffner SF, Yu F, Peltonen L, Dermitzakis E, Bonnen PE, Altshuler DM, Gibbs RA, de Bakker PIW, Deloukas P, Gabriel SB, Gwilliam R, Hunt S, Inouye M, Jia X, Palotie A, Parkin M, Whittaker P, Yu F, Chang K, Hawes A, Lewis LR, Ren Y, Wheeler D, Gibbs RA, Marie Muzny D, Barnes C, Darvishi K, Hurles M, Korn JM, Kristiansson K, Lee C, McCarroll SA, Nemesh J, Dermitzakis E, Keinan A, Montgomery SB, Pollack S, Price AL, Soranzo N, Bonnen PE, Gibbs RA, Gonzaga-Jauregui C, Keinan A, Price AL, Yu F, Anttila V, Brodeur W, Daly MJ, Leslie S, McVean G, Moutsianas L, Nguyen H, Schaffner SF, Zhang Q, Ghori MJR, McGinnis R, McLaren W, Pollack S, Price AL, Schaffner SF, Takeuchi F, Grossman SR, Shlyakhter I, Hostetter EB, Sabeti PC, Adebamowo CA, Foster MW, Gordon DR, Licinio J, Cristina Manca M, Marshall PA, Matsuda I, Ngare D, Ota Wang V, Reddy D, Rotimi CN, Royal CD, Sharp RR, Zeng C, Brooks LD, McEwen JE, The International HapMap C, Principal i, Project coordination l, Manuscript writing g, Genotyping, Qc, sequencing E, discovery SNP, Copy number variation t, analysis, Population a, Low frequency variation a, Linkage d, haplotype sharing a, Imputation, Natural s, Community e, sample collection g, Scientific m (2010) Integrating common and rare genetic variation in diverse human populations. Nature 467(7311):52–58. https://doi.org/10.1038/nature09298

Ardlie KG, Deluca DS, Segrè AV, Sullivan TJ, Young TR, Gelfand ET, Trowbridge CA, Maller JB, Tukiainen T, Lek M, Ward LD, Kheradpour P, Iriarte B, Meng Y, Palmer CD, Esko T, Winckler W, Hirschhorn JN, Kellis M, MacArthur DG, Getz G, Shabalin AA, Li G, Zhou Y-H, Nobel AB, Rusyn I, Wright FA, Lappalainen T, Ferreira PG, Ongen H, Rivas MA, Battle A, Mostafavi S, Monlong J, Sammeth M, Mele M, Reverter F, Goldmann JM, Koller D, Guigó R, McCarthy MI, Dermitzakis ET, Gamazon ER, Im HK, Konkashbaev A, Nicolae DL, Cox NJ, Flutre T, Wen X, Stephens M, Pritchard JK, Tu Z, Zhang B, Huang T, Long Q, Lin L, Yang J, Zhu J, Liu J, Brown A, Mestichelli B, Tidwell D, Lo E, Salvatore M, Shad S, Thomas JA, Lonsdale JT, Moser MT, Gillard BM, Karasik E, Ramsey K, Choi C, Foster BA, Syron J, Fleming J, Magazine H, Hasz R, Walters GD, Bridge JP, Miklos M, Sullivan S, Barker LK, Traino HM, Mosavel M, Siminoff LA, Valley DR, Rohrer DC, Jewell SD, Branton PA, Sobin LH, Barcus M, Qi L, McLean J, Hariharan P, Um KS, Wu S, Tabor D, Shive C, Smith AM, Buia SA, Undale AH, Robinson KL, Roche N, Valentino KM, Britton A, Burges R, Bradbury D, Hambright KW, Seleski J, Korzeniewski GE, Erickson K, Marcus Y, Tejada J, Taherian M, Lu C, Basile M, Mash DC, Volpi S, Struewing JP, Temple GF, Boyer J, Colantuoni D, Little R, Koester S, Carithers LJ, Moore HM, Guan P, Compton C, Sawyer SJ, Demchok JP, Vaught JB, Rabiner CA, Lockhart NC, Ardlie KG, Getz G, Wright FA, Kellis M, Volpi S, Dermitzakis ET (2015) The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science 348(6235):648–660. https://doi.org/10.1126/science.1262110

Battle A, Mostafavi S, Zhu X, Potash JB, Weissman MM, McCormick C, Haudenschild CD, Beckman KB, Shi J, Mei R, Urban AE, Montgomery SB, Levinson DF, Koller D (2014) Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. Genome Res 24(1):14–24. https://doi.org/10.1101/gr.155192.113

Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J (2017) A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. Stat Med 36(11):1783–1802. https://doi.org/10.1002/sim.7221

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol 34(5):525–527. https://doi.org/10.1038/nbt.3519

Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, Daly MJ, Price AL, Neale BM, Schizophrenia Working Group of the Psychiatric Genomics C (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47(3):291–295. https://doi.org/10.1038/ng.3211

Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean Irina M, Austine-Orimoloye O, Azov Andrey G, Barnes I, Bennett R, Berry A, Bhai J, Bignell A, Billis K, Boddu S, Brooks L, Charkhchi M, Cummins C, Da Rin FL, Davidson C, Dodiya K, Donaldson S, El Houdaigui B, El Naboulsi T, Fatima R, Giron CG, Genez T, Martinez Jose G, Guijarro-Clarke C, Gymer A, Hardy M, Hollis Z, Hourlier T, Hunt T, Juettemann T, Kaikala V, Kay M, Lavidas I, Le T, Lemos D, Marugán JC, Mohanan S, Mushtaq A, Naven M, Ogeh Denye N, Parker A, Parton A, Perry M, Piližota I, Prosovetskaia I, Sakthivel Manoj P, Salam Ahamed Imran A, Schmitt Bianca M, Schuilenburg H, Sheppard D, Pérez-Silva José G, Stark W, Steed E, Sutinen K, Sukumaran R, Sumathipala D, Suner M-M, Szpak M, Thormann A, Tricomi FF, Urbina-Gómez D, Veidenberg A, Walsh Thomas A, Walts B, Willhoft N, Winterbottom A, Wass E, Chakiachvili M, Flint B, Frankish A, Giorgetti S, Haggerty L, Hunt Sarah E, IIsley Garth R, Loveland Jane E, Martin Fergal J, Moore B, Mudge Jonathan M, Muffato M, Perry E, Ruffier M, Tate J, Thybert D, Trevanion Stephen J, Dyer S, Harrison Peter W, Howe Kevin L, Yates Andrew D, Zerbino Daniel R, Flicek P (2021) Ensembl 2022. Nucleic Acids Res 50(D1):D988–D995. https://doi.org/10.1093/nar/gkab1049

Deng W, Mou T, Kalari KR, Niu N, Wang L, Pawitan Y, Vu TN (2019) Alternating EM algorithm for a bilinear model in isoform quantification from RNA-seq data. Bioinformatics 36(3):805–812. https://doi.org/10.1093/bioinformatics/btz640

Frankish A, Diekhans M, Jungreis I, Lagarde J, Loveland Jane E, Mudge JM, Sisu C, Wright JC, Armstrong J, Barnes I, Berry A, Bignell A, Boix C, Carbonell Sala S, Cunningham F, Di Domenico T, Donaldson S, Fiddes Ian T, García Girón C, Gonzalez JM, Grego T, Hardy M, Hourlier T, Howe KL, Hunt T, Izuogu OG, Johnson R, Martin FJ, Martínez L, Mohanan S, Muir P, Navarro FCP, Parker A, Pei B, Pozo F, Riera FC, Ruffier M, Schmitt BM, Stapleton E, Suner M-M, Sycheva I, Uszczynska-Ratajczak B, Wolf MY, Xu J, Yang Yucheng T, Yates A, Zerbino D, Zhang Y, Choudhary Jyoti S, Gerstein M, Guigó R, Hubbard TJP, Kellis M, Paten B, Tress ML, Flicek P (2020) GENCODE 2021. Nucleic Acids Res 49(D1):D916–D923. https://doi.org/10.1093/nar/gkaa1087

Gazal S, Finucane HK, Furlotte NA, Loh P-R, Palamara PF, Liu X, Schoech A, Bulik-Sullivan B, Neale BM, Gusev A, Price AL (2017) Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection. Nat Genet 49(10):1421–1427. https://doi.org/10.1038/ng.3954

Gibbs JR, van der Brug MP, Hernandez DG, Traynor BJ, Nalls MA, Lai S-L, Arepalli S, Dillman A, Rafferty IP, Troncoso J, Johnson R, Zielke HR, Ferrucci L, Longo DL, Cookson MR, Singleton AB (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. PLOS Genetics 6(5):e1000952. https://doi.org/10.1371/journal.pgen.1000952

Glass D, Viñuela A, Davies MN, Ramasamy A, Parts L, Knowles D, Brown AA, Hedman ÅK, Small KS, Buil A, Grundberg E, Nica AC, Di Meglio P, Nestle FO, Ryten M, Durbin R, McCarthy MI, Deloukas P, Dermitzakis ET, Weale ME, Bataille V, Spector TD, the UKBEc, the Mu Tc, (2013) Gene expression changes with age in skin, adipose tissue, blood and brain. Genome Biol 14(7):R75. https://doi.org/10.1186/gb-2013-14-7-r75

Hashimoto S, Kohsaka M, Morita N, Fukuda N, Honma S, Honma K-i (1996) Vitamin B12 enhances the phase-response of circadian melatonin rhythm to a single bright light exposure in humans. Neurosci Lett 220(2):129–132. https://doi.org/10.1016/S0304-3940(96)13247-X

Lappalainen T, Sammeth M, Friedländer MR, t Hoen PA, Monlong J, Rivas MA, Gonzàlez-Porta M, Kurbatova N, Griebel T, Ferreira PG, Barann M, Wieland T, Greger L, van Iterson M, Almlöf J, Ribeca P, Pulyakhina I, Esser D, Giger T, Tikhonov A, Sultan M, Bertier G, MacArthur DG, Lek M, Lizano E, Buermans HP, Padioleau I, Schwarzmayr T, Karlberg O, Ongen H, Kilpinen H, Beltran S, Gut M, Kahlem K, Amstislavskiy V, Stegle O, Pirinen M, Montgomery SB, Donnelly P, McCarthy MI, Flicek P, Strom TM, Lehrach H, Schreiber S, Sudbrak R, Carracedo A, Antonarakis SE, Häsler R, Syvänen AC, van Ommen GJ, Brazma A, Meitinger T, Rosenstiel P, Guigó R, Gut IG, Estivill X, Dermitzakis ET (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature 501(7468):506–511. https://doi.org/10.1038/nature12531

Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, Pritchard JK (2018) Annotation-free quantification of RNA splicing using LeafCutter. Nat Genet 50(1):151–158. https://doi.org/10.1038/s41588-017-0004-9

Mayer G, Kröger M, Meier-Ewert K (1996) Effects of vitamin B12 on performance and circadian rhythm in normal subjects. Neuropsychopharmacology 15(5):456–464. https://doi.org/10.1016/s0893-133x(96)00055-3

Monlong J, Calvo M, Ferreira PG, Guigó R (2014) Identification of genetic variants associated with alternative splicing using sQTLseekeR. Nat Commun 5(1):4698. https://doi.org/10.1038/ncomms5698

Ng B, White CC, Klein H-U, Sieberts SK, McCabe C, Patrick E, Xu J, Yu L, Gaiteri C, Bennett DA, Mostafavi S, De Jager PL (2017) An xQTL map integrates the genetic architecture of the human brain's transcriptome and epigenome. Nat Neurosci 20(10):1418–1426. https://doi.org/10.1038/nn.4632

Nilsen TW, Graveley BR (2010) Expansion of the eukaryotic proteome by alternative splicing. Nature 463(7280):457–463. https://doi.org/10.1038/nature08909

O'Brien HE, Hannon E, Hill MJ, Toste CC, Robertson MJ, Morgan JE, McLaughlin G, Lewis CM, Schalkwyk LC, Hall LS, Pardiñas AF, Owen MJ, O'Donovan MC, Mill J, Bray NJ (2018) Expression quantitative trait loci in the developing human brain and their enrichment in neuropsychiatric disorders. Genome Biol 19(1):194. https://doi.org/10.1186/s13059-018-1567-1

O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, Rajput B, Robbertse B, Smith-White B, Ako-Adjei D, Astashyn A, Badretdin A, Bao Y, Blinkova O, Brover V, Chetvernin V, Choi J, Cox E, Ermolaeva O, Farrell CM, Goldfarb T, Gupta T, Haft D, Hatcher E, Hlavina W, Joardar VS, Kodali VK, Li W, Maglott D, Masterson P, McGarvey KM, Murphy MR, O'Neill K, Pujar S, Rangwala SH, Rausch D, Riddick LD, Schoch C, Shkeda A, Storz SS, Sun H, Thibaud-Nissen F, Tolstoy I, Tully RE, Vatsan AR, Wallin C, Webb D, Wu W, Landrum MJ, Kimchi A, Tatusova T, DiCuccio M, Kitts P, Murphy TD, Pruitt KD (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res 44(D1):D733–D745. https://doi.org/10.1093/nar/gkv1189

Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nat Genet 40(12):1413–1415. https://doi.org/10.1038/ng.259

Park E, Pan Z, Zhang Z, Lin L, Xing Y (2018) The expanding landscape of alternative splicing variation in human populations. Am J Hum Genet 102(1):11–26. https://doi.org/10.1016/j.ajhg.2017.11.002

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14(4):417–419. https://doi.org/10.1038/nmeth.4197

Planell N, Lagani V, Sebastian-Leon P, van der Kloet F, Ewing E, Karathanasis N, Urdangarin A, Arozarena I, Jagodic M, Tsamardinos I, Tarazona S, Conesa A, Tegner J, Gomez-Cabrero D (2021) STATegra: multi-omics data integration—a conceptual scheme with a bioinformatics pipeline. Front Genet 12:1–12. https://doi.org/10.3389/fgene.2021.620453

Porter RS, Jaamour F, Iwase S (2018) Neuron-specific alternative splicing of transcriptional machineries: implications for neurodevelopmental disorders. Mol Cell Neurosci 87:35–45. https://doi.org/10.1016/j.mcn.2017.10.006

Nature Portfolio (2022) Alternative splicing. https://www.nature.com/subjects/alternative-splicing. Accessed 4 Aug 2022

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575. https://doi.org/10.1086/519795

Qi T, Wu Y, Fang H, Zhang F, Liu S, Zeng J, Yang J (2022) Genetic control of RNA splicing and its distinct role in complex trait variation. Nat Genet 54(9):1355–1363. https://doi.org/10.1038/s41588-022-01154-4

Safran M, Rosen N, Twik M, BarShir R, Stein TI, Dahary D, Fishilevich S, Lancet D (2021) The GeneCards suite. In: Abugessaisa I, Kasukawa T (eds) Practical guide to life science databases. Springer, Singapore, pp 27–56

Schafer S, Miao K, Benson CC, Heinig M, Cook SA, Hubner N (2015) Alternative splicing signatures in RNA-seq data: percent spliced in (PSI). Curr Prot Hum Genet 87(1):11.16.11–11.16.14. https://doi.org/10.1002/0471142905.hg1116s87

Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, Burgess S, Jiang T, Paige E, Surendran P, Oliver-Williams C, Kamat MA, Prins BP, Wilcox SK, Zimmerman ES, Chi A, Bansal N, Spain SL, Wood AM, Morrell NW, Bradley JR, Janjic N, Roberts DJ, Ouwehand WH, Todd JA, Soranzo N, Suhre K, Paul DS, Fox CS, Plenge RM, Danesh J, Runz H, Butterworth AS (2018) Genomic atlas of the human plasma proteome. Nature 558(7708):73–79. https://doi.org/10.1038/s41586-018-0175-2

Takata A, Matsumoto N, Kato T (2017) Genome-wide identification of splicing QTLs in the human brain and their enrichment among schizophrenia-associated loci. Nat Commun 8(1):14519. https://doi.org/10.1038/ncomms14519

Thakur PK, Rawal HC, Obuca M, Kaushik S (2019) Bioinformatics approaches for studying alternative splicing. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C (eds) Encyclopedia of bioinformatics and computational biology. Academic Press, Oxford, pp 221–234

Walker RL, Ramaswami G, Hartl C, Mancuso N, Gandal MJ, de la Torre-Ubieta L, Pasaniuc B, Stein JL, Geschwind DH (2019) Genetic control of expression and splicing in developing human brain informs disease mechanisms. Cell 179(3):750-771.e722. https://doi.org/10.1016/j.cell.2019.09.021

Xu Q, Modrek B, Lee C (2002) Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. Nucleic Acids Res 30(17):3754–3766. https://doi.org/10.1093/nar/gkf492

Yeo G, Holste D, Kreiman G, Burge CB (2004) Variation in alternative splicing across human tissues. Genome Biol 5(10):R74. https://doi.org/10.1186/gb-2004-5-10-r74

Zaghlool A, Ameur A, Cavelier L, Feuk L (2014) Chapter five—splicing in the human brain. In: McWeeney S (ed) Hitzemann R. International Review of Neurobiology, Academic Press, pp 95–125

Zhang Y, Yang HT, Kadash-Edmondson K, Pan Y, Pan Z, Davidson BL, Xing Y (2020) Regional variation of splicing QTLs in

human brain. Am J Hum Genet 107(2):196–210. https://doi.org/10.1016/j.ajhg.2020.06.002

Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, Hemani G, Tansey K, Laurin C, Genetics E, Consortium LEE, Pourcain BS, Warrington NM, Finucane HK, Price AL, Bulik-Sullivan BK, Anttila V, Paternoster L, Gaunt TR, Evans DM, Neale BM (2016) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics 33(2):272–279. https://doi.org/10.1093/bioinformatics/btw613