Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

Alternative transcription initiation expands the genetic regulatory landscape of the human brain

Hui Chen[1,3#], Ting Shen[2#], Ting Zhang[1], Xing Li[1], Xin Meng[1], Dinglin Luo[1], Xudong Zou[1*], Mireya Plass[3,4*], Chao Chen[2*], Lei Li[1*]

1. Institute of Systems and Physical Biology, Shenzhen Bay Laboratory, Shenzhen 518055, China
2. MOE Key Laboratory of Rare Pediatric Diseases & Hunan Key Laboratory of Medical Genetics, School of Life Sciences, and Department of Psychiatry, The Second Xiangya Hospital, Central South University, Changsha 410000, Hunan, China
3. Physiological Sciences Department, Faculty of Medicine and Health Sciences, University of Barcelona, L'Hospitalet del Llobregat 08907, Spain.
4. Center for Networked Biomedical Research on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Madrid 28029, Spain.

[#]These authors contributed equally to this study.

[*]Correspondence: lei.li@szbl.ac.cn; chenchao@sklmg.edu.cn; mplass@ub.edu; zouxd@szbl.ac.cn

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

Abstract

Alternative transcription initiation (ATI) plays an important role in transcript diversity and protein production. However, its landscape and genetic regulation in the human brain remain largely unexplored. Here, we perform a systematic analysis of ATI across seven brain regions and construct a cross-ancestry atlas of ATI quantitative trait loci (atiQTLs) using 7,710 RNA sequencing samples. We identify 22,838 alternative first exon events in 12,462 genes and 17,981 tandem transcription start site events in 10,846 genes, most of which show region-specific usage that is often independent of gene expression. We find 688,204 atiQTLs across brain regions, on average 26.96% of which are independent of expression QTLs (eQTLs). By integrating atiQTLs with genome-wide association studies of 18 brain-related diseases and traits through colocalization and transcriptome-wide association analyses, we further identify 204 candidate susceptibility genes predicted to modulate disease risk via ATI. As proof of principle validation, we show that the Parkinson's disease risk locus at *SCARB2* regulates tandem TSS usage, and the short *SCARB2* 5′ untranslated region (5′UTR) leads to higher reporter activity than the long 5′UTR in neural progenitor cells, whereas mutating in the long 5′UTR partially relieves this repression. Moreover, integration with protein QTLs reveals that approximately 40% of atiQTLs affecting alternative 5′ end usage are associated with protein abundance, 37% to 46% of which are independent of eQTLs. Together, our study highlights the significant role of ATI as a distinct layer of genetic regulation in the human brain and provides a strong foundational framework for understanding the etiology underlying brain disorders.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

Introduction

Neuropsychiatric disorders are widely recognized as complex diseases with a substantial genetic component[1-3]. Large-scale genetic studies have shown that common variants across the human genome can account for at least 10% of the SNP-based heritability of neuropsychiatric disorders[4,5]. Understanding how noncoding risk variants identified in genome-wide association studies (GWAS) influence molecular function in the human brain remains a central challenge in the study of neuropsychiatric disorders. Expression quantitative trait locus (eQTL) studies in brain tissues have linked a subset of disease risk variants to candidate genes and revealed substantial effects on cellular processes[6-8]. Recent studies have also linked non-coding variants to alternative splicing (sQTL) and 3′ UTR regulation (3′aQTL)[9]. However, these studies of genetic regulation did not examine alternative transcription initiation, another key process in gene regulation, and many non-coding disease variants remained uncharacterized.

Control of transcription initiation is the primary layer of gene expression regulation. In the mammalian genome, over 50% of protein-coding genes have alternative transcription start sites (TSSs)[10,11]. Alternative transcription initiation (ATI) generates transcripts that differ in their first exons or in the length of their 5′ untranslated region (5′ UTR), giving rise to alternative first exons (AFEs) or tandem transcription start sites (ATTSSs)[12], respectively. The use of AFE produces transcripts with different open reading frames (ORFs), expanding the repertoire of encoded proteins through isoforms with alternative N termini[13], whereas ATTSS can produce transcripts that largely share the same coding region but have distinct 5′ UTRs, thereby determining translation efficiency through regulatory elements such as short upstream ORFs (uORFs)[14,15]. Studies across human tissues have established widespread variation in promoter usage and tandem transcription initiation[16,17], and our previous work showed that tandem-initiation phenotypes are genetically regulated and can reveal disease-associated effects that are not captured by conventional expression analyses[18]. ATI plays a crucial role in neuronal development and function[19,20]. Recent work has suggested that alternative promoter usage occurs across brain regions and may play a role in neuropsychiatric disorders[21]. Together with studies of isoform and splicing regulation[22,23], these findings indicate that transcript-level phenotypes can uncover layers of genetic regulation beyond total gene expression. Yet transcript initiation and its genetic regulation have not been systematically characterized in the human brain, leaving its contribution to brain regulatory variation and disease genetics largely unknown.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

Here we systematically characterize the genetic regulation of transcript initiation in the human brain using 7,710 RNA-sequencing profiles spanning seven brain regions, including a uniquely generated Chinese brain cohort. By analyzing AFE and ATTSS usage together, we characterize regional and genetic variation in transcript initiation and determine what it reveals beyond total gene expression. We use data from diverse populations to refine the interpretation of initiation-associated genetic effects and integrate these signals with brain-related genome-wide association studies and protein-associated genetic variation.[12] These analyses identify disease-associated loci whose regulatory effects are preferentially captured through transcript initiation. At the Parkinson's disease-associated *SCARB2* locus, we link genetically regulated tandem initiation to alternative 5′ UTR architecture and its functional consequences in neural progenitor cells. Together, our study provides a genetic map of transcript initiation in the human brain across populations and identifies a little-studied layer of regulation that links noncoding variation to brain disease.

Results

Alternative transcription initiation diversifies the human brain transcriptome

To characterize variation in transcription initiation in the human central nervous system, we integrated 11 brain transcriptomic resources, including AMP-AD (ROSMAP, MAYO and MSBB)[24], PsychENCODE[25] (BrainGVEX, UCLA_ASD, Bipseq[26], CMC and CMC_HBCC[27]), GTEx[28], NABEC[29] and ChinaBrain[30] (Fig. 1a and Supplementary Fig. 1). After removing duplicate samples and applying quality control (Supplementary Figs. 2-3), we retained 7,710 RNA-seq samples and assigned them to seven anatomical groups: amygdala, basal ganglia, cerebellum, cortex, hippocampus, hypothalamus and spinal cord (Supplementary Table 1). The 11 resources included control samples and samples from individuals with major neurological and psychiatric disorders, including Alzheimer's disease (AD), schizophrenia (SCZ) and bipolar disorder (Supplementary Fig. 1). We quantified alternative first exon (AFE) and alternative tandem transcription start site (ATTSS) usage using proActiv[31] and DATTSS[17], respectively. For ATTSS events, DATTSS calculates the distal transcription start site usage index (DTUI), which reflects the relative contribution of transcripts initiating from the upstream (distal) TSS; lower DTUI values indicate greater usage of the downstream (proximal) TSS.

Across all samples and brain regions, we identified 22,838 non-redundant AFE events across 12,462 genes and 17,981 ATTSS events across 10,846 genes (Fig. 1b). Although most

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

AFE events were associated with a single isoform, 38.9% of genes had at least two distinct first exons, indicating widespread diversity in transcript initiation across the brain (Supplementary Fig. 4a). The median distal transcription start site usage index (DTUI) was 0.30 across brain regions, reflecting substantial variation in the relative usage of distal and proximal TSSs (Supplementary Fig. 4b). We next examined the relationship between ATI usage and total gene expression. Only 9.04% of ATTSS–gene pairs and 22.96% of AFE–gene pairs had an absolute Spearman correlation above 0.3, suggesting that ATI usage was generally only modestly correlated with total gene expression (Fig. 1c). For example, distal TSS usage of *NTRK2* varied across amygdala samples but was not significantly correlated with total *NTRK2* expression (Spearman's $\rho = -0.064$, $P = 0.49$; Fig. 1c). These observations show that changes in total gene expression do not necessarily capture variation in transcript initiation.

To assess how ATI affects gene products in the brain, we compared functional regions (5' UTR, coding sequence [CDS], 3' UTR) with those associated with the major promoter using GENCODE (v47) annotations (Supplementary Fig. 5a). The proximal TSSs inferred by DATTSS were predominantly located within the first quarter of the first exon region, with a median retained fraction of 42.4% (Supplementary Fig. 5b). ATI also substantially affects transcript coding regions. In protein- coding genes with ATTSS events, only 40% of the CDS remains unchanged on average, indicating extensive coding- level consequences of alternative promoter usage (Supplementary Fig. 5c), whereas almost 64.7% of such events give rise to isoforms that may use a different 3' UTR sequence based on annotations (Supplementary Fig. 5c).

To assess how often ATI usage contributes to context-specific regulation, we looked for events whose usage varied significantly across brain regions and was independent of overall expression levels (FDR-adjusted $P < 0.05$, Fig. 1d**)**. ATI events that regulate distinct gene isoforms can show marked regional specificity while remaining largely independent of total gene expression (Fig. 1e-f). For example, the gene encoding calcyon neuron specific vesicular protein (*CALY)* showed higher distal TSS usage in the cerebellum than in all other regions (Figures 1g-h**)**, yet usage was only weakly correlated with total gene expression (Spearman's $\rho = -0.24$, FDR = 2.36×10^{-27}). Similarly, usage of the second alternative first exon of *CHST12* (carbohydrate sulfotransferase 12) was lowest in the cerebellum (Figures 1g-h) and was independent of its overall expression level (Spearman's $\rho = -0.011$, FDR = 0.70).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

Comparison with non-brain tissues identified 410 ATTSS events with brain-associated shifts in start-site usage, of which 228 favored lengthening and 182 favored shortening (Supplementary Fig. 6a). Genes exhibiting brain-associated lengthening were enriched for nucleotide biosynthetic processes, whereas those exhibiting shortening were enriched for processes including forebrain and cerebral cortex cell migration, mitochondrial calcium ion homeostasis and selective autophagy (Supplementary Fig. 6b). These findings link tissue-associated initiation patterns to distinct gene sets, including those involved in neural development (Supplementary Fig. 6c). Differential-usage analysis also identified events with altered ATTSS usage and comparatively stable expression of the corresponding gene (Supplementary Fig. 6d). ATTSS usage differed between disease and control samples, and the direction of these differences varied across disorders. Shortening events predominated in SCZ, lengthening events were more frequent in BIP, and AD showed similar numbers in both directions (Supplementary Figs. 6e, f).

These findings highlight ATI as an additional layer of transcriptomic diversity in the human brain, with regional and disease-associated usage patterns that are largely decoupled from total gene expression.

Genetic variation shapes transcript initiation across brain regions

To identify genetic determinants of ATI across brain regions (atiQTL), we tested for associations between initiation phenotypes and variants within a *cis* window extending 1 Mb on either side of the relevant 5′ UTR region. In each brain region, we normalized AFE and ATTSS phenotypes and adjusted for known covariates, including donor age, sex and study batch, and for covariates inferred from DTUI using PCAForQTL[32] (Supplementary Figs. 7-8). We tested associations between 9.93 million variants with minor allele frequency (MAF) > 1% and 22,156 AFE events and 15,293 ATTSS events. We defined genes with at least one significant association at FDR < 0.05 as atiGenes. Across all brain regions, we identified 688,204 *cis*-atiQTL in 2, 060 ATTSS and 2,891 AFE events (Fig. 2a). These atiQTLs showed substantial genetically driven differences in ATI among individuals and affected genes involved in neuronal function, including *BIN1* (Bridging Integrator 1), a key regulator of endocytic trafficking and membrane dynamics (Supplementary Fig. 9). Across brain regions, jointly estimated atiQTL effects agreed more often in direction than in magnitude (Supplementary Fig. 10). To assess the genetic contribution to differences in ATI usage among individuals, we estimated the *cis*-heritability of each ATI event using GCTA

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

(Genome-wide Complex Trait Analysis)[33]. Although ATI events showed widespread genetic associations, estimated cis-heritability was generally modest (Supplementary Fig. 11), averaging 18.3% for AFE and 8.4% for ATTSS events. Most atiQTL variants lay near their target ATI events, with 70.68% of significantly associated SNPs within 100kb of the ATI event TSS (Fig. 2b). Based on computational fine-mapping analysis utilizing sum of single effects (SuSiE) regression[34], 41.59% of these atiQTLs have a single credible set, while only 11.23% of the credible sets harboring one single variant (median 11 variants per credible set; mean = 43.2; Supplementary Figs. 12a-b). When we stratified fine-mapped atiQTLs by posterior probability, high-confidence atiQTL SNPs (with higher PIPs) clustered preferentially near transcription start sites (Supplementary Fig. 12c). atiQTLs were enriched in 5′ UTRs and promoters, with mean $\log_2$(odds ratios) of 1.50 ± 0.23 and 0.92 ± 0.15, respectively, across regions (mean ± s.d.; Fig. 2c). These positional and functional patterns link the associations to the local sequence context of transcription initiation. Among ATI subtypes, atssQTLs were particularly enriched in promoter-proximal regions, including core promoters and their 100bp flanking sequences (Supplementary Fig.13) and in uORF regions (Fig. 2d), highlighting the importance of local transcriptional regulatory elements in shaping transcription initiation patterns.

Within the same brain regions, atiQTLs and eQTLs shared some target genes, indicating regulatory differences between the two types of variants. In genes with both eQTLs and atiQTLs, the lead SNPs tended to be separated, with a median distance of 22,340bp (Fig. 2e), and 37.8% of atiQTLs were poorly tagged (LD < 0.5) by eQTLs in matched tissues. These results suggest that atiQTLs can capture an additional layer of gene regulation.

To investigate the potential regulatory basis of atiQTLs, we integrated fine-mapping results with chromatin, transcription factor (TF) and RNA-related annotations. TF and RNA-binding protein (RBP) binding sites accounted for 45.7% and 9.8% of the total posterior inclusion probability, respectively (Supplementary Fig. 14a). Enrichment analysis highlighted binding sites of RBPs that regulate translation, including DDX3X and GRSF1, as well as sites for proteins involved in splicing and RNA stability (Supplementary Figs. 14b,c). Chromatin-state analysis across five brain tissues showed that ATTSS-associated variants were enriched in active promoters, whereas AFE-associated variants tended to occur in transcribed regions (Supplementary Fig. 14d). Analysis of 407 ENCODE ChIP–seq datasets identified enrichment for 57 TF binding annotations across at least three brain regions, including promoter-associated factors MAZ, YY1 and SP1 (Supplementary Figs. 14e). Sequence analysis identified nine enriched TF motifs, and predicted changes in motif affinity

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

correlated with ATI effect sizes for 44 TFs (FDR < 0.10; Supplementary Figs. 14f,g). For SP1, reduced predicted binding affinity was associated with increased ATTSS usage (Pearson's r = −0.57, P = 6×10^{-9}). These annotations implicate transcriptional and post-transcriptional processes as candidate mechanisms underlying ATI-associated genetic effects, although direct binding and functional consequences require experimental validation.

To improve fine-mapping resolution using data from multiple ancestries, we used SuSiEx to fine-map European and East Asian samples jointly and compared the resulting credible sets with those from ancestry-specific analyses. Joint fine-mapping yielded significantly smaller credible sets than single-ancestry analysis (two-sided Wilcoxon signed-rank tests, P = 4.5×10^{-5}, Fig. 2f).

These results show that combining ancestry data can refine ATI-associated signals while preserving the distinction between regulatory sharing and differences in statistical power.

Transcript initiation links brain disease risk to regulation beyond gene expression

We explored the extent to which atiQTL could explain to the genetic variants associated with brain disorders. We collected GWAS summary statistics for 18 brain-related diseases and traits (Supplementary Data 2), including neurological and psychiatric disorders, behavioral traits and other complex traits. We used stratified linkage disequilibrium score regression (S-LDSC)[35,36] to assess heritability enrichment for atiQTLs while controlling for 53 other functional annotations in the baseline-LD model. Enrichment was significantly higher for atiQTLs than for eQTLs in several brain-related diseases (Fig. 3a). For example, in the brain cortex, Parkinson's disease GWAS variants showed greater enrichment for atiQTLs (fold-enrichment = 2.23, $P = 4.04 \times 10^{-3}$) than for eQTLs (fold-enrichment = 1.38, $P = 2.06 \times 10^{-6}$).

To prioritize ATI events potentially involved in the risk of brain-related diseases, we integrated atiQTLs with GWAS summary statistics for 18 traits using colocalization[37,38] and SMR[39]. atiQTLs colocalized with 489 brain-related disease risk loci. Across all traits, SuSiE-COLOC and SMR jointly prioritized 173 ATI event–trait associations, whereas 265 and 342 associations were supported only by COLOC and SMR, respectively (Fig. 3b and Supplementary Data 3). A median of 74.3% of atiGenes implicated by both methods were not eGenes across the 18 traits (Fig. 3c), suggesting that conventional eQTL analysis might miss the regulatory effects captured by most ATI-linked susceptibility genes.

To identify ATI-associated susceptibility genes for brain disorders systematically, we extended the conventional FUSION TWAS framework [40] to model the contribution of ATI-

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

associated regulatory variants to trait-associated risk loci. We used a mixed-linear model to estimate the heritability of ATI levels attributable to *cis*-SNPs near the corresponding TSSs of each gene in a reference panel. We retained only genes with significant heritability estimates (*P*-value <0.05) for subsequent analyses. We then fitted predictive linear models (including Elastic Net, LASSO, BLUP, and Top1) for each ATI event using *cis*-SNPs and used five-fold cross-validation to select the model with the highest prediction accuracy. We obtained 2,901 ATI level prediction models, spanning 1,190 unique ATTSS events and 1,711 unique AFE events in 1,605 unique genes. Of note, we found that 34.4% (552/1,605) of atiTWAS models are in at least two brain regions, with an average of 484 atiTWAS models per brain region.

Because heritability constrains the predictive power of linear models, we evaluated model performance using normalized R^2 (defined as $R^2/cis\text{-}h^2$, Supplementary Fig. 15a). The atiTWAS models achieved an average predictive accuracy of 78.4% (Supplementary Fig. 15b), indicating that they captured a substantial proportion of the variance in *cis*-regulated ATI levels. These results show the broad coverage and high predictive utility of the atiTWAS framework for studying gene regulation in the brain.

Because LD structure and/or pleiotropic effects can confound TWAS associations[41], we applied causal-TWAS (cTWAS)[42] to GWAS for the same 18 brain-related diseases and traits. Unlike conventional TWAS, cTWAS uses a generalized Bayesian fine-mapping model that includes both imputed gene expression and genetic variants to correct for LD and pleiotropic effects, assigning posterior inclusion probabilities to candidate causal genes and thereby substantially reducing false discoveries in standard TWAS. We applied the FUSION-derived ATI prediction model to cTWAS and identified 86 high-confidence ATI events in 74 protein-coding genes with PIP > 0.8 (Figs 3d-e). Comparing these susceptibility genes across diseases revealed distinct disease clusters and gene pleiotropy, pointing to shared genetic mechanisms and potential common therapeutic targets (Supplementary Figs. 16a-b). The analysis recovered several well-established risk genes for brain disorders, including *SNCA*[43], *CTSF*[44], and *GPM6A*[45]. Notably, 63.51% of these identified genes were cross-validated and with target-disease evidence in the Open Targets Platform[46] (Supplementary Figs. 16c), exhibiting significantly higher disease association scores compared to the randomly selected null gene sets ($P = 3.5 \times 10^{-8}$, Supplementary Figs. 16d). We found robust support for 164 susceptibility gene-disease pairs from at least two of the three methods (colocalization, SMR and atiTWAS) (Fig. 3f). Functional enrichment analysis identified a highly interconnected

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

pathway network underlying the etiology of brain disorders, with mitochondrial bioenergetics, proteostasis and intracellular trafficking as prominent components (Fig. 3g).

Genetic regulation of transcript initiation is linked to protein abundance

Circulating protein levels and profiles can directly reflect broader pathologies, making them informative about health and disease. Genome-proteome-wide association studies provide substantial insights into the biology of diverse human diseases and inform prioritization of potential drug targets. To investigate the regulatory cascade linking ATI and protein abundance, we compared our atiQTLs with 8,799 protein abundance QTL (pQTLs) identified in previous studies (Supplementary Fig. 17a). We integrated publicly available summary statistics for (1) eQTLs from the Metabrain [6] resource, (2) plasma pQTLs[47-49], and (3) pQTLs from neurologically relevant tissues (brain and cerebrospinal fluid)[50,51]. Agreement among pQTLs was high across studies (π_1 estimates = 0.55−0.97, Supplementary Fig.17b**)**, although the degree of sharing varied by tissue context. Genes associated with the consolidated pQTLs were enriched for protein homeostasis and intracellular trafficking (e.g., protein transport and metabolic processes), cellular motility and structural organization (e.g., cell migration and membrane/organelle organization), and neuronal development and signaling pathways (e.g., axonogenesis and synapse organization) (Supplementary Figs. 17c-d), highlighting the pivotal roles of the processes captured by these pQTLs in maintaining proteostasis and regulating neuronal structural plasticity.

To characterize the relationship between atiQTLs and pQTLs, we first tested for colocalization across brain regions. ATI and protein abundance shared a substantial proportion of genetic regulatory signals, with 11.59% (1,020/8,799) of pQTL signals overlapping atiQTLs identified in brain tissues. Among these pairs, an average of 51.65% of atiQTLs colocalized with pQTLs (at a threshold of PP4>0.70, Figures 4a). We identified a subset of pQTLs (42.6%) that colocalized with atiQTLs in the absence of eQTL signals (Figure 4a), indicating that a subset of ATI events can drive variation in protein abundance independently of changes in total gene expression. For example, the pQTL (rs388294) for asparaginyl-tRNA synthetase 1 (*NARS1*), a gene known to influence progenitor proliferation in cortical brain organoids[52], colocalized exclusively with an atiQTL (PP4 = 0.994, Figure 4b). While eQTL–pQTL correlations were predominantly positive, atiQTL–pQTL correlations had a bimodal distribution with a substantial fraction of negative correlations (Fig. 4c).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

To further identify and prioritize putative disease risk genes with functional effects on the proteome, we integrated atiQTLs, pQTLs and GWAS summary statistics using HyPrColoc[53]. We identified 182 candidate genes that colocalized with brain-related diseases and traits, of which 143 colocalized through gene expression and 39 through ATI (Fig. 4d). Across diseases and traits, eQTLs showed a strong, statistically significant bias toward positive directional concordance with pQTLs ($P = 6.0 \times 10^{-6}$, Wilcoxon signed-rank test), whereas no such bias was observed for atiQTLs. Instead, atiQTLs showed a slight, statistically nonsignificant trend toward negative associations with pQTLs. (Fig. 4e**)**. Among these candidates, we identified 59 genes previously characterized as potentially druggable targets by cross-referencing the Therapeutic Target Database (TTD) [54]and the druggable genome[55] project (Figure 4f). We then combined data on ATI-colocalized pGenes with pharmacological information to assess their potential relevance to drug development and repurposing. We obtained drug-gene interactions from DGIdb v5.0[56]. On average, each pGene was linked to 18 drugs per brain-disorders, and up to 44 interactions were found in cross-disorder (Figure 4g). Representative pGenes in CDG, such as *SNAP91*, *GNL3*, and *POR,* were linked to several therapeutic classes, including antipsychotics, anxiolytics, and hypnotic and sedative agents (Figure 4h). These findings suggest that existing compounds could be investigated for alternative therapeutic applications involving the identified ATI-colocalized pGenes.

Alternative SCARB2 transcript initiation alters 5′UTR-dependent regulation

SCARB2 was a particularly strong candidate at a Parkinson's disease locus, with all three approaches supporting its ATI association (Fig. 5a). The agreement among methods prioritized a specific transcript-initiation phenotype at a disease-associated locus in a gene with an established role in lysosomal function. The *SCARB2* locus thus provided a candidate for studying the potential molecular consequences of genetically regulated ATI. The findings also raised the broader question of whether ATI-associated genetic effects extend beyond RNA-level regulation to protein abundance. We therefore integrated ATI associations with protein quantitative trait data to examine their relationship with protein-level variation.

SCARB2 encodes lysosomal integral membrane protein 2 (LIMP-2), whose role in glucocerebrosidase activity and α-synuclein clearance links it to Parkinson's disease biology[57]. At the *SCARB2* locus, the PD association[58] colocalized with the ATI signal (PP4 = 0.93), with additional support from SMR ($P = 1.68 \times 10^{-6}$; HEIDI P = 0.68; Figs. 5a, b). The

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

variant rs11547135 was associated with lower distal-start usage in the cerebellum (effect estimate = −0.46, P = 3.29×10^{-18}), corresponding to increased proximal-start usage (Fig. 5c-d). ATI effect directions varied across regions, so the cerebellar effect should not be generalized to all sampled tissues.

We first asked whether alternative *SCARB2* starts could be detected experimentally. In SH-SY5Y cells, 5′ rapid amplification of cDNA ends identified two transcript-start products, with sequences confirmed by Sanger sequencing (Fig. 5e). After transcriptional inhibition with actinomycin D, the long-5′ UTR transcript signal decreased more rapidly than the total *SCARB2* RNA signal over the measured time course (Fig. 5f). The long-transcript signal and the signal from the combined transcript pool thus showed different decay patterns. Because total RNA includes both forms, the experiment does not directly measure the half-life of the short transcript alone.

We then tested the functional effects of the alternative 5′ UTR sequences in M6-derived neural progenitor cells. The short-UTR reporter produced higher output than the long-UTR reporter ($P < 0.001$; Fig. 5g-h). Alternative-allele and deletion constructs within the long 5′ UTR increased reporter output relative to the unmodified long form, partially relieving repression. Both the short-versus-long comparison and the sequence perturbations indicate that the extended *SCARB2* 5′ UTR contains sequence that suppresses reporter output in this model.

The functional assay supports an effect of *SCARB2* 5′-UTR configuration at a genetically nominated PD locus. On its own, the assay cannot distinguish changes in reporter RNA abundance from changes in translation or show that rs11547135 changes endogenous TSS selection in these cells. Tandem-start usage also correlated with internal intron usage (Supplementary Fig. 18), but these correlations do not establish full-length isoform structures or coordinated regulation on individual RNA molecules. The assay therefore supports a functional difference between alternative *SCARB2* 5′ UTR sequences, which human genetic evidence links to a disease-associated initiation phenotype.

Discussion

Understanding how noncoding variants influence brain disorders requires molecular phenotypes that capture the diversity of gene regulation. We generated a large-scale resource for investigating alternative transcription initiation in the human brain. By integrating 7,710 RNA-sequencing profiles, we jointly characterized alternative first exon and tandem

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

transcription start site usage, revealing widespread ATI with distinct regional patterns. These complementary forms of initiation expand transcript diversity through alternative first-exon configurations and differences in the sequences retained at transcript 5′ends. Our work connects ATI with genetic variation, protein abundance and brain disease associations, providing a foundation for investigating the functional significance of transcript initiation in the nervous system.

The regional organization of ATI suggests a role in shaping the molecular diversity of the brain. Alternative starts can change the regulatory sequences retained in an RNA and distinguish transcripts with different coding configurations, providing potential routes to altered RNA stability and protein production. The enrichment of atiQTLs in promoters, 5′UTRs and TF binding annotations supports local regulation of initiation, whereas RBP-binding enrichment suggests connections with post-transcriptional processes. Incorporating multiple ancestries improved fine-mapping resolution, illustrating the value of population diversity for resolving regulatory signals.

Integration with GWAS prioritized ATI-linked candidates across brain-related diseases and traits, many of which were not recovered by conventional expression analysis. These results provide specific transcript features for experimental investigation at disease-associated loci. Shared ATI–protein genetic signals further suggest potential consequences for protein abundance, with heterogeneous effect directions consistent with sequence-dependent relationships involving RNA stability, translation or isoform composition.

The *SCARB2* locus illustrates how genetic prioritization can guide functional investigation. Alternative transcript starts were confirmed experimentally, and the short 5′ UTR produced higher reporter output than the long form in neural progenitor cells. Sequence perturbations partially relieved repression by the long UTR, implicating regulatory features within the extended region. Together with the genetic association at the PD locus, these findings motivate investigation of how endogenous start-site selection affects *SCARB2* regulation and lysosomal function.

This study has several limitations that should be acknowledged. First, the sample size in some brain regions may limit the statistical power to detect atiQTLs with small or region-specific effects, hindering downstream inference. Large samples from underrepresented regions and ancestries are needed. Second, bulk tissue likely underestimated cell-type specific regulatory effects. Public single-cell RNA-seq datasets were based on 3′ end enriched protocols with limited 5′ coverage, and deconvolution reference profiles inherit

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

the same bias, limiting resolution of 5′ end dynamics. Third, short read sequencing does not establish complete transcript structures. Fourth, external protein datasets differ in tissue context and lack matched RNA measurements. Future studies combining cell-resolved 5′-end profiling, long-read sequencing and endogenous perturbation will help connect genetic effects on initiation to downstream molecular consequences. Together, this work establishes a resource for studying transcript initiation as a complementary component of brain genetic regulation and for developing more specific hypotheses about the molecular consequences of disease-associated variation.

Methods

Data collection and RNA sequencing quality control

RNA-sequencing alignments and genotype data were obtained from AMP-AD[24] (MAYO, ROSMAP and MSBB), PsychENCODE[25], GTEx[59], NABEC[29] and ChinaBrain[30]. PsychENCODE sources included BipSeq, BrainGVEX, CMC, CMC_HBCC, UCLA_ASD and YALE_ASD. GTEx and NABEC accessions were phs000424.v8.p2 and phs001300.v4.p1, respectively. ChinaBrain[30] data were obtained from Central South University. Samples were grouped by brain region (amygdala, basal ganglia, cerebellum, cortex, hippocampus and hypothalamus), with spinal cord analyzed as an additional tissue. Where replicate samples were available, we selected the sample with the highest RNA integrity number. RNA alignment metrics were assessed using Picard CollectRnaSeqMetrics v3.1.1 and samtools flagstat v1.10. Samples were excluded if fewer than 10% of aligned reads mapped to coding bases and the overall alignment rate was below 60% (Supplementary Fig. 2).

Quantification of alternative transcription initiation

The tandem-TSS reference combined GENCODE v44, RefSeq v200 and FANTOM5 CAGE annotations. Transcripts with the same first-exon end coordinate were grouped, and the most upstream annotated start on the corresponding strand defined the nonredundant first-exon region. Tandem TSSs had to lie on the same strand as the transcript and within 500 nucleotides upstream of the first exon. Transcripts with at least two TSSs in this region were classified as harboring tandem TSSs. We used DATTSS[17] to estimate contributions from distal- and proximal-start transcripts based on RNA-sequencing coverage extracted with

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

pyBigWig v0.3.22. DTUI was defined as distal-transcript abundance divided by the combined distal- and proximal-transcript abundance. Values were retained when mean 5′-UTR read coverage was at least 10, and observations below this threshold were treated as missing.

We quantified AFE activity with proActiv[31] using first-exon junction counts at annotated, CAGE-supported promoters. Single-exon genes and first exons overlapping internal exons were excluded. Absolute activity was derived from log2-transformed normalized junction counts, and relative activity represented each promoter's contribution to the gene's promoter activity. Within each region, the first exon with the highest mean activity was designated the major first exon. Other first exons were classified as minor if their mean activity was at least 0.25 and as inactive if it was lower.

Consequence of gene product analysis

To investigate the potential functional impact of ATI events, we compared the genomic regions associated with major and alternative events for the samples of each brain region. For AFE events, the major promoter was defined as the one with highest mean promoter activity across samples from a given brain region. The corresponding transcript region were then divided into sequences specific to the major promoter, sequences specific to the alternative promoter, and sequences shared by both. For ATTSS events, transcript regions were similarly classified as specific to the isoform with proximal TSS, isoform with distal TSSs and the shared regions. The functional composition, encompassing 5'UTR, exon, coding sequence (CDS) and 3'UTR of each region was then characterized using GENCODE. For each region, we recorded both the presence of these functional elements and the proportion of the total region they occupied.

Identification of the brain-region specific ATI events

Brain region specificity of ATI events was evaluated using all the normal samples from the GTEx cohort. For each brain region i, we fitted the following linear model:

$$R_{p,i} = \beta^{R}_{0,i,p} + \beta^{R}_{1,i,p} x_i + \varepsilon^{R}_{p,i}$$

Where $R_{p,l}$ denotes the relative usage of ATI event p in sample l and $x_{i,l}$ is an variable indicates whether sample l is from brain region i. Statistical significance was assessed using the t-statistic for the coefficients of the linear regression for the relative usage of ATTSS (distal TSS usage) and AFE (relative promoter usage). Multiple testing correction was

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

performed across all ATI events by using the Benjamini-Hochberg (BH) method. Events with a BH adjusted *P*-value≤0.05 were retained as a candidate tissue specific event.

To quantify the regional specificity, let S_i represent the set of samples of tissue i and $S_i^{'}$ denoting samples from all other brain regions. We calculated $\overline{R}_{p,S_i}$, the mean relative usage of ATI event p for the samples in the tissue i and $\overline{R}_{p,S_i^{'}}$, the mean across other regions.

Genotype processing and QC

Genotype data were generated using different platforms for the included cohorts, including whole-genome sequencing (AMP-AD, GTEx, and ChinaBrain), genotyping arrays (NABEC) and haplotype reference consortium-imputed genotypes (PsychENCODE).

To assign the genetic ancestry for each individual sample, a PCA approach with PLINK (v1.9) [60]were utilized by combining our overlapping SNPs with the genotype data from the 1KGP[61], and pruning with PLINK (using the command *–indep-pairwise 50 5 0.2*; These parameters correspond to a 50 SNP window, a five-SNP step size, and an r^2 threshold of 0.2, respectively). Ancestry labels were assigned by applying a *k*-nearest neighbors (k=7) to the first two PCs. No sample was excluded. Most samples were assigned EUR ancestry (n = 3,631 samples), whereas 309 and 561 samples were classified as AFR and EAS ancestry, respectively (Supplementary Fig. 3).

Since the different datasets have varying read depths, the resulting genotypes still have many missing values. Therefore, Michigan Imputation Server 2 (ref[62]) was used for imputing genotypes into the TOPMed [63] (version r3; hg38) reference panel (https://imputation.biodatacatalyst.nhlbi.nih.gov/) by chromosome and subsequently merged. Variant with imputation quality R^2 <0.3, Hardy-Weinberg equilibrium p value < 10^{-6}, and MAF <0.05 were excluded, yielding approximately 6 million autosomal SNPs.

atiQTL mapping and regional sharing

To minimize potential confounding arising from latent batch effects and other unobserved factors, we applied PCAForQTL [32] to determine the optimal number of molecular phenotype-derived PCs (Supplementary Figs 7-8). and included the top five genotype principal components (PCs) to account for population structure. For atiQTL mapping, we included the top five genotype PCs, optimal phenotype-derived PCs, chromosomal sex, age and diagnostic information as covariates. We defined *cis*-atiQTL as variants affecting nearby ATI events,

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

with a mapping window of 1Mb upstream and downstream of the 5′ UTR region. Association testing was performed using the linear regression model implemented in tensorQTL (v1.0.8). We utilized quantile-normalized ATI events usage (DTUI values for ATTSS events and Major promoter usage for AFE events) as the response variable. To correct for multiple testing, permutation analysis(n=1,000) was conducted to derive empirical P-values for each ATI event, which were subsequently adjusted to Storey's q-values using the *qvalue* R package (v2.28.0). ATI events with a q-value < 0.05 were considered significant.

To characterize the sharing and specificity of atiQTL effects across the brain regions and ancestry groups, mashr [64] was applied to lead *cis*-atiQTL across all brain region and ancestry available. We used the nominal P values of variants from atiQTLs for different brain region, and different ancestry specifically for Brain cortex region. For each ATI event, the variant with the most significant P value was defined as the lead *cis*-atiQTL, and the resulting lead variants were combined to construct the strong association set. A random set was generated by sampling 20,000 *cis*-atiQTL from nominal results. For each selected variant-ATI pair, the estimated effect size and its corresponding standard error from every condition were supplied to mashr. The posterior effect estimates were used to quantify QTL effect sizes, while the local false sign rate (LFSR) was used to assess statistical significance. Effects with an LFSR below 0.05 were considered significant.

Statistical fine-mapping of atiQTLs

The identified significant ATI-QTLs may not be the causal variants but rather tag with the true functional SNPs through linkage disequilibrium. To better estimate the independent associations and potential causal SNPs, the state-of-art fine-mapping method susieR (v0.14.21)[34,65] was applied, which implements the Iterative Bayesian Stepwise Selection (IBSS) framework. A maximal of three causal variants was allowed within each locus, with estimate_prior_variantce = TRUE. For each signal, a 90% credible set was then defined as the smallest collection of variants accounting for more than 90% of the posterior probability.

Furthermore, divergent LD structures between populations can be leveraged to enhance causal variant identification[66,67]. To this end, we applied SuSiEx[68] to conduct cross-ancestry fine-mapping for each pair of ancestries (EAS-EUR, EAS-AFR, and EUR-AFR), respectively. By integrating atiQTL summary statistics with population-matched LD references, SuSiEx estimates both shared and specific causal effects within defined genomic regions (e.g., 1 Mb around atiGenes). Specifically, within a given region, SuSiEx jointly

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

evaluate variants from both tested ancestries and prioritizes the variant with the highest PIP as the most likely causal variant in that region.

Enrichment of atiQTLs for Functional annotations

The genomic feature annotation was performed on all variants using Ensembl Variant Effect Predictor (VEP, v115.2), including enhancer, promoter, "CTCF binding site", "TF binding site", "3 prime UTR variant", "5 prime UTR variant", "frameshift variant", "intron variant", "missense variant", "non coding transcript exon variant", "splice acceptor variant", "splice donor variant", "splice region variant", "stop gained" and "synonymous variant". Annotations for 16 human core promoter elements were retrieved from publicly available publication[69]. These comprised GAGA, BREu, TATA box, BREd, XCPE1, XCPE2, Ohler Motif 1, dTCT, hTCT, BBCABW, hInr, dInr, MTE, bridge, DPE and Pause button. In addition, the 15 chromatin states from the brain-related samples were obtained from the Roadmap Epigenomics Mapping Consortium (REMC) [70](https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html; https://roadmapepigenomics.s3.us-west-2.amazonaws.com/index.html?prefix=hg19/uniformly_processed/data/byFileType/chromhmmSegmentations/ChmmModels/coreMarks/jointModel/final/). We evaluated the enrichment of atiQTLs across these annotations using Torus[71]. Torus fits a hierarchical logistic regression to jointly evaluate multiple annotations and estimate an enrichment coefficient for each annotation category, which reflect the increased likelihood that atiQTL variants reside in a given . The public available uORF annotation were downloaded from uORFdb[72]， Phase I consensus sets of Ribo-seq ORFs from GENCODE[73].

Characterizing RNA features and TF binding associated with atiQTLs

The curated functional annotations comprise RBP binding sites determined through enhanced cross-linking immunoprecipitation (eCLIP) in K562 and HepG2 cells from ENCODE[74], ChIP-seq peaks for TFs and histone modifications (in LCLs) from ENCODE[75], experimentally determined RiboSNitch that affect RNA secondary structure by [76] and TargetScan predicted microRNA binding site[77]. For RBP eCLIP annotations, we used the reproducible and significantly enriched peaks of eCLIP data determined from irreproducible discovery rate (IDR) approach[78]. Detailed annotations for RBP functions were downloaded from ref[78].

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

We used the independent SNPs from the fine-mapped atiQTLs by selecting SNPs with the highest PIP within the 95% credible set. To generate the random control SNP set, each variant was matched to randomly sampled control variants across 100 sets from the same chromosome, allele frequency, numbers of SNPs in LD, distance to the nearest genes, gene density and type of genomic region (5'UTR, CDS, 3'UTR, intron and intergenic regions) using vSampler[79]. We then calculated the enrichment of QTLs overlapping with RNA feature or TF-binding annotations compared to the control SNPs using the two-tailed Fisher's exact test. A pseudo count of one was added to avoid division by zero errors.

To evaluate the relative contributions of RNA features (RBP binding, secondary structure and microRNA binding) and transcriptional features (TF binding) to ATI regulation, we estimated the fraction of putative causal atiQTLs falling into each annotation class.

Assessing the effect of atiQTLs on TF binding

TF binding motifs within atiQTL associated promoter regions were identified using Fimo (v5.5.5)[80]. The potential effect of atiQTL variants on TF binding were subsequently predicted with motifBreakR (v 2.27.2)[81], which identifies allelic changes likely to alter sequence-specific TF binding affinity.

Curation of GWAS summary statistics for brain-related traits

GWAS summary statistics for 18 brain-related diseases and traits were curated from publicly available sources (as detailed in Supplementary Data 2). The trait set included Alzheimer's disease (AD)[82,83], Parkinson's disease (PD)[58], amyotrophic lateral sclerosis (ALS)[84], schizophrenia (SCZ)[85,86], bipolar disorder[87,88], autism spectrum disorder[89], attention-deficit/hyperactivity disorder (ADHD)[90], anxiety[91], depression[91], major depressive disorder (MDD)[92,93], post-traumatic stress disorder (PTSD)[94], obsessive–compulsive disorder (OCD)[95], cross-disorder analyses[96,97], intelligence[98], insomnia[99], neuroticism[100], smoking initiation[101] and age at menarche[102].

Heritability enrichment analysis

atiQTL enrichment in the 18 brain-related traits was estimated using stratified LD score regression [35,36] (S-LDSC,v1.0.1) with the 53 other functional annotations in the baseline-LD model [36] (v1.2) in a window of ±100 kb around each atiGene with at least one independent *cis*-atiQTL. Ancestry-matched LD scores were computed respectively for the European and

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

East Asian populations from the 1KGP (phase 3) [103]with a 1 cM window. For each annotation category, heritability enrichment was defined as the proportion of SNP heritability attributable to that category relative to the proportion of SNPs assigned to it.

Colocalization analysis

Genome-wide significant loci were extracted from GWAS summary statistics data using PLINK (v1.9)[60]. We applied the *–clump* function to identify independent variants, utilizing the following parameters: *--clump-r2* 0.6 *--clump-p1* 5e-8 *--clump-p2* 0.05 *–clump-kb* 1000). Specifically, independent loci were identified using PLINK with a LD threshold of r^2<0.6. Adjacent loci separated by less than 250kb were merged into a single region to consolidate proximal signals and minimize the fragmentation of potentially shared genetic effects. For each GWAS locus, colocalization analysis was performed using the *coloc.abf* function in the COLOC (v5.2.3) [37]and SuSiE-coloc [38] to evaluate whether the GWAS and atiQTL signals were consistent with a shared underlying genetic variant. A posterior probability greater than 0.75 for hypothesis 4 (PP4), which represents a shared signal between the two traits tested, was considered evidence of colocalization.

SMR analysis

Summary-data-based Mendelian Randomization (SMR) analysis [39] was performed using the LD reference panel derived from the 1KGP. The lead atiQTL were use as the instrumental variables when its association reached the threshold $P < 5 \times 10^{-8}$. Additionally, GWAS summary statistics for 18 brain disorders (Supplementary Data 2) were used in the analysis. To provide a more comprehensive assessment, we also conducted the multi-SNP SMR analysis, incorporating eligible variants within 500-kb of each lead atiQTL after LD pruning at $r^2 > 0.1$. The HEIDI (heterogeneity in dependent instruments) test was used to evaluate whether the observed SMR association was more consistent with a shared genetic variant than with two distinct variants in LD. HEIDI test utilized between three and 20 *cis*-atiQTL SNPs ($P_{\text{atiQTL}} < 1.57 \times 10^{-3}$) with LD bounds of 0.05–0.9. In each brain region, SMR *P*-values were corrected for multiple testing using the Bonferroni method, with significance defined as $P_{\text{SMR}} < 0.05/m$, where m is the total number of tests performed. Associations with a P_{HEIDI}<0.01 were excluded because they show evidence of heterogeneity. Additionally, significant SMR associations with $P_{GWAS} > 1\times 10^{-5}$ were removed.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

Transcriptome-wide association analysis

We trained FUSION[40] models using *cis*-variants within 1 Mb of ATI events, with sex, age, genotype principal components and phenotype principal components as covariates. Candidate models were BLUP, elastic net, LASSO and the single best variant. Five-fold cross-validation selected the model with the highest out-of-sample R^2, retaining events meeting the nominal heritability criterion $P < 0.05$. This yielded 2,901 prediction models. Normalized predictive performance was defined as R^2 divided by *cis*-heritability. The extended MHC interval, chr6:28,510,120–33,480,577 in hg38, was excluded.

With cTWAS[42], we modeled predicted ATI and nearby variants jointly using GWAS statistics and LD reference data. The European reference comprised 1,700 LD blocks and 8,942,640 variants from 1000 Genomes. Prediction weights and GWAS *Z* scores were harmonized to the LD reference using preprocess_weight and preprocess_z_snp. Multiallelic variants and variants with ambiguous effect directions were excluded. Missing GWAS variants were imputed using the LD reference. We used PIP > 0.8 for prioritization.

Association score of gene in 18 brain-related diseases

To evaluate the gene-disease associations, we utilized the Open Targets Platform[46] database (https://platform.opentargets.org/, release 26.06) which is a valuable open resource for identifying novel targets. We focused on aforementioned 18 brain-related diseases, for which the number of associated genes ranged from 543 to 7,359. The Open Targets Platform provides comprehensive gene-disease association scores (ranging from 0 to 1), representing the strength of evidence across multiple domains. These scores integrate diverse evidence types, including genetics, genomics, transcriptomics, pharmacology, animal models, and scientific literature, providing a robust resource for identifying and prioritizing novel therapeutic targets.

pQTL data source

To assess downstream propagation of atiQTL effects, we collected five genome-proteome-wide association studies for genetic variant- protein target associations. These included plasma proteomics studies [47-49] (obtained from Synapse: https://www.synapse.org/#!Synapse:syn51824537 and https://www.synapse.org/#!Synapse:syn51365303, and GWAS catalog with Accession ID: GCST90274758 to GCST90274848) and brain/CSFs [50,51] (obtained from Synapse:

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

https://www.synapse.org/#!Synapse:syn64600176 and https://www.niagads.org/datasets/ng00102). Brain eQTLs from MetaBrain [6] were also included for comparative analysis. We defined *cis*-pQTLs as variants located within a 1Mb window of the gene TSS and restricted the analysis to variants present in our genotype data. Following an assessment of Pearson correlation of *P*-values for overlapping variant-protein pairs across datasets and removal of redundant targets, we ultimately identified 8,799 genes with proteo-genomic summary statistics.

Cross-molecular-trait integration analysis

Statistical colocalization was performed to assess whether protein abundance and expression or ATI levels shared common genetic signals. For each protein sentinel *cis*-pQTL, results from corresponding ± 500kb locus were extracted and matched to the identical genomic window in the *cis*-atiQTL dataset. Colocalization between *cis*-pQTLs and atiQTLs was considered supported if it met following criteria: PP4 > 0.7, *P*-value < 5×10^{-5} in pQTLs, and FDR < 0.05 in atiQTLs. We further investigated the potential convergence of genetic effects on protein abundance and disease risk for 18 brain-related GWAS data (Supplementary Data 2) and 8,799 proteo-genomic summary data. We conducted multi-trait colocalization analysis to detect shared genetic regulation across disease, and transcriptional phenotypes. A posterior probability of PP4 > 0.7 was used as evidence for shared genetic associations.

Drug-gene interaction analysis

We compiled a comprehensive list of approved and potentially druggable targets from two public resources: the Therapeutic Target Database (TTD; https://db.idrblab.net/ttd) [54] and the druggable genome[55]. The TTD encompasses 3,395 genes, including 426 approved targets, 1,014 clinical-stage targets, 212 preclinical or patented targets, and 1,479 literature-reported targets. Additionally, the druggable genome comprises 4,479 genes classified into three tiers: **Tier 1** contains 1,427 genes corresponding to approved drugs or clinical-phase candidates; **Tier 2** includes 682 genes encoding targets with drug-like compounds that share ⩾ 50% identity with approved drug targets; and **Tier 3** consists of 2,370 potentially druggable genes exhibiting lower similarity to established drug targets. Candidate genes were additionally queried against the Drug Gene Interaction database (DGIdb v5.0, https://dgidb.org)[56] to identify reported drug-gene relationships and druggability evidence collected from multiple publications and databases. Drugs with the identified genes were subsequently assigned

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

Anatomical Therapeutic Chemical (ATC) codes based on the drug information available in the KEGG drug section (https://www.genome.jp/kegg/drug/).

5′RACE PCR

5′ rapid amplification of cDNA ends was performed in SH-SY5Y cells to identify endogenous TSS isoforms of *SCARB2*. Products were resolved by gel electrophoresis and sequenced. The primers used for 5'-RACE are listed in Supplementary Data 4**.**

mRNA stability assay

SH-SY5Y cells were treated with actinomycin D at a final concentration of 5 μg/ml. RNA was harvested at serial time points and quantified by qRT-PCR to assess mRNA half-life differences between long and short 5′UTR isoforms. The primers used for qRT-PCR are listed in Supplementary Data 4**.**

Dual luciferase reporter assay

A dual-luciferase reporter system was used to evaluate the effect of rs11547135 on tandem TSS usage. *SCARB2* 5'UTR fragments of different lengths harboring either allele of rs11547135 were placed upstream of the firefly luciferase coding sequence. M6-derived neural progenitor cells were then transfected with the corresponding reporter plasmids. Luciferase signals were detected 48 h after transfection, and firefly activity was adjusted relative to Renilla activity. Differences between the allelic constructs were analyzed using an unpaired two-tailed Student's t-test.

Data availability

Raw GTEx project V8 data used in this study are available through the Database of Genotypes and Phenotypes (dbGaP, accession code: phs000424.v7.p2). Data from CMC (Synapse accession syn2759792), CMC_HBCC (syn10623034), AMP-AD (syn2580853), BrainSeq (syn12299750), UCLA_ASD (syn4587609), BrainGVEx (syn4590909) and BipSeq (syn5844980) were obtained from Synapse (https://www.synapse.org/). NABEC data are available through dbGaP (accession code: phs001301.v1.p1). Raw sequencing data from ChinaBrain are available through the Genome Sequence Archive–Human at the Beijing Institute of Genomics, Chinese Academy of Sciences (https://bigd.big.ac.cn/gsa-human,

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

accession code: HRA000108). GWAS summary statistics are provided in Supplementary Data 2.

Acknowledgments

We thank the members of the Li laboratory for helpful discussions and the Shenzhen Bay Laboratory High-Performance Computing and Informatics Core Facility for high-performance computing support.

Funding Statement

This work was supported by the Shenzhen Medical Research Fund (grant no. C2503001 to L.L.), the Major Program of Shenzhen Bay Laboratory (grant no. C1012524001 to L.L.), the National Natural Science Foundation of China (grant no. 32370721, 32100533 to L.L., grant no. 32570730 to X.Z.), and the Shenzhen Clinical Research Center for Oral Diseases (grant no. 20210617170745001 to X.Z.).

Author Contributions Statement

L.L. conceived and designed the study. L.L. and X.Z. acquired funding. L.L. was responsible for project administration. H.C. curated the data and performed bioinformatic analyses with help from X.M., X.L. and D.L., and T.S. conducted the experiments. H.C. prepared all figures with help from T.Z, X.L. and X.Z. X.Z. and H.C. wrote the original draft of the manuscript; M.P., X. Z., C.C. and L.L. reviewed and commented on the manuscript; L.L., X.Z., and H.C. revised the manuscript; all authors contributed to review and editing.

Competing Interests Statement

The authors declare no competing interests.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

Figure legends

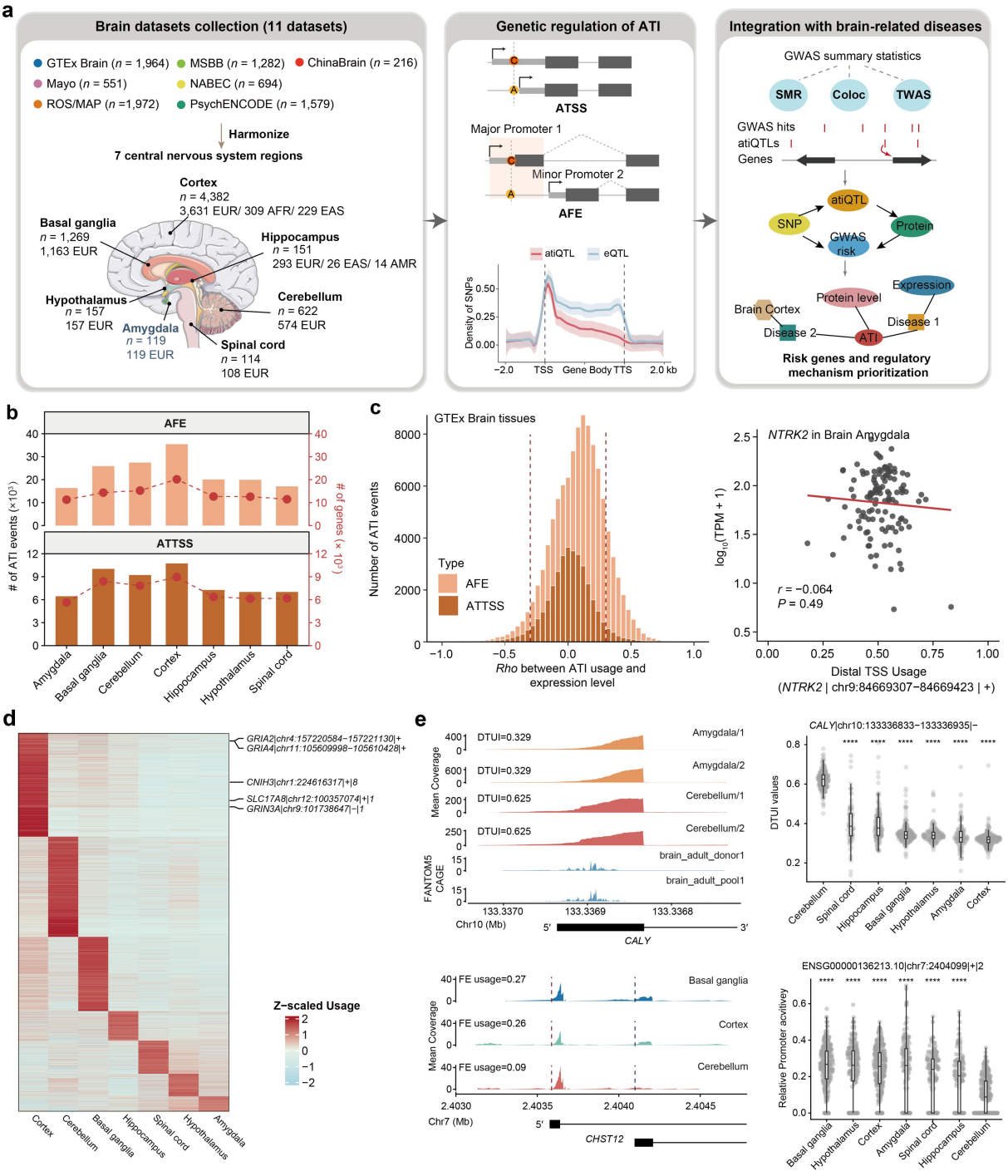

Fig. 1 Regional transcript initiation captures information beyond gene expression

a, Study design and definitions of alternative first exons (AFE) and alternative tandem transcription start sites (ATTSS), based on RNA-sequencing data from 7,710 samples spanning brain regions and spinal cord. **b,** Numbers of quantified ATI events and associated genes by region and ATI class. **c,** Distributions of Spearman correlations between ATI usage and total gene expression in GTEx brain tissues. The absolute-correlation threshold of 0.3

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

indicates effect magnitude, not statistical significance. **d,** Mean relative usage of region-associated ATI events. **e,f,** RNA-sequencing coverage and CAGE support at *CALY* (e) and distal TSS usage across brain regions (f). **g,h,** First-exon junction coverage at *CHST12* (g) and relative activity of the alternative first exon across regions (h). Regional associations were tested using linear models with Benjamini–Hochberg correction. Regional comparisons in f and h are labeled ****P ≤ 0.0001.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

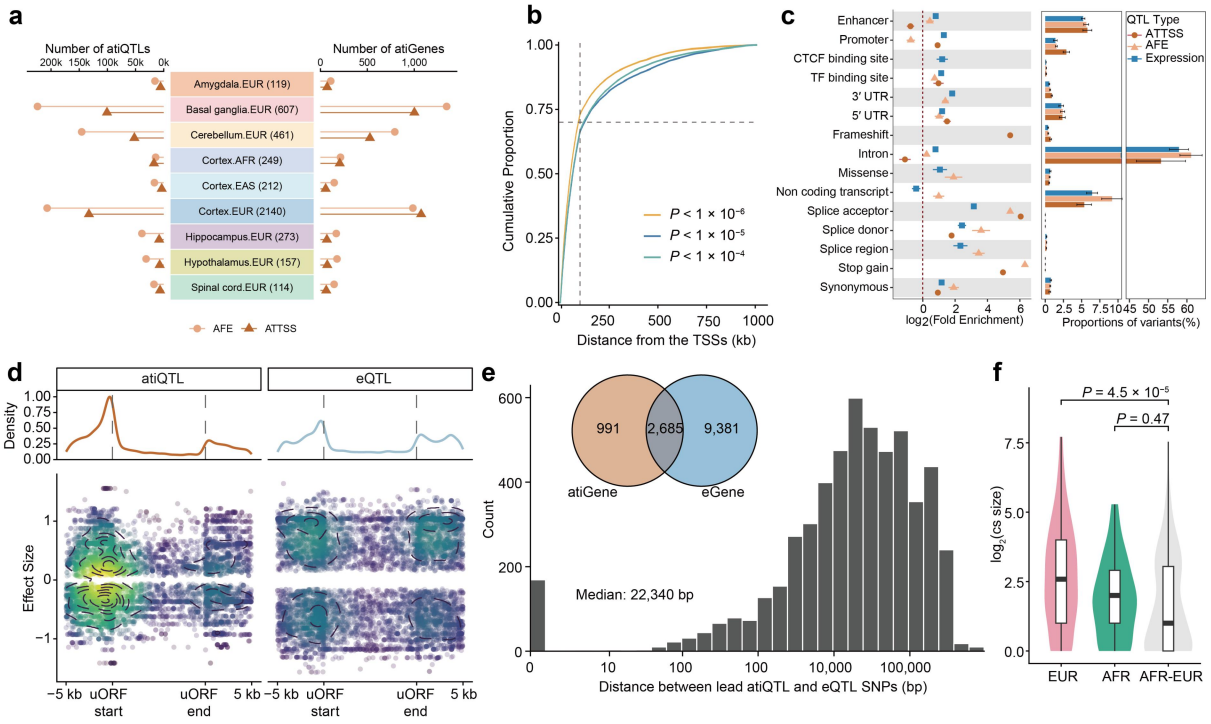

Fig. 2 Identification and characterization of atiQTL

a, Numbers of genes with significant cis-atiQTL associations by region and ATI class (q < 0.05). **b,** Cumulative distribution of variant distances from the associated TSS at different association thresholds. **c,** Enrichment of ATI QTLs and eQTLs in functional genomic annotations, shown as mean ± s.d. across regions. **d,** Overlap of genes associated with ATI QTLs and eQTLs. **e,** Distance between lead atiQTL and eQTL variants for genes detected by both analyses. **f,** Distribution of the size of 95% credible sets (CSs) and comparisons between the cross-ancestry fine-mapping and single-population fine-mapping. Shared atiGenes among the three groups and single-CS genes were used for comparison. *P* values are two-tailed and based on the Wilcoxon signed-rank test.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

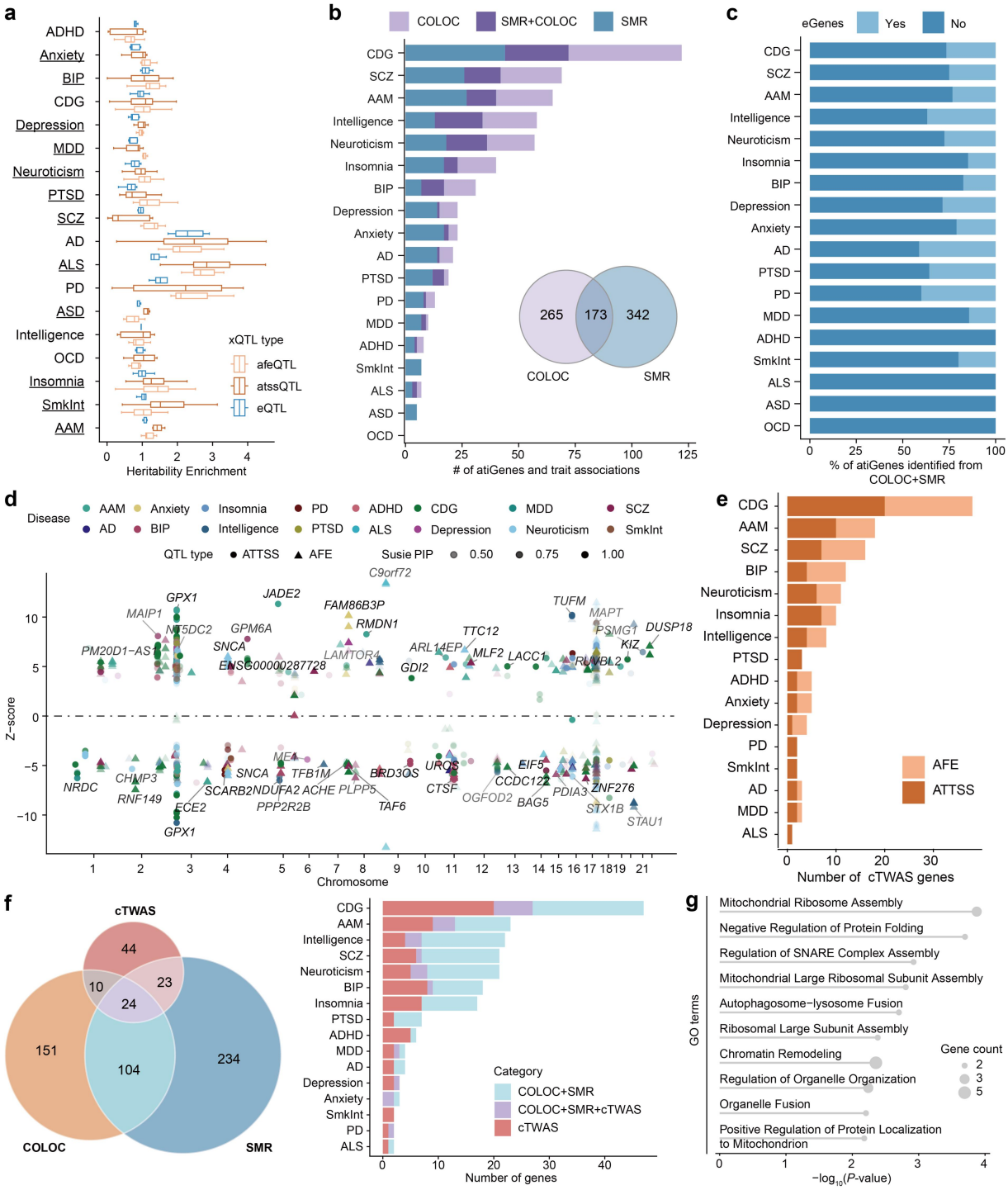

Fig. 3 Transcript initiation links genetic variation to brain-related diseases and traits

a, Trait-heritability enrichment for ATI-QTL and eQTL annotations. Stratified LD score regression conditions on baseline annotations, and the separate enrichment tests do not assess the difference between ATI and expression estimates. **b,** Trait-associated ATI nominations and overlap between colocalization and SMR. **c,** Proportion of jointly nominated genes also detected as eGenes. **d,** Genome-wide distribution of Z-score values from atiTWAS across 18

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

brain-related diseases and traits. Top-associated genes for each chromosome are labeled. Only the genes achieving a PIP threshold greater than 0.8 were prioritized as significant and retained for visualization. **e,** Total number of susceptibility genes identified by atiTWAS for each brain-related trait. **f,** (Left) Venn diagram showing the overlap of susceptibility genes supported by colocalization, SMR, and atiTWAS analyses; (Right) Overlap of susceptibility genes identified by cTWAS and those supported by both colocalization and SMR per trait. **g,** GO biological process enrichment for genes identified from atiTWAS, and the dot size corresponds to the gene count.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

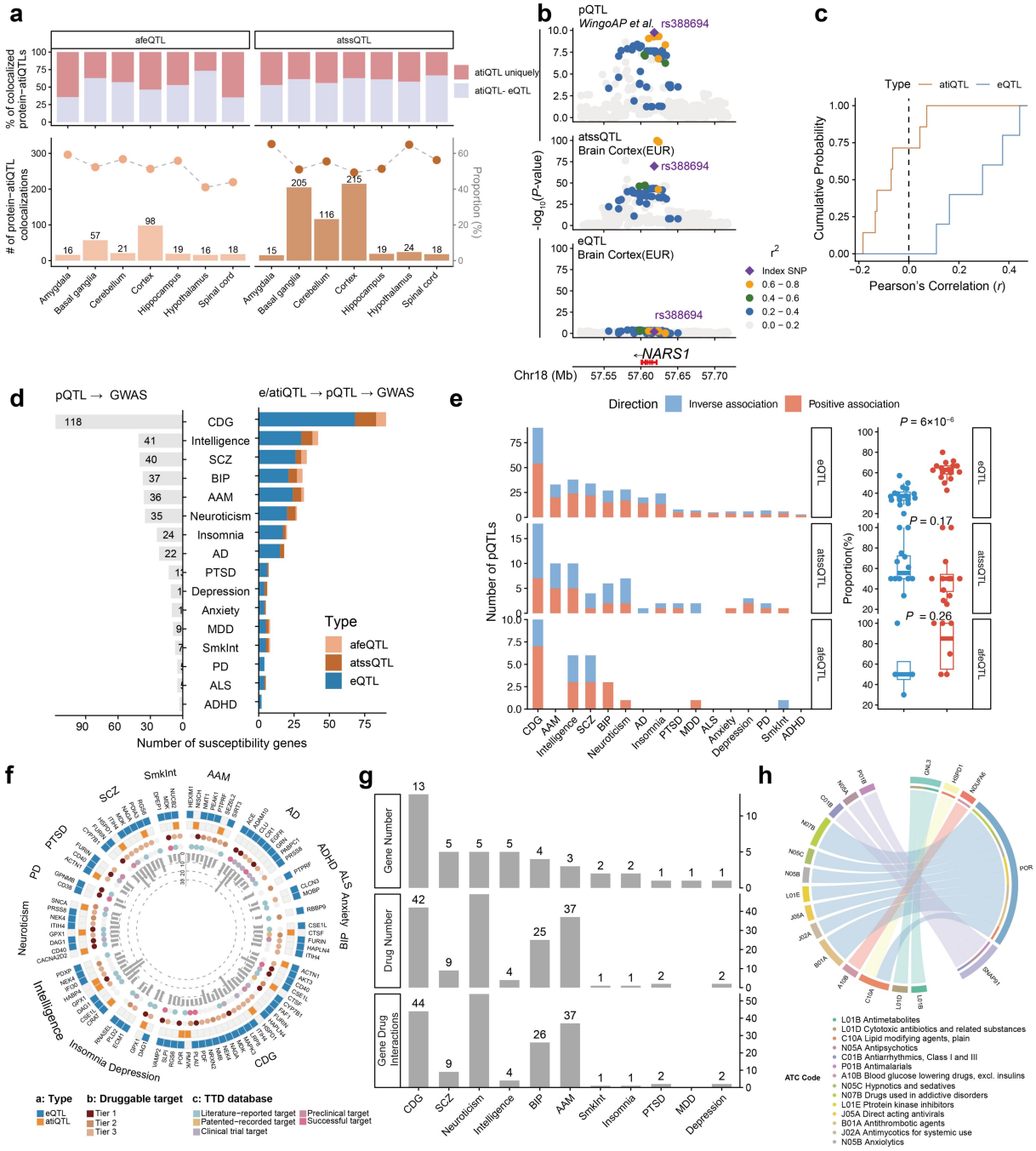

Fig. 4 Protein associations provide complementary evidence for ATI-linked loci

a, Protein-QTL sources and their overlap across brain, CSF and plasma, covering 8,799 genes. **b,** ATI–pQTL colocalization counts and proportions by brain region and ATI class, with the fraction also supported by the expression comparator. Pairwise colocalization required PP4 > 0.7, pQTL $P < 5 \times 10^{-5}$ and ATI FDR < 0.05. **c,** Regional association patterns for *NARS1* protein abundance, cortex ATTSS usage and cortex total expression. ATI–protein PP4 = 0.994. **d,** Protein–trait and RNA–protein–trait colocalization summaries, distinguishing ATI and expression. Protein measurements are distinguished by source compartment.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

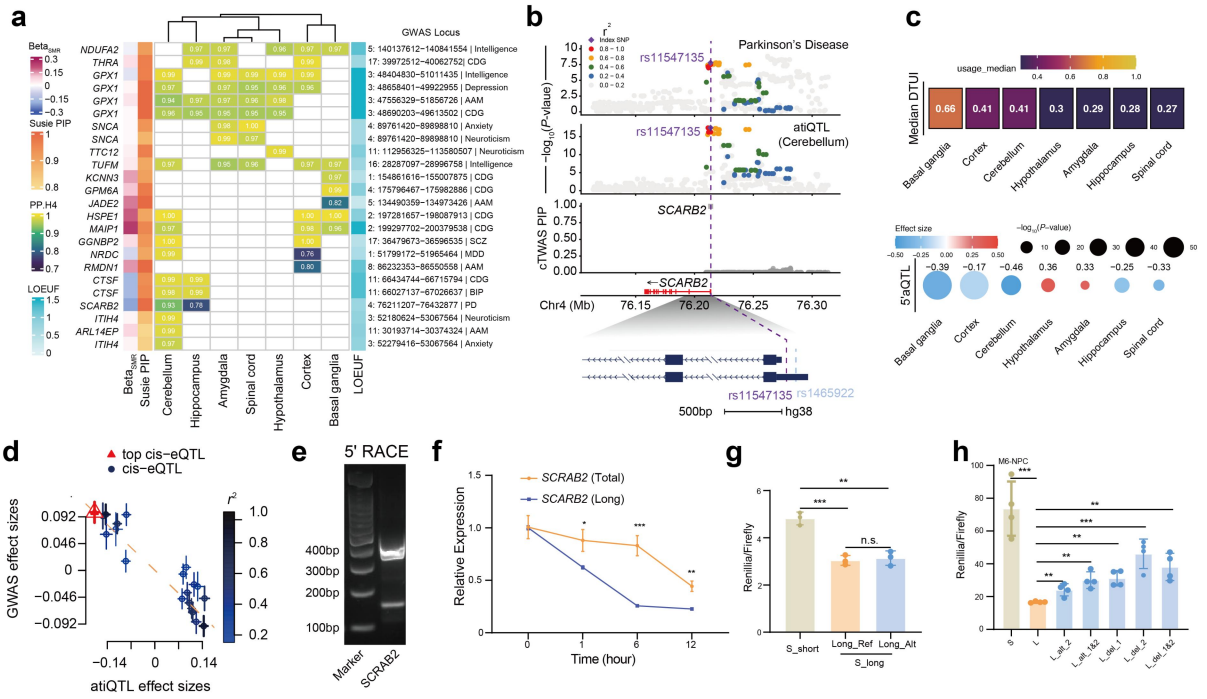

Fig. 5 Alternative *SCARB2* 5′ UTR sequences alter reporter output at a PD-associated locus

a, Regional colocalization evidence for candidates supported by all three approaches, including *SCARB2*. Colocalization required PP4 > 0.75, and SMR used regional Bonferroni correction and HEIDI filtering. **b,** Regional PD GWAS and *SCARB2* ATI association plots with cTWAS prioritization and transcript-start annotations. Colors indicate LD with rs11547135. **c,** Regional *SCARB2* distal-start usage and ATI-QTL effect estimates. The cerebellar effect estimate for rs11547135 is −0.46 ($P = 3.29 \times 10^{-18}$). **d,** PD and *SCARB2* ATI effect estimates for variants used in the HEIDI analysis. **e,** 5′ RACE products from SH-SY5Y cells supporting two *SCARB2* transcript starts, with product sequences confirmed by Sanger sequencing. **f,** Long-transcript and total *SCARB2* RNA measurements at 0, 1, 6 and 12 h after actinomycin D treatment (5 μg ml^{-1}). This assay does not directly quantify short-transcript half-life. **g,** Reporter output in M6-derived neural progenitor cells for short, long and altered long 5′-UTR constructs, measured 48 h after transfection. Renilla/firefly activity is shown as mean ± s.d.; n = 4 independent experiments; two-sided paired t-tests. Short versus long, ***P < 0.001; other displayed comparisons are indicated by **P < 0.01 or ***P < 0.001.

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

References

1. Brainstorm, C. *et al.* Analysis of shared heritability in common disorders of the brain. *Science* **360**(2018).
2. Blauwendraat, C., Nalls, M.A. & Singleton, A.B. The genetic architecture of Parkinson's disease. *Lancet Neurol* **19**, 170–178 (2020).
3. Ridge, P.G. *et al.* Assessment of the genetic variance of late-onset Alzheimer's disease. *Neurobiol Aging* **41**, 200 e13–200 e20 (2016).
4. Romero, C. *et al.* Exploring the genetic overlap between twelve psychiatric disorders. *Nat Genet* **54**, 1795–1802 (2022).
5. Grotzinger, A.D. *et al.* Genetic architecture of 11 major psychiatric disorders at biobehavioral, functional genomic and molecular genetic levels of analysis. *Nat Genet* **54**, 548–559 (2022).
6. de Klein, N. *et al.* Brain expression quantitative trait locus and network analyses reveal downstream effects and putative drivers for brain-related diseases. *Nat Genet* **55**, 377–388 (2023).
7. Bryois, J. *et al.* Cell-type-specific cis-eQTLs in eight human brain cell types identify novel risk genes for psychiatric and neurological disorders. *Nat Neurosci* **25**, 1104–1112 (2022).
8. Wang, D. *et al.* Comprehensive functional genomic resource and integrative model for the human brain. *Science* **362**(2018).
9. Li, L. *et al.* An atlas of alternative polyadenylation quantitative trait loci contributing to complex trait and disease heritability. *Nat Genet* **53**, 994–1005 (2021).
10. Davuluri, R.V., Suzuki, Y., Sugano, S., Plass, C. & Huang, T.H. The functional consequences of alternative promoter use in mammalian genomes. *Trends Genet* **24**, 167–77 (2008).
11. Baek, D., Davis, C., Ewing, B., Gordon, D. & Green, P. Characterization and predictive discovery of evolutionarily conserved mammalian alternative promoters. *Genome Res* **17**, 145–55 (2007).
12. Alfonso-Gonzalez, C. & Hilgers, V. (Alternative) transcription start sites as regulators of RNA processing. *Trends Cell Biol* **34**, 1018–1028 (2024).
13. de Klerk, E. & t Hoen, P.A. Alternative mRNA transcription, processing, and translation: insights from RNA sequencing. *Trends Genet* **31**, 128–39 (2015).
14. Hinnebusch, A.G., Ivanov, I.P. & Sonenberg, N. Translational control by 5'-untranslated regions of eukaryotic mRNAs. *Science* **352**, 1413–6 (2016).
15. Leppek, K., Das, R. & Barna, M. Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them. *Nat Rev Mol Cell Biol* **19**, 158–174 (2018).
16. Yuan, J. *et al.* A compendium of genetic variations associated with promoter usage across 49 human tissues. *Nat Commun* **15**, 8758 (2024).
17. Zhao, Z. *et al.* Pan-cancer transcriptome analysis reveals widespread regulation through alternative tandem transcription initiation. *Sci Adv* **10**, eadl5606 (2024).
18. Zou, X. *et al.* Alternative tandem transcription initiation links noncoding variants to human disease through translational control. *Nat Commun* **17**(2026).
19. Haigh, J.L. *et al.* Deletion of a non-canonical regulatory sequence causes loss of Scn1a expression and epileptic phenotypes in mice. *Genome Med* **13**, 69 (2021).
20. Pruunsild, P., Kazantseva, A., Aid, T., Palm, K. & Timmusk, T. Dissecting the human BDNF locus: bidirectional transcription, complex splicing, and multiple promoters. *Genomics* **90**, 397–406 (2007).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

21. Dong, P. *et al.* A multi-regional human brain atlas of chromatin accessibility and gene expression facilitates promoter-isoform resolution genetic fine-mapping. *Nat Commun* **15**, 10113 (2024).
22. Wen, C. *et al.* Cross-ancestry atlas of gene, isoform, and splicing regulation in the developing human brain. *Science* **384**, eadh0829 (2024).
23. Raj, T. *et al.* Integrative transcriptome analyses of the aging brain implicate altered splicing in Alzheimer's disease susceptibility. *Nat Genet* **50**, 1584–1592 (2018).
24. Hodes, R.J. & Buckholtz, N. Accelerating Medicines Partnership: Alzheimer's Disease (AMP-AD) Knowledge Portal Aids Alzheimer's Drug Discovery through Open Data Sharing. *Expert Opin Ther Targets* **20**, 389–91 (2016).
25. Psych, E.C. Revealing the brain's molecular architecture. *Science* **362**, 1262–1263 (2018).
26. BrainSeq, A.H.B.G.C.E.a.d.l.o. & BrainSeq, A.H.B.G.C. BrainSeq: Neurogenomics to Drive Novel Target Discovery for Neuropsychiatric Disorders. *Neuron* **88**, 1078–1083 (2015).
27. Fromer, M. *et al.* Gene expression elucidates functional impact of polygenic risk for schizophrenia. *Nat Neurosci* **19**, 1442–1453 (2016).
28. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
29. Gibbs, J.R. *et al.* Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet* **6**, e1000952 (2010).
30. Chen, Y. *et al.* Cross-ancestry analysis of brain QTLs enhances interpretation of schizophrenia genome-wide association studies. *Am J Hum Genet* **111**, 2444–2457 (2024).
31. Demircioglu, D. *et al.* A Pan-cancer Transcriptome Analysis Reveals Pervasive Regulation through Alternative Promoters. *Cell* **178**, 1465–1477 e17 (2019).
32. Zhou, H.J., Li, L., Li, Y., Li, W. & Li, J.J. PCA outperforms popular hidden variable inference methods for molecular QTL mapping. *Genome Biol* **23**, 210 (2022).
33. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76–82 (2011).
34. Wang, G., Sarkar, A., Carbonetto, P. & Stephens, M. A simple new approach to variable selection in regression, with application to genetic fine mapping. *J R Stat Soc Series B Stat Methodol* **82**, 1273–1300 (2020).
35. Gazal, S. *et al.* Linkage disequilibrium-dependent architecture of human complex traits shows action of negative selection. *Nat Genet* **49**, 1421–1427 (2017).
36. Finucane, H.K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat Genet* **47**, 1228–35 (2015).
37. Giambartolomei, C. *et al.* Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genet* **10**, e1004383 (2014).
38. Wallace, C. A more accurate method for colocalisation analysis allowing for multiple causal variants. *PLoS Genet* **17**, e1009440 (2021).
39. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481–7 (2016).
40. Gusev, A. *et al.* Integrative approaches for large-scale transcriptome-wide association studies. *Nat Genet* **48**, 245–52 (2016).
41. Wainberg, M. *et al.* Opportunities and challenges for transcriptome-wide association studies. *Nat Genet* **51**, 592–599 (2019).
42. Zhao, S. *et al.* Adjusting for genetic confounders in transcriptome-wide association studies improves discovery of risk genes of complex traits. *Nat Genet* **56**, 336–347 (2024).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

43. Tian, L. *et al.* iTRAQ-based quantitative proteomic analysis provides insight for molecular mechanism of neuroticism. *Clin Proteomics* **16**, 38 (2019).
44. Bayir, F.S. *et al.* Catatonia and cognitive decline in a patient with bipolar disorder: a diagnostic challenge leading to neuronal ceroid lipofuscinosis type 13. *Neurol Sci* **46**, 6139–6143 (2025).
45. Formoso, K., Garcia, M.D., Frasch, A.C. & Scorticati, C. Evidence for a role of glycoprotein M6a in dendritic spine formation and synaptogenesis. *Mol Cell Neurosci* **77**, 95–104 (2016).
46. Buniello, A. *et al.* Open Targets Platform: facilitating therapeutic hypotheses building in drug discovery. *Nucleic Acids Res* **53**, D1467–D1475 (2025).
47. Sun, B.B. *et al.* Plasma proteomic associations with genetics and health in the UK Biobank. *Nature* **622**, 329–338 (2023).
48. Pietzner, M. *et al.* Mapping the proteo-genomic convergence of human diseases. *Science* **374**, eabj1541 (2021).
49. Zhao, J.H. *et al.* Genetics of circulating inflammatory proteins identifies drivers of immune-mediated disease risk and therapeutic targets. *Nat Immunol* **24**, 1540–1551 (2023).
50. Yang, C. *et al.* Genomic atlas of the proteome from brain, CSF and plasma prioritizes proteins implicated in neurological disorders. *Nat Neurosci* **24**, 1302–1312 (2021).
51. Wingo, A.P. *et al.* Multiancestry brain pQTL fine-mapping and integration with genome-wide association studies of 21 neurologic and psychiatric conditions. *Nat Genet* **57**, 2156–2165 (2025).
52. Wang, L. *et al.* Loss of NARS1 impairs progenitor proliferation in cortical brain organoids and leads to microcephaly. *Nat Commun* **11**, 4038 (2020).
53. Foley, C.N. *et al.* A fast and efficient colocalization algorithm for identifying shared genetic risk factors across multiple traits. *Nat Commun* **12**, 764 (2021).
54. Zhou, Y. *et al.* TTD: Therapeutic Target Database describing target druggability information. *Nucleic Acids Res* **52**, D1465–D1477 (2024).
55. Finan, C. *et al.* The druggable genome and support for target identification and validation in drug development. *Sci Transl Med* **9**(2017).
56. Cannon, M. *et al.* DGIdb 5.0: rebuilding the drug-gene interaction database for precision medicine and drug discovery platforms. *Nucleic Acids Res* **52**, D1227–D1235 (2024).
57. Rothaug, M. *et al.* LIMP-2 expression is critical for beta-glucocerebrosidase activity and alpha-synuclein clearance. *Proc Natl Acad Sci U S A* **111**, 15573–8 (2014).
58. Nalls, M.A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol* **18**, 1091–1102 (2019).
59. Consortium, G.T. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
60. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* **4**, 7 (2015).
61. Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
62. Das, S. *et al.* Next-generation genotype imputation service and methods. *Nat Genet* **48**, 1284–1287 (2016).
63. Taliun, D. *et al.* Sequencing of 53,831 diverse genomes from the NHLBI TOPMed Program. *Nature* **590**, 290–299 (2021).
64. Urbut, S.M., Wang, G., Carbonetto, P. & Stephens, M. Flexible statistical methods for estimating and testing effects in genomic studies with multiple conditions. *Nat Genet* **51**, 187–195 (2019).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

65. Zou, Y., Carbonetto, P., Wang, G. & Stephens, M. Fine-mapping from summary data with the "Sum of Single Effects" model. *PLoS Genet* **18**, e1010299 (2022).
66. Schaid, D.J., Chen, W. & Larson, N.B. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nat Rev Genet* **19**, 491–504 (2018).
67. Asimit, J.L., Hatzikotoulas, K., McCarthy, M., Morris, A.P. & Zeggini, E. Trans-ethnic study design approaches for fine-mapping. *Eur J Hum Genet* **24**, 1330–6 (2016).
68. Yuan, K. *et al.* Fine-mapping across diverse ancestries drives the discovery of putative causal variants underlying human complex traits and diseases. *Nat Genet* **56**, 1841–1850 (2024).
69. Adato, O. *et al.* ElemeNT 2023: an enhanced tool for detection and curation of core promoter elements. *Bioinformatics* **40**(2024).
70. Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–30 (2015).
71. Wen, X. Molecular QTL discovery incorporating genomic annotations using Bayesian false discovery rate control. *Ann. Appl. Stat.* **10**, 1619–1638 (2016).
72. Wethmar, K., Barbosa-Silva, A., Andrade-Navarro, M.A. & Leutz, A. uORFdb--a comprehensive literature database on eukaryotic uORF biology. *Nucleic Acids Res* **42**, D60–7 (2014).
73. Mudge, J.M. *et al.* Standardized annotation of translated open reading frames. *Nat Biotechnol* **40**, 994–999 (2022).
74. Van Nostrand, E.L. *et al.* Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP). *Nat Methods* **13**, 508–14 (2016).
75. Consortium, E.P. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
76. Wan, Y. *et al.* Landscape and variation of RNA secondary structure across the human transcriptome. *Nature* **505**, 706–9 (2014).
77. Agarwal, V., Bell, G.W., Nam, J.W. & Bartel, D.P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**(2015).
78. Van Nostrand, E.L. *et al.* A large-scale binding and functional map of human RNA-binding proteins. *Nature* **583**, 711–719 (2020).
79. Huang, D. *et al.* vSampler: fast and annotation-based matched variant sampling tool. *Bioinformatics* **37**, 1915–1917 (2021).
80. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–8 (2011).
81. Coetzee, S.G., Coetzee, G.A. & Hazelett, D.J. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. *Bioinformatics* **31**, 3847–9 (2015).
82. Wightman, D.P. *et al.* A genome-wide association study with 1,126,563 individuals identifies new risk loci for Alzheimer's disease. *Nat Genet* **53**, 1276–1282 (2021).
83. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat Genet* **54**, 412–436 (2022).
84. van Rheenen, W. *et al.* Common and rare variant association analyses in amyotrophic lateral sclerosis identify 15 risk loci with distinct genetic architectures and neuron-specific biology. *Nat Genet* **53**, 1636–1648 (2021).
85. Trubetskoy, V. *et al.* Mapping genomic loci implicates genes and synaptic biology in schizophrenia. *Nature* **604**, 502–508 (2022).
86. Lam, M. *et al.* Comparative genetic architectures of schizophrenia in East Asian and European populations. *Nat Genet* **51**, 1670–1678 (2019).
87. O'Connell, K.S. *et al.* Genomics yields biological and phenotypic insights into bipolar disorder. *Nature* **639**, 968–975 (2025).

Langtaosha (LTS) Preprint doi: https://doi.org/10.65215/LTSpreprints.2026.11.24.000401. This version posted November 24, 2026. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. Creative Commons license: CC Attribution-NonCommercial-NoDerivatives 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0

88. Mullins, N. *et al.* Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology. *Nat Genet* **53**, 817–829 (2021).
89. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat Genet* **51**, 431–444 (2019).
90. Demontis, D. *et al.* Genome-wide analyses of ADHD identify 27 risk loci, refine the genetic architecture and implicate several cognitive domains. *Nat Genet* **55**, 198–208 (2023).
91. Thorp, J.G. *et al.* Symptom-level modelling unravels the shared genetic architecture of anxiety and depression. *Nat Hum Behav* **5**, 1432–1442 (2021).
92. Howard, D.M. *et al.* Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nat Neurosci* **22**, 343–352 (2019).
93. Giannakopoulou, O. *et al.* The Genetic Architecture of Depression in Individuals of East Asian Ancestry: A Genome-Wide Association Study. *JAMA Psychiatry* **78**, 1258–1269 (2021).
94. Nievergelt, C.M. *et al.* Genome-wide association analyses identify 95 risk loci and provide insights into the neurobiology of post-traumatic stress disorder. *Nat Genet* **56**, 792–808 (2024).
95. Strom, N.I. *et al.* Genome-wide analyses identify 30 loci associated with obsessive-compulsive disorder. *Nat Genet* **57**, 1389–1401 (2025).
96. Grotzinger, A.D. *et al.* Mapping the genetic landscape across 14 psychiatric disorders. *Nature* **649**, 406–415 (2026).
97. Bipolar, D., Schizophrenia Working Group of the Psychiatric Genomics Consortium. Electronic address, d.r.v.e., Bipolar, D. & Schizophrenia Working Group of the Psychiatric Genomics, C. Genomic Dissection of Bipolar Disorder and Schizophrenia, Including 28 Subphenotypes. *Cell* **173**, 1705–1715 e16 (2018).
98. Savage, J.E. *et al.* Genome-wide association meta-analysis in 269,867 individuals identifies new genetic and functional links to intelligence. *Nat Genet* **50**, 912–919 (2018).
99. Jansen, P.R. *et al.* Genome-wide analysis of insomnia in 1,331,010 individuals identifies new risk loci and functional pathways. *Nat Genet* **51**, 394–403 (2019).
100. Nagel, M. *et al.* Meta-analysis of genome-wide association studies for neuroticism in 449,484 individuals identifies novel genetic loci and pathways. *Nat Genet* **50**, 920–927 (2018).
101. 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. *Nat Genet* **51**, 237–244 (2019).
102. Day, F.R. *et al.* Genomic analyses identify hundreds of variants associated with age at menarche and support a role for puberty timing in cancer risk. *Nat Genet* **49**, 834–841 (2017).
103. Genomes Project, C. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).