Transcriptional Determinism and Stochasticity Contribute to the Complexity of Autism Associated *SHANK* Family Genes

Xiaona Lu[1], Pengyu Ni[2], Paola Suarez-Meade[6], Yu Ma[7], Emily Niemitz Forrest[1], Guilin Wang[5], Yi Wang[7], Alfredo Quiñones-Hinojosa[6], Mark Gerstein[2,5], Yong-hui Jiang[1,3,4]

Department of Genetics[1], Biomedical Informatics & Data Science[2], Neuroscienc[3], Pediatrics[4], & Yale Center for Genome Analysis[5], Yale University School of Medicine New Haven, CT, 06520 USA

Department of Neurosurgery[6], Mayo Clinic, Jacksonville, FL, 32224 USA

Department of Neurology[7], Children's Hospital of Fudan University, Shanghai, 201102 China

Correspondence:

Yong-hui Jiang, MD, PhD

Department of Genetics

Yale University School of Medicine

WWW 313, 333 Cedar Street,

New Haven CT 06520

Tel: 203 785 2429

Fax:203 785 3404

Email:yong-hui.jiang@yale.edu

Abstract

Precision of transcription is critical because transcriptional dysregulation is disease causing. Traditional methods of transcriptional profiling are inadequate to elucidate the full spectrum of the transcriptome, particularly for longer and less abundant mRNAs. *SHANK3* is one of the most common autism causative genes. Twenty-four *Shank3* mutant animal lines have been developed for autism modeling. However, their preclinical validity has been questioned due to incomplete *Shank3* transcript structure. We applied an integrative approach combining cDNA-capture and long-read sequencing to profile the *SHANK3* transcriptome in human and mice. We unexpectedly discovered an extremely complex *SHANK3* transcriptome. Specific *SHANK3* transcripts were altered in *Shank3* mutant mice and postmortem brains tissues from individuals with ASD. The enhanced *SHANK3* transcriptome significantly improved the detection rate for potential deleterious variants from genomics studies of neuropsychiatric disorders. Our findings suggest the stochastic transcription of genome associated with *SHANK* family genes.

Introduction

In the central dogma of molecular biology, RNA transcription acts as a rheostat, orchestrating the cellular functions of the genes in response to intrinsic and extrinsic signals. The complex functions in the organs such as brain require a diverse proteome from a relatively small gene pool. This diversity is facilitated by transcriptional regulations involving alternative promoter usage and splicing, occurring in >90% of neuronal genes in mammalian brains[1-4]. Disruption in transcript-specific regulatory

elements due to DNA mutations can lead to diseases. Transcriptome-wide changes are implicated in neuropsychiatric conditions, including autism spectrum disorder (ASD)[5-8]. Accurate annotation and interpretation of these changes relied on a comprehensive transcriptomic profile, either for a given gene or on a genome-wide scale. However, popular short-read sequencing is suboptimal for delineating longer transcripts and discovering novel exons and splicing events[9]. Standard long-read sequencing techniques are not sufficiently sensitive to detect transcripts with lower abundance. A theoretical solution lies in the combination of mRNA/cDNA-capture methods[10] and long-read sequencing, could identify both long and low abundant transcripts. However, this approach has been sparingly reported, probably due to the technical challenge of preserving the mRNA integrity. The current inability to construct a complete transcriptome fuels a continuing debate over the extent of pervasive transcription across the genome and the significance of transcriptional "dark matter" endogenously[11-15]. The incomplete transcriptome impedes accurate annotation of disease-linked variants and interpretation of transcriptomic data. This shortfall affects the validation of genetically modified disease models used in preclinical research to develop molecular therapies. Previous studies have indicated specific functions of *SHANK3* mRNA transcripts at synapses[16-29]. An incomplete human *SHANK3* transcriptome could underestimate the contribution of the genetic risk for ASD and other neuropsychiatric disorders. Similarly, the incomplete mouse transcriptome complicate interpretations of their relevance to human *SHANK3* disorders from studies of more than 24 lines of genetically modified animal models[30 18,31]. To bridge these substantial gaps in knowledge, we performed standard Iso-Seq (**SIS**) for whole transcriptome analysis and paired with targeted cDNA

capture and long-read sequencing techniques (Capture-Iso-Seq, **CIS**) to specifically investigate the *SHANK* family genes in humans and mouse brain. We discovered a drastically intricate *SHANK3* transcript structure and a broad transcriptomic diversity across the human and mouse genomes. We identified unexpected extensive fusion transcripts and atypical patterns of transcripts in *Shank3* mutant mice. The enhanced *SHANK3* transcriptome has significantly improved the discovery rate of deleterious variants in genomic and transcriptomic studies of neuropsychiatric disorders. Our study advocates for a paradigm shift in experimental design and evaluation of genetic disease models using genetically modified animals, emphasizing the need to carefully evaluate the molecular validity of these mutant animal models in preclinical research.

Results

Dataset overview and experimental strategy evolution and optimization

We sequenced 56 SMRT libraries of human and mouse brains using the PacBio Sequel II System (**Fig. 1A-B**). Sixteen libraries proceeded using the SIS method. Forty libraries were constructed following the CIS method, which employed targeted capture enrichment with specific oligonucleotide probe panels, that covered the full genomic regions of *SHANK/Shank* family genes (*SHANK1-3*, **Supplementary Table 1-2**). A non-neuronal gene, *TP53,* was included as a comparison. Twenty libraries were synthesized from cerebral cortex of neurotypical children aged 5-6 years, and young adults, aged 24-30 years. For mice, 35 libraries were derived from striatum (ST) and prefrontal cortices (PFC) of 21-day-old wild-type (WT) C57BL/6J and *Shank3* mutants (*Shank3*$^{\Delta e4-9}$, *Shank3*$^{\Delta e21}$ and *Shank3*$^{\Delta e4-22}$)[19,20,32-34]. We processed only the RNA with an Integrity

Number (RIN) above 7 for human and above 8 for mouse samples for subsequent sequencing. The quality and reproducibility of the SIS and CIS platforms were optimized (**Supplementary Fig.1A-I)**. For experimental validation, RT-PCR and Sanger sequencing confirmed novel *SHANK3* transcripts from CIS. We performed *in silico* transcriptome analyses using short-read bulk RNA-seq (**srRNA-seq**) and single-cell RNA-seq (**scRNA-seq**) data, and gene discovery analyses of exome sequencing (ES) and whole genome sequencing (WGS) data from PsychENCODE project along with other genomics studies[5,35-39].

Standard Iso-Seq uncovered more diverse transcriptome genome-wide in mouse and human brains

From the SIS of 12 SMRT libraries of human brain, we uncovered 131,585 unique transcripts across 15,308 annotated genes, including 311 novel transcripts (**UCSC Track 1**). Distribution of unique transcripts and sequencing reads per gene are shown in **Fig. 1C**. The number of unique transcripts for a given gene was significantly correlated (Pearson r=0.8871, p<0.001) with its abundance (**Fig. 1D**). From 4 SIS of mouse ST and PFC, we uncovered 154,492 unique transcripts from 16,556 annotated genes, with 1,570 being novel (**Fig.1E** and **UCSC Track 2 and 3)**.

In human brains, the average number of isoforms per gene was 19, with an average sequence read count of 63. Notably, 595 genes exhibited over 100 isoforms (**Fig.1E** and **Supplementary Table 3a)**. *SEPTIN4* has the highest number of isoforms at 692, a gene encoding a presynaptic scaffold and GTP-binding protein, involved in exocytosis,

and interacted with alpha-synuclein, implicated in Parkinson's disease[40]. In mouse brains, the average number of unique transcripts per gene was found to be 8, with an average of 17 sequence reads per transcript. *Sorbs1* had the highest number of isoforms at 158; this gene encodes a Sorbin and SH3 domain containing protein involved in insulin signaling and stimulation[41] **(Supplementary Table 3b)**. We identified 182 genes with more than 50 isoforms and 19 genes with over 100 isoforms in mouse brains. Of these, seven have human orthologs that also exhibit more than 100 isoforms. Our studies revealed a greater transcript diversity than other studies using the same sequencing platform and analytic algorithm[8,42]. We examined the transcript diversity of 213 highly confident ASD risk genes consolidated from 3 recent extensive ASD genomics studies using our SIS data[43-45] (**Fig. 1F** and **Supplementary Table 4**). On average, individual ASD risk genes exhibited 56 transcripts, with a 90% CI ranging 8-140. *ANK2* was noted for having the highest number of transcripts of 372. Remarkably, the expression level of *SHANK3* was one of the lowest, ranking 212 of 213 ASD risk genes (**Fig. 1F**). Genes associated with brain disorders, especially ASD and neurodevelopmental disorders (NDD), have significantly greater numbers of transcripts compared to genes implicated in disorders not related to the brain (**Fig. 1G-H).**

A complex mouse *Shank3* transcriptome from CIS

We noted that the longest annotated *SHANK3/Shank3* transcripts in humans (NM_001372044.2, 7,691 bp, hg38) and mice (NM_021423.4, 7,380 bp, mm39) have not been detected in any published long-read RNA-seq datasets[6,8,42]. From 4 SIS of mouse ST and PFC, we identified only 5 *Shank3* transcripts (ranging 5,625-6,463 bp) in

ST, with none detected in PFC upon validation. The discrepancy in transcript number and the variation between ST and PFC were consistent with the highest expression level of *Shank3* in ST and lower expression in neocortex at P21 days[25]. The failure to detect longer *Shank3* mRNAs by SIS was most likely due to their low abundance, as transcripts up to 14.5 kb were successfully sequenced in our libraries (**Supplementary Fig. 1F-G**).

With CIS, we detected 545 *Shank3* transcripts in the mouse ST (**Fig. 2A**) and 345 in PFC (**Fig. 3A**), including the longest annotated transcript (NM_021423.4). We successfully validated 51 (85%) out of 60 representative novel transcripts by RT-PCR and sequencing (**Fig. 2E-H** and **Supplementary Table 5**). To evaluate the quality of each transcript, we employed a confidence metric that integrates the transcript abundance, the length of predicted open reading frame (ORF), and validations with srRNA-seq data (**Supplementary Fig. 2A**). In ST, 223 (41%) of *Shank3* transcripts were classified as high confidence, while 382 (59%) were in moderate confidence. In PFC, 168 (49%) transcripts were in high confidence, with the remaining and 176 (51%) of moderate confidence. Analysis revealed 36 and 26 potential transcription start sites (TSS) in ST and PFC, respectively. In the ST, 142 *Shank3* transcripts originated at exon 1 of the annotated referenced transcript (NM_021423.4) and terminated at 26 different sites **(Fig. 2B)**. Thirty-five transcripts terminated within exon 21, each presenting different ORFs. Exon 21, the largest coding exon of 2,257 bp, was spliced out in many transcripts. Over 90% of transcripts terminated within 100-500 bp of annotated transcription termination sites (TTS) and poly A signals (**Supplementary Fig. 3**). This

161 indicates that the early terminations are not artifacts of RNA degradation or cDNA
162 synthesis errors. Intron retentions were observed in introns 1, 2, 11, 12, and 19, leading
163 to altered ORFs and earlier stop codons. While some transcript structure variations
164 were subtle, they are predicted to encode different ORFs (**Fig. 2C-D)**.
165
166 In the PFC, we identified 59 *Shank3* transcripts initiating from 19 different exons and
167 terminating within the last coding exon 22 (**Fig. 3A-B**). Notably, 28 of these transcripts
168 started within exon 21 with different ATG codons. This finding aligns with our prior
169 results obtained from 5′ RACE experiments[25]. We discovered 12 new exons in ST and
170 17 in PFC, with 11 being shared to both regions (**Fig. 2A** and **Fig. 3A**). Additionally, we
171 discovered 4 new untranslated exons, U1-4, located 5′ upstream of the annotated
172 *Shank3* exon 1 **(Fig. 2A)**. Six new and alternative spliced exons, E9a-f, were identified
173 between exons 9 and 10. The spliced variants between exons 9 and 10 were the most
174 abundant with 4,326 reads in ST and 641 in PFC, while exon 12e was exclusive to the
175 PFC (**Fig. 2E**).
176
177 Surprisingly, we observed a considerable number of novel fusion transcripts, in which
178 different *Shank3* exons were joined to downstream exons 2-5 of the *Acr* gene, which
179 encodes the acrosin protein in the acrosome of spermatozoa[46] (**Fig. 2A** and **Fig. 3A**).
180 These fusion transcripts were validated by PCR and sequencing (**Fig.2E-F**). We noted
181 that splice events linking *Shank3* exons 17 and 21 to *Acr* exons occurred more
182 frequently than others. Specifically, fusions from *Shank3* exon 21 to *Acr* exon 2 (208
183 reads) and exon 3 (243 reads) were the most abundant. We also identified splice

products from *Shank3* exons 17 and 21 to three novel exons/transcripts (T1-3) situated downstream of *Acr* (**Fig. 2A**). These transcripts in ST and PFC are predicted to yield five ORFs, extending the SHANK3 protein by an additional 64 amino acids (NP_001358973).

The transcriptomic architecture of *Shank3* revealed by CIS in ST and PFC displayed both shared and unique characteristics. Overall, 230 transcripts (42% of ST, 67% of PFC) were common to both brain regions **(Fig. 3C).** We analyzed the tissue specific usage of TSS and coding sequence starting sites (CDS). Transcripts were categorized as follows: overlapping with the annotated *Shank3* mRNA, U1-4 to *Shank3*, *Shank3-Acr* fusion, and *Shank3*-T1-3. In ST, 75% of transcripts belonged to the category overlapping with the annotated *Shank3*, and 24% fell within the *Shank3-Acr* fusion category (**Fig. 3E**). In PFC, 52% of the transcripts were overlapping with annotated *Shank3*, while 43% were classified as *Shank3-Acr* fusion transcripts **(Fig. 3F).**

Protein domain specific mouse SHANK3 proteome

SHANK3 and its family encode proteins possess 6 domains of ubiquitin-like (Ubl), ankyrin repeats (ANKYR), postsynaptic density protein 95/discs large homologue 1/zonula occludens 1 (PDZ), an Src homology 3 (SH3), a proline-rich region containing Homer and Cortactin-binding sites (Pro), and a sterile alpha motif (SAM)[47-49]. As scaffold proteins in the postsynaptic density (PSD) of synapses, SHANK3 protein interacts with various synaptic proteins via these domains, contributing to synaptic architecture and function. There are 474 ORFs predicted from 545 *Shank3* transcripts in

ST and 270 ORFs in PFC using GeneMarkS-T[50], with 261 ORFs being common to both brain regions (**Fig. 3D).** ORFs of novel transcripts were further corroborated by proteome data derived from various *in silico* datasets, utilizing graded criteria for sequence identity and overlap **(Supplementary Fig. 2B**).

Among the 125 ORFs predicted from 140 *Shank3* transcripts starting from exon 1 in ST, only 4 ORFs encompassed all 6 protein domains (**Fig. 3G**). Among the 270 ORFs predicted from 345 *Shank3* transcripts in PFC, only one ORF contained the complete set of 6 protein domains, while 37 ORFs have more than 3 protein domains **(Fig. 3H-K).** One hundred nineteen SHANK3 ORFs (30%) in PFC comprised only a single protein domain, typically the Pro domain. Approximate 15% of the predicted ORFs lacked recognized protein domains. The protein domain combinations were found to be non-random and tissue-specific; for instance, no predicted ORFs included the SAM-SH3 combination. The SAM-Pro-SH3 and SAM-SH3-ANKYR domains combinations were exclusive to PFC, while the Ubl-ANKYR-Pro-SAM and ANKYR-SH3-PDZ-Pro combination were identified only in ST (**Fig. 3L**).

Uniquely altered *Shank3* transcriptome in *Shank3* mutant mice

Sixteen *Shank3* mutant mouse lines and 8 mutant rat, dog, and non-human primate featuring various exonic deletions or point mutations, have been generated to model *SHANK3* associated ASD[30] (**Fig. 4A**). Using the same *Shank3* probe design, we conducted CIS on *Shank3* mutant mice: those with deletions of exons 4-9 (*Shank3*$^{\Delta e4\text{-}9}$), exons 4-22 (*Shank3*$^{\Delta e4\text{-}22}$), and exon 21 (*Shank3*$^{\Delta e21}$)[19,20,32,33]. In *Shank3*$^{\Delta e4\text{-}9}$

230 homozygous mice, we detected 69 *Shank3* transcripts in ST and 56 in PFC.
231 Representative mutant and residual transcripts are diagramed in **Fig. 4B**, with details
232 provided in **Supplementary Fig. 4A-B**. In ST and PFC of *Shank3*$^{\Delta e4\text{-}9}$ mice, we
233 identified 3 long transcripts (~7.3 kb), harboring a deletion of exons 4-9. Interestingly,
234 the first exon of these transcripts, with the exon 4-9 deletion, was in intron 1 of the
235 annotated *Shank3*, a TSS was not utilized in WT mice, suggesting an alternative TSS
236 due to the exon 4-9 deletion. These transcripts also lacked coding exon 22 and
237 exhibited fusions between exon 21 of *Shank3* and exon 2 of *Acr*. ORF prediction
238 suggests that the resultant SHANK3-ACR fusion proteins for these mutant transcripts is
239 1254 aa for PB.6361.147, 1073 aa for PB.6623.114, and 833aa for PB.6623.199.
240 Approximately 70% of the residual transcripts are initiated from intron 16/exon 17 and
241 terminate within exon 21/intron 21 of *Shank3* or exon 5 of *Acr*. Transcripts starting at
242 exon 11 were exclusively detected in ST. The proportion of transcripts initiated from
243 intron 16/exon 17 was increased in *Shank3*$^{\Delta e4\text{-}9}$ mice compared to WT. A total of 54
244 ORFs (ranging 113 to 1,327 aa), were predicted in ST, with a similar pattern observed
245 in PFC from residual transcripts of *Shank3*$^{\Delta e4\text{-}9}$ mice**.**
246
247 In *Shank3*$^{\Delta e21}$ homozygous mice, we identified 401 *Shank3* transcripts in ST and 148 in
248 PFC (**Fig. 4C**, **Supplementary Fig. 4C-D**). In *Shank3*$^{\Delta e4\text{-}22}$ homozygous mice, the
249 number were 436 in ST and 792 in PFC (**Fig. 4D**, **Supplementary Fig. 4E-F**).
250 Remarkably, over 99% of these transcripts were *Shank3-ACR* fusion events in both
251 brain regions of *Shank3*$^{\Delta e21}$ and *Shank3*$^{\Delta e4\text{-}22}$ mice. The predominant transcripts in
252 *Shank3*$^{\Delta e21}$ mice was from the intron 16/exon 17 region in both ST and PFC. Conversely,

in *Shank3*$^{\Delta e4\text{-}22}$ mice, transcription primarily initiated from intron 1/exon 2. We also detected multiple novel exons interposed between *Shank3* and *Acr* genes (**Supplementary Fig. 4C*-D***), exclusive to these *Shank3* mutant lines and absent in WT. Fusion transcripts of *Shank3-Acr* were more prevalent in *Shank3*$^{\Delta e4\text{-}22}$, *Shank3*$^{\Delta e4\text{-}9}$, and *Shank3*$^{\Delta e21}$ mutants. Moreover, a significant overexpression of *Acr* transcripts was found in neocortex and hippocampus of *Shank3*$^{\Delta e4\text{-}22\text{-}/\text{-}}$ mice (**Fig. 4E-F**). Bulk RNA-seq data analysis from ST of *Shank3*$^{\Delta e4\text{-}22\text{-}/\text{-}}$ mice also indicated a compensatory expression from *Shank1* and *Shank2*, which was protein domain-specific (**Fig. 4H-K**).

The *Shank3* transcriptomic findings from CIS prompted us to extend our approach to include all *Shank* family genes (*Shank1-3*) using a joint capture strategy. This joint CIS for the *Shank* family genes identified 664 *Shank1* and 495 *Shank2* transcripts in PFC, and 320 *Shank1* and 326 *Shank2* transcripts in ST (**UCSC Tracks 4 and 5**). The overall transcript structures and patterns of *Shank3* from both single-gene and joint CIS were similar. We discovered 7 novel exons upstream of the annotated exon 1 of *SHANK1* (**Supplementary Fig. 5A**). Fusion transcripts involving *Shank1* and *Shank2* with adjacent genes were also detected. The most upstream novel exon of *Shank1* overlapped with the last exon of *Clec11a* gene (NM_009131.3), which is transcribed in the reverse direction to *Shank1* **(Supplementary Fig. 5B)**. The fusion transcripts between *Shank1* and *Josd2*, a gene located approximately 100 kb downstream, were exclusively detected in PFC. Two new untranslated exons, U1 and U2, were found about 24kb upstream from the annotated 5′ exon 1 of *SHANK2* (**Supplementary Fig. 5C**)

Transcript diversity of *SHANK* family genes in human brains

In current reference genome (hg38), an annotated human *SHANK3* mRNA (7,691 bp, NM_001372044) is displayed, yet it has not been experimentally validated. With CIS on *SHANK* family genes, we discovered 472 unique *SHANK3* transcripts (**Fig. 5A-C, UCSC Track 6**), with the longest was 6,824 bp. Notably, the annotated 7,691 bp *SHANK3* transcript (NM_001372044) was absent. The absence of the longest *SHANK3* transcript is unlikely a result of RNA degradation, because a 10.8 kb *SHANK2* transcript was detected in the same captured sample. Instead, it appeared due to extremely low or no expression of the full-length *SHANK3* transcript in adult frontal and temporal cortices. Most of the 472 unique *SHANK3* transcripts clustered within regions spanning exons 1-9 and 10-22. None incorporated splicing between exon 9 and 10, an area characterized by high GC content (77% of GC) and a CpG island (hg38). Accordingly, 43 unique transcripts initiated from this CpG island, implying a TSS within intron 9. *In silico* analysis using a parameter-free assembly approach (Cufflinks-Cuffmerge)[51] applied to srRNA-seq data also failed to detect any transcripts connecting exon 9 and 10.

Similar to mouse *Shank3*, we detected 66 fusion transcripts between *SHANK3* and *ACR* (**Fig. 5C**). These fusion transcripts, intron retention, and novel exons were validated by RT-PCR and sequencing **(Fig. 5D)**. Fifty-eight of them were fusion transcripts constituted exon 19/exon 20 of *SHANK3* (exon 20 is the largest exon in human equivalent to exon 21 in mice) to exons 2-5 of *ACR*. Nine transcripts started within

SHANK3 exon 20 were found to be fused with *ACR*. We observed splicing events connecting *SHANK3* exons 19-20 to uncharacterized downstream exons, T1-2, of *ACR*. We also detected 3 novel untranslated exons (U1-U3) upstream exon 1 of *SHANK3* mRNA (**Fig. 5E**). The sequence of U2 is highly conserved in mouse.

With the joint capture for *SHANK* family genes, we detected 86 *SHANK1* and 277 *SHANK2* transcripts (**UCSC Track 6**), from which 69 ORFs for *SHANK1* and 165 ORF for *SHANK2* were predicted. Across these *SHANK* family ORFs, we observed 17 different combinations of six functional domains, with the PDZ domain appearing most frequently (**Fig. 5F**). A complete set of all six functional domains (Ubl, ANKYR, SH3, PDZ, Pro, and SAM) was predicted only in one *SHANK2* transcript.

The unexpected discovery of extensive fusion transcripts between *SHANK3* and *ACR* in human brain tissue led to a comprehensive genome-wide analysis for fusion transcripts in SIS data. We detected 2,265 fusion transcripts (1.7% of the total transcripts) associated with 3,499 genes in the brains of children and adults, with 963 fusion transcripts common to both groups. About 98% of fusion transcripts are between two adject genes. A small number of fusion transcripts are among 3 adjacent genes. No fusion transcript is from distant genes or genes from two chromosomes. Gene Ontology enrichment analysis revealed a significantly enrichment of fusion transcripts in genes associated with ASD (**Fig. 5G-H**).

To access the functional constraint of novel *SHANK3/Shank3* exons in humans and mice identified by CIS, we utilized Evolutionary Rate Profiling (GERP)[52,53] and PhyloP[54] conservation scores. In mice, GERP and PhyloP scores for most *Shank3* novel exons were significantly higher than those of non-transcribed region, but they were lower than scores for known coding exons in both PFC and ST (**Fig. 5I-J, supplementary Table 6A-B**). A concordant pattern was observed in human *SHANK3* (**Fig. 5K-L, supplementary Table 6C-D**). These results suggest that the novel exons of *SHANK3/ Shank3* uncovered by CIS are evolutionarily constrained elements, underscoring their potential functional significance.

Transcript diversity and novel transcripts of *TP53* gene in human and mice

To examine whether the transcriptional complexity is exclusively associated with synaptic genes, we applied SIS and CIS to *TP53* in human brain, and to *Trp53* in mouse brains and thymus, where *Trp53* expression is the highest. SIS detected only 5 *Trp53* transcripts in mouse ST and 3 in mouse PFC that is consistent with the data in literature [55,56],. In contrast, CIS identified a comprehensive set of 243 transcripts from thymus, 164 from PFC, and 188 from ST (**Supplementary Fig. 6A-C, UCSC Track 7**). The pattern of unique *Trp53* transcripts is similar amongst the 3 tissues, with 18 alternative TSS deduced from thymus transcripts. A significantly higher percentage of transcripts exhibited intron retention in *Trp53* compared to *Shank3*. Additionally, novel tissue-specific 5′ exons unique to brain (bU1) and thymus (tU1/tU2) were discovered.

In human brain, CIS detected 106 *TP53* transcripts, which predicted 60 ORFs, 18 TSSs, and three 3′ transcriptional ends (**Supplementary Fig. 6D, UCSC tracks)**. We also discovered 3 novel exons (hT1-3) at the 3′ end, which extended the C-terminus of TP53 ORF by 72aa and is conserved with the mouse *TRP53* (77% identical). These observations underscore the complexity of the *TP53/Trp53* transcriptome, which is complex but less heterogenous than *SHANK* family genes.

Developmental, tissue, and tell type specificity of SHANK3/Shank3 transcripts from CIS

To investigate the developmental specificity of *Shank3* transcriptome, we aligned mouse srRNA-seq data of cerebral cortex at different ages from E14.5 to P180 day[57-59] to *Shank3* transcripts from CIS (**Fig. 6A**). The E14.5 embryos exhibited the least diversity of *Shank3* transcripts of. As development, the number of unique *Shank3* transcripts increased, reaching a maximum at P56 day before declining at P180 day. Further analysis on cell type specificity aligning scRNA-seq data from anterior cingulate area (ACA) of 8-week-old mice[39] to *Shank3* transcripts identified by CIS, demonstrated a significantly higher abundance of *Shank3* transcripts in glutamatergic neurons compared to GABAergic neurons. The *Shank3* transcripts including exon 18 was exclusively found in endothelial cells (**Fig. 6B**).

To investigate tissue specificity, we analyzed the exon usage in mouse *Shank3* transcripts from CIS against scRNA-seq data from 5 cerebral cortex subregions[39]. The exon usage patterns of *Shank3* CIS transcripts within the same cell type exhibited

unique variations across different brain subregions (**Supplementary Fig.7**). The pattern of human *SHANK3* transcripts in infants and children was distinct from adults, when aligned human srRNA-seq data to *SHANK3* transcripts from CIS. The *SHANK3* exon usages were also changes with age.

We mapped *Shank3* transcripts to 10x Genomics Visium spatial transcriptome of mouse to visualize the expressive pattern *in situ*[60]. Two probes targeting *Shank3* exons 11 and 22, and one for *Acr* exon 5, facilitated this analysis. Three *Shank3* transcripts identified by CIS were enriched to distinct anatomical regions (**Fig. 6C-F**). *Shank3* transcript TALONT000202476 containing exon 11, and TALONT000200721 incorporating exon 22, have a similar cell-specific expression pattern, albeit at different levels of abundance. Transcript TALONT000200852, a fusion transcript connecting *Shank3* exon 21 and *Acr* exon 5, displayed a cell type-specific expression pattern**.** Furthermore, we found a cellular compartment-specific preference for the *Shank3* transcripts. The inclusion of *Shank3* largest exon 21 is significantly more common in synapses than in nuclei from mouse brain scRNA-seq data[61] (**Fig.6G**). The exon 2 of *Acr* gene, frequently fused with *Shank3* exons, was significantly less present in nucleus of AD models compared to WT. The splicing events involving 5′ segment of *Acr* exon 5 were more common across both nucleus and synapses in AD mice, while splicing involving the latter 2/3 of *Acr* exon 5 was more frequent in nucleus of WT (**Fig. 6H).**

Applications of *SHANK3* transcriptome from CIS to genome sequencing and transcriptome analyses of ASD and other neuropsychiatric disorders

Human *SHANK3* transcripts identified through CIS exhibit expression patterns that are specific to developmental stages and brain regions, such as the cerebral cortex and cerebellum (**Fig. 7A-D** and **Supplementary Fig. 8**). We extended the *in-silico* transcriptome diversity analysis to 213 highly confident ASD risk genes consolidated from 3 recent extensive ASD genomics studies (**Supplementary Table 4)**[43-45]. The transcriptome diversity of ASD risk genes was significantly greater than of non-ASD associated genes (Pearson r=0.386, p<0.001)**.** Specifically, ASD risk genes associated with gene expression regulation and neuronal communication showed significantly higher level of transcriptome complexity, compared to genes in other functional categories (Pearson r=0.825 and Pearson r=0.793, respectively, both p<0.001). *SHANK3*, consistently reported as one of the top 5 ASD causing genes in these studies[43-45], is also implicated in schizophrenia (SCZ)[62], bipolar disorder (BPD)[63], and major depressive disorder (MDD)[64]. To investigate alterations in *SHANK3* transcriptomes across these disorders, we analyzed srRNA-seq data from the PsychENCODE project[35]. Principle component analysis (PCA) revealed unique transcript patterns for each disorder, especially for ASD and SCZ **(Fig. 7E).** The expressions of a subset of *SHANK3* transcripts varied across ASD, MDD, BD, SCZ, and controls **(Fig. 7E-I)**. Brain region and age-specific expression of *SHANK*3 transcripts formed distinct cluster in PCA analysis (**Fig. J**). The exons 12, 15, 20, and 22 of *SHANK3* transcripts in BA7 were significantly more represented in ASD brains than controls **(Fig. 7K)**, and exon 10 showed a higher expression in BA38 of ASD brains **(Fig. 7L).**

While *SHANK3* genetic mutations are implicated in 1-2% of ASD cases and to a lesser extent in other neuropsychiatric disorders[43-45,65,66], we sought to examine whether incorporating the enhanced *SHANK3* transcript structure from CIS into public available ES and WGS of ASD/SCZ/BPD datasets could uncover additional disease-associated single nucleotide variants (SNVs)[45,67-70]. We re-analyzed sequence variants on a large cohort of 177,000 samples of both controls and disease subjects, including ES data from the Autism Sequencing Consortium[45], BPD Exomes[68], SCZ Exome Meta-analysis Consortium[67], as well as WGS of ASD, SCZ, and BP cohorts from BrainVar[69] and BrainGVEX[70]. Variant identifications and annotations of were previously based on the mRNA reference NM_001372044.2 and hg38 genome assembly. We used Variant Effect Predictor (VEP, release 107)[71] and Genome Aggregation Database (gnomAD, v3.1.2)[72], for annotation and filtering, including variants with a population allele frequency of <= 0.01 for protein truncating variants (PTVs), and excluding missense and synonymous variants for further analysis. SpliceAI[73] and SnpEff[74] were used to analyze splice variants and evaluate the pathogenic potential of stop-loss, stop-gain, and frameshift variants. This re-annotation identified 1,530 new SNVs across 55,000 cases pooled from ASD (11,986 ES; 923 WGS), BP (14,210 ES), and SCZ (27,648 ES) cohorts **(Fig. 7M)**, resulting in the discovery of 27 stop-loss, 60 stop-gain, 52 frameshift, and 53 splice variants in *SHANK3* considered potentially deleterious or PTVs using CIS annotation in disease subjects but not in controls. This was a marked contrast to the variants analyzed using current reference (0 stop-loss, 1 stop-gain, 4 frameshift, and 16 splice variants). Accordingly, the detection rate for potential deleterious SNVs of *SHANK3* increased from 1.3% when using current reference (NM_001372044) to 12.5%

435 when annotated with the *SHANK3* CIS transcripts, highlighting the significance of
436 comprehensive transcriptome annotation in uncovering genetic contributions to
437 neuropsychiatric disorders **(Fig. 7N).**
438
439 **Discussion**
440 Diverse transcription is crucial for generating proteomic diversity and facilitating
441 complex cellular functions. Precision of transcription is critical because mutations in the
442 transcriptional regulatory DNA elements can cause numerous single gene disorders.
443 Despite the recent report of the completed human genome[75], the transcriptome remains
444 largely uncharted. Our work applying SIS on human and mouse brains discovered
445 unprecedented transcriptome diversity[8,42]. Glinos *et al*[14] reported a maximum of 178
446 isoforms for a single gene, with only 5 genes exhibiting more than 100 isoforms, and a
447 median of 2 isoforms per gene across various tissues and cell lines. Leung *et al*.'s
448 study[42] noted a peak of 40 isoforms per gene in the human cortex. Furthermore, Chau
449 *et al*[76] assembled an average of 4 isoforms per gene from bulk RNA-seq of human
450 developing brains. Significantly, these studies uncovered only a few incomplete
451 *SHANK3* mRNA isoforms. However, our study identified as many as 692 isoforms for a
452 single gene, with 595 genes having more than 100 isoforms, and an average of 19
453 isoforms per gene in the human cerebral cortex. Our results suggest that the extent of
454 transcript complexity described in existing literature is significantly underestimated,
455 particularly for genes like *SHANK3*.
456

Our targeted capture and long-read sequencing have mapped the SHANK family transcriptomes in detail, with the majority of novel transcripts likely endogenously expressed. This is supported by our strict identification process, validation through RT-PCR and Sanger sequencing, consistency across experiments and brain regions, and conservation between species. Additionally, the specificity of these transcripts was confirmed in Shank3 mutant mice. Despite the high confidence, it remains a possibility that a small fraction might not be expressed endogenously. The discovery of a substantial number of fusion transcripts for *SHANK3/Shank3* in our study was unexpected, with a prevalence that surpassed the findings from other studies[8,42]. Until recently, fusion transcripts have been largely investigated in cancer-related studies because of their oncogenic properities[77,78]. Yet, their presence in normal cells has only recently been acknowledged [8,42,79]. Two recent studies using the SIS method[8,42] reported a mere 136 fusion transcripts (0.41% of total transcripts) in human brains. In contrast, our study identified 2,265 fusion transcripts in human brains, constituting 1.7% of total transcripts. Interestingly, these fusion transcripts were found to be particularly more enriched in the human ASD-associated transcriptome.

The enhanced *SHANK3* transcript structure from CIS has significantly increased the detection rate of PTVs or predicted LOF variants in ES and WGS data for neuropsychiatric disorders. Further functional validations are warranted to determine the pathogenicity of these new identified PTVs. Our findings highlight the significance of employing fully characterized transcript structures in genomics studies of disease gene discovery. Transcriptional dysregulations in the brain have been implicated in

480 neuropsychiatric disorders[5,80]. By integrating the *SHANK3* transcriptome data from CIS
481 and the transcriptome data from PsychENCODE, we discovered brain region-specific
482 dysregulations in *SHANK3* transcriptome associated with ASD and other
483 neuropsychiatric disorders. Notably, brain region-specific DNA methylation in intragenic
484 CpG islands, which show altered methylation in ASD brains[81,82], suggest that epigenetic
485 changes could be instrumental in *SHANK3* transcript variations.
486
487 In *Shank3* mutant mice, stable transcripts with exonic deletions indicated truncated
488 protein production or upregulated non-mutant isoforms[30,83]. Cryptic promoters,
489 especially within intron 16/exon 17, suggest alternative initiation and potential novel
490 protein isoforms. These could perturb the postsynaptic density protein interactome,
491 indicating possible loss and gain of function in *Shank3* mutants. Such complexities
492 question the molecular and phenotypic consistency of *Shank3* mouse models with
493 SHANK3 disorders[16-18,30,84,85]. For example, differential behavioral phenotypes and
494 receptor subunit alterations are noted across mutant lines[19,30,34]. Specifically,
495 *Shank3*$^{\Delta e21}$ mutants show unique upregulation of alternative transcripts and fusion
496 transcripts, diverging behaviorally from *Shank3*$^{\Delta e4\text{-}22}$ mutants[30,86]. These molecular
497 nuances challenge the translational fidelity of *Shank3* mouse models for preclinical
498 studies and necessitate re-evaluation, particularly for models in therapeutic
499 development.
500
501 Our study's detailed alignment of *SHANK3/Shank3* transcripts underscores its
502 proteomic diversity at the PSD, essential for complex synaptic functions[48,49,87]. However,

503 about 15% of the transcripts, possibly arising from cryptic promoters or alternative
504 splicing, lack substantial ORFs or are lowly expressed, hinting at stochastic transcription
505 events previously noted in other species[88-97]. Challenges to the ENCODE projects'
506 findings on genome transcription by subsequent short-read RNA-seq studies[11-13,15,98-100],
507 align with our discovery that *SHANK3/Shank3* and *TP53* transcription involves
508 intragenic promoters and frequent intron retention. These regions, less conserved
509 evolutionarily, affirm pervasive transcription and suggest a more deterministic
510 transcriptional landscape for these genes in humans and mice.
511

512 **Limitations of the study**

513 Several limitations of study that are warranted for discussion. We will not be able
514 quantify the extent of stochasticity of transcription from current analysis. The extensive
515 functional validation of transcripts at the protein level remains a challenge, as some
516 transcripts may function uniquely at the RNA level, eluding protein-interaction analyses.
517 Also, our capture-based method trades sensitivity for efficiency when scaling up, as
518 increased gene targets reduce sequence depth, necessitating careful experimental
519 design for quality data.
520

521 **Figure legends**

522 **Fig. 1. Genome wide transcript diversity and abundance in brains detected by SIS.**

523 **A.** Experimental design of SIS and CIS of human and mouse tissues.

524 **B.** Schematic of experimental procedure of RNA capture and long read-sequencing.

C. Number of unique transcripts (transcript diversity) for individual genes (blue) and the number of sequences reads (abundance) (red) for an individual transcript detected in human cerebral cortex by SIS with projected chromosome coordinates and ideograms. **D.** Transcript diversity was significantly correlated with the sequence reads (abundance) of the transcripts. **E.** Number of transcripts per gene genome-wide from SIS in human and mouse brains. **F.** Number of unique transcripts (Trans_Div) and abundance (Gene_FL) for 213 ASD risk genes, shown an average of 56 transcripts per gene and a median of 35. **G-H.** Human SIS data showed heightened transcript diversity in genes associated with brain disorders, especially ASD and NDD, compared to other diseases. We observed a strong correlation between transcript diversity and abundance in all gene clusters except for those related to dementia/Alzheimer's.

Fig. 2. Novel *Shank3* transcriptome in mouse striatum (ST) by CIS.

A. CIS revealed a refined *Shank3* gene structure and splicing patterns in WT mouse striatum. The established *Shank3* structure (NM_001034115, mm39) is expanded with newly detected exons shared between striatum and PFC, depicted in purple. Unique splicing events, represented by grey lines and thickness indicating read quantity, include novel striatum-specific exons in dark blue and alternative splices in light blue. Fusion transcript exons near Gm41381 and *Acr*, shown in green and orange, respectively, feature unique splicing with newly identified red exons (T1-T3) exclusive to *Shank3*. New exon U3 is shared between striatum and PFC. U4 is linked to Gm4138

and striatum specific. 21e is a new in-frame exon and 21c is a new exon harbor a stop codon.

B. 142 unique transcripts started with the canonical exon 1 of annotated *Shank3* (NM_001034115) in ST and terminated at different positions. Pink bar plots on the left are the abundance (log2 counts). Arrows describe the features of given transcripts.

C. Example of transcripts with similar structures in panorama but different at the sequence level with predicted ORFs and ATG codons. The transcripts of PB.13560.548, PB.13560.628, and PB.13560.547 are similar but the predicted ORFs show different ATG codons and protein domains.

D. Details of the split exon 1. There is a cryptic splicing of 127 bp (non-capitalized sequence in black) within the annotated exon 1 of transcript PB.106071.171 which resulted in a predicted upstream ATG codon and additional 134 amino acids. Other transcripts have transcriptional starting sites (TSS) in exon 1 but predicted ATG codon in exon 2. Variability in TSS and intron 1 retention, as seen in transcripts PB.13554.484, PB.13554.580, and PB.13554.668, leads to ORFs of 304 aa, 106 aa, and 1,290 aa, respectively.

E. Validations new transcripts from paired mouse PFC and ST samples. Pair 1, novel exon U1; Pair 2, fusion transcript between *Shank3* exon 21 and *Acr* exon2; Pair 3, splicing event between *Shank3* exon 9 and exon19; Pair 4, splicing event between *Shank3* exon 5 and exon 21; Pair 5, novel exon 9b of *Shank3*; Pair 6, *Shank3* exon11 extension/intron11 retention. The red arrows are the novel products confirmed by Sanger sequencing. Other bands are products from known transcripts.

F. Sanger sequencing confirmation of a fusion transcript between *Shank3* exon21 and *Acr* exon2 in mouse brain (pair 2 of E)

G. Fusion transcripts in other tissues. Forward and reverse primers were from exon 20 of *Shank3* and exon 5 of mouse *Acr* respectively. lane1, liver in P21 mouse; lane 2, thymus in P21 mouse; lane 3, ovary in P21 mouse; lane 4, ovary in 3 months old mouse; lane 5, testis in P21 mouse; lane 6, testis in 3-month-old mouse. The red arrows are the novel products confirmed by Sanger sequencing as indicated. Other bands are known products.

H. Sanger sequencing of *Shank3* exon 11 extension/intron 11 retention in mouse brain (lane 6 of G).

Fig. 3. Novel *Shank3* transcriptome in mouse PFC by CIS and predicted domain structures of ORFs

A. New *Shank3* transcript structure and conch plot of splicing events discovered in WT mouse PFC by CIS. Color code is the same as **Fig. 2A**. The novel exon 9a (chr15:89394416-89394465, mm39) is shared between PFC and ST. Other novel exons such as exon 12e (chr15:89414330-89414640, mm39) were unique to PFC. Novel exons 21a, 21b and 21c are predicted to result in an early stop codon and shorter ORFs (chr15: 89394416-89394465, chr15: 89408698-89408784, chr15: 89418571-89418609, mm39).

B. Structure of 59 transcripts with different TSSs but terminating at annotated exon 22 of *Shank3*. Pink bar plot represents the abundance (log2 counts) of each transcript.

C-D. The comparison of transcripts and predicted ORFs between mouse ST and PFC.

E-F. The pattern of deduced TSS and predicted starting sites of the coding sequence (CDS) for all *Shank3* transcripts including new 5′ and 3′ fusion transcripts from CIS in mouse ST (E) and PFC (F). Each filament represents an individual transcript in different classes of GM41381(U1-U2)-*Shank3*, *Shank3*-T1-3, *Shank3,* Shank3-Acr (first column), deduced TSS (middle column), and predicted starting sites of CDS (third column).

G. A total of 125 unique ORFs are predicted from 142 transcripts starting with exon 1 in ST. The pattern of the combination of 6 protein domains is shown in the outermost ring of the windmill plot. The middle layer shows the abundance of each RNA transcript and the p value of its expression level compared to other transcripts. Only 4 ORFs of transcripts contained all 6 protein domains.

H-K. Four windmill plots showing 270 predicted ORFs from all 345 transcripts detected in PFC classified by the combination of functional domains.

L. Spiral plot showing an aggregated functional domain coverage of the transcripts captured by *Shank1-3* joint probe panel by CIS of mouse PFC and ST. Each dot represents a unique transcript. Each color represents a unique combination of functional domain. The dots are ordered from the longest to the shortest transcript, while the colors are arranged from the SAM to UBL domain.

Fig. 4. The summary and illustration of altered *Shank3* transcripts in *Shank3*$^{\Delta e4\text{-}9}$, *Shank3*$^{\Delta e21}$ and *Shank3*$^{\Delta e4\text{-}22}$ mutant mice from CIS

A. Current annotated mouse *Shank3* and *Acr* (NM_013455, mm39) gene structure. The annotations of genetically targeted mutations in mice, rat, monkey, and dog are shown. (KO: exonic deletions; KI; knock-in mutation)

B. The gene structure of *Shank3*$^{\Delta e4\text{-}9}$ mutant mice in grey and representative mRNA transcripts from *Shank3*$^{\Delta e4\text{-}9\text{-}/\text{-}}$ mice are in pink. No transcript using first annotated exon 1 was detected. Instead, the first exon, presumably a cryptic TSS (arrow), was detected in intron 1. The exon 4-9 deleted transcript missed exon 11, 12, and 22 but with fusion between *Shank3* and *Acr*. The transcripts starting at intron 16/exon 17 (arrows) as first exon were most abundant. Extensive fusion transcripts between *Shank3* exon 21 and *Acr* exon 2 were observed. The last coding exon 22 was not detected in any transcripts.

C. The gene structure of *Shank3*$^{\Delta e21}$ mutant mice and *Acr* gene in grey and representative3 mRNA transcripts from of *Shank3*$^{\Delta e21\text{-}/\text{-}}$ mice in blue. The splicing between exon 4 of *Shank3* and exons of *Acr* that resulted in fusion transcripts were observed. The transcripts starting at intron 16/exon 17 (arrows) as first exon and fusion between *Shank3* and *Acr* were most common. The coding exon 22 were not detected in any transcript.

D. The gene structure of *Shank3*$^{\Delta e4\text{-}22}$ mutant mice and *Acr* gene in grey and representative mRNA transcripts in purple. The number of fusion transcripts between *Shank3* and *Acr* is significantly increased in *Shank3*$^{\Delta e4\text{-}22\text{-}/\text{-}}$ mutant mice.

E-F. Increased expression of *Acr* transcript in *Shank3*$^{\Delta e4\text{-}22\text{-}/\text{-}}$ mutant mouse by RT-qPCR. The expression of *Acr* gene was significantly increased in both striatum and hippocampus by >100 folds.

G-J. Compensatory expression of the functional domains of *SHANK* family proteins in striatum of *Shank3*$^{\Delta e4\text{-}22}$ mutant mice**.** The bulk RNA-seq data of *Shank3*$^{\Delta e4\text{-}22}$ were analyzed for the compensatory expression of other functional domains of *Shank1* and *Shank2* genes. The deficiency of ANKRY and SH3 domains of SHANK3 was

638 compensated by SHANK1 but the deficiency of PDZ and SAM domains were
639 compensated by both SHANK1 and SHANK2. The deficiency of SAM and SH3 domain
640 was fully compensated but the deficiency of ANKRY and PDZ domains was partially
641 compensated.
642
643 **Fig. 5. The novel transcripts of human *SHANK3* genes detected by CIS and**
644 **predicted ORFs**
645 **A.** New *SHANK3* transcript structure and Conch plot of *SHANK3* transcripts discovered
646 by CIS in normal human cortex. Black backbone is the annotated *SHANK3* transcript of
647 NM_001372044 (hg38). Blue rectangles represent novel exons of *SHANK3*. The exons
648 of *ACR* are in orange rectangles. The new and uncharacterized exons distal to *ACR* are
649 in red rectangles. The grey line connects adjacent exons while the light blue line
650 illustrates alternative splicing events. The number of sequences reads for the splicing
651 event is shown in the middle of connecting lines and reflected in the thickness of the
652 connecting lines.
653 **B.** Zoomed view of the splicing events between exons 10 and 20 in the human cortex.
654 Exons 16 and 20 of *SHANK3* in humans corresponds to exons 17 and 21 of *Shank3* in
655 mice.
656 **C.** Structure and abundance of the fusion transcripts between *SHANK3* and *ACR* in the
657 human cortex. Majority of fusion transcripts are initiated after exon 10, mainly from
658 introns 16, 17, and exon 21. The fusion transcripts are notably skipping exon 20 (the
659 largest exon) of *SHANK3* and exon 1 of ACR.

D. Validations novel *SHANK3* transcripts in in human brain tissue by RT-PCR and Sanger sequencing. Diagram for the primer design of L1 is shown. RT-PCR gel: L1, fusion transcript between *SHANK3* exon 20 and *ACR* exon 2; L2, fusion transcript between *SHANK3* exon 20 and *ACR* exon 4; Lane 3, fusion transcript between SHANK3 exon 19 and ACR exon 2; L 3, novel exon U3; Lane 4; L5, intron14 retention; L6, intron 15 retention. M, DNA marker. Sanger sequence of RT-PCR product of *SHANK3* exon 20 and *ACR* exon 2 fusion from L1

E. Three new exons upstream of the annotated exon 1 of *SHANK3* mRNA (NM_001372044) (U1, chr22:50672853-50672979; U2, chr22:50674076-50674097; U3, chr22:50674642-50674705, hg38). A new ATG codon is in U2.

F. Dandelion plot shows functional domain combinations of the *SHANK1*, *SHANK2*, and *SHANK3* transcripts from CIS. Each dot represents a unique transcript, and each color is a unique combination of functional domains. There are 17 combinations of functional domains of human *SHANK* family genes. The PDZ domain was significantly more present (~70%) in predicted ORFs.

G-H. Significant enrichment of fusion transcripts in transcriptome data of ASD and schizophrenia**.** For Gene Ontology enrichment analysis with Enrichr95 in 41 disease-related datasets. The fusion transcripts were significantly enriched in ASD and schizophrenia in Disease Perturbations form GEO dataset (G) and the ClinVar2019 dataset (H).

I-J. Distribution of GERP (G) and PhyloP (H) scores across human *SHANK3* genomic regions of known coding exons, novel exons from CIS, and non-transcribed region in cerebral cortex. **I**. GERP score for novel exons from CIS in cerebral cortex is

683 significantly high than non-transcribed region (D=0.097; p<0.001) but significantly lower
684 than that of *SHANK3* known exons (D=0.299; p<0.001). **J.** PhyloP score for novel exons
685 from CIS in cerebral cortex is significantly higher than non-transcribed region (D=0.133,
686 p<0.001) but significantly lower than that of *SHANK3* known coding exons (D=0.296,
687 p<0.001).
688 **K-L.** Distribution of GERP and PhyloP scores across mouse *Shank3* genomic regions of
689 known coding exons, novel exons from CIS, and non-transcribed region in PFC and ST.
690 **K**. GERP score for novel exons from CIS in PFC and ST is significantly high than that of
691 non-transcribed region (PFC: D=0.548, p<0.001; ST:D=0.602, p<0.001) but significantly
692 lower than known *Shank3* coding exons (PFC:0.15, p<0.001; ST:D=0.0960; p<0.001). **L**.
693 PhyloP score for novel exons from CIS in PFC and ST is significantly higher than that of
694 non-transcribed region (PFC:D=0.385, p<0.001; ST:D=0.439, p<0.001) but significantly
695 lower than known *Shank3* coding exons (PFC:D=0.184, p<0.001; ST:D=0.128,
696 P<0.001).
697

698 **Fig. 6. Developmental, cell type, cell compartment specific, and spatial**
699 **transcriptome of *Shank3* in mouse brains.**
700 **A.** Developmental specific *Shank3* transcripts in mouse cerebral cortex.
701 **B**. Cell type specific *Shank3* transcripts in mouse brains. The scRNA-seq of anterior
702 cingulate cortex (ACA)[5] was aligned to *Shank3* transcripts detected by CIS.
703 Glutamatergic neurons, especially the L2/3, L4/5, and L6 CTX, have more diverse
704 *Shank3* transcripts compared to GABAergic neuron and non-neuronal cells. Certain

transcripts were cell type specific. *Shank3* transcript (PB.10607.933) including exon 18 was only detected in endothelial cells.

C-F. Mouse *Shank3* transcripts in Visium spatial transcriptome. C. Visium spatial anatomy (CA: Cornu Ammonis, DG: Dentate Gyrus, TH: Thalamus, PIR: Piriform cortex, MEA: Medial Amygdala, CP: choroid plexus, CTX: Cortex, HPF: Hippocampal Formation, HY: Hypothalamus).

G. Cellular compartment specific changes of *Shank3* exon usage in the hippocampus of Alzheimer's disease (AD) mouse model from scRNA-seq data from different cellular compartment. The nucleus, compared to synapses, expressed significantly fewer splicing events of 32 and 33 that correspond to the exon 21, the largest exon of mouse *Shank3.*

H. Different pattern of *Shank3-Acr fusion* transcripts in nucleus and synapse between WT and AD mice.

Fig.7 Improved transcriptome analysis of ASD transcriptome and sequence variant annotations of genome sequence data using *SHANK3* transcript structure from CIS

A-D. The pattern of human *SHANK3* transcripts from CIS changed at different ages and brain regions**.** Bulk RNA-seq data of normal controls was aligned to *SHANK3* transcripts detected using CIS (BA, Brodmann area; CBL; cerebellum).

E-I. PCA of human *SHANK3* transcripts from CIS and bulk RNA-seq data of 2,474 cases with ASD, BPD, MDD), or SCZ, and normal controls from PsychENCODE (only data from prefrontal cortex is included). The clusters of MDD and BPD overlapped but

are separate from ASD and SCZ. The volcano plots for individual disorders ASD (n=68), MDD (n=87), BPD (n=297), and SCZ (n=736) compared to controls (n=1,286).

J. PCA analysis of *SHANK3* transcripts in different brain regions and age (BA, Brodmann area; CBL, cerebellum)

K-L: Brain region-specific change of *SHANK3* transcripts in ASD brains. Bulk RNA-seq data of subregions of the brain from ASD and controls were aligned to *SHANK3* transcripts from CIS. **K.** Exons 11, 15, 20, and 22 of *SHANK3* transcripts were significantly more represented in the BA7 region of ASD. **L.** Exon 10 of *SHANK3* transcripts is significantly more represented in BA38 of ASD brain.

M. Utilizing the updated *SHANK3* transcript structure from CIS enhanced PTV detection in ASD, SCZ, and BPD exome and genome sequencing data. From 55,000 cases, we identified 1,530 new PTVs, a significant increase from previous annotations using the SHANK3 transcript NM_001372044.2 in hg38. Of these, 192 variants were likely deleterious, including 27 stop-loss, 60 stop-gain, 52 frameshift, and 53 splice variants, compared to the earlier finding of 22 such variants.

N. The discovery rate of PTVs for *SHANK3* is increased from 1.3% using NM_001372044.2/hg38 as a reference to 12.5% using the transcript structure from CIS in this study.

Acknowledgement

We thank the valuable discussion and assistance of Antonio Giraldez, Hongyu Zhao, Gang Peng. And Antonio Jorge Forte. We thank Emily Qian and Rao Nivedita for editorial assistant. YHJ is supported by NIH Grants R01MH117289, R01HD088007, and R01HD088626. XL is the postdoctoral fellow of Foundation for Angelman Syndrome Therapeutics (FAST).

Author contributions

XL and YHJ conceived and designed the project. XL performed most of data collection and data analysis. PSM, YM, YW and AQH prepared and process human brain tissues. GW assisted the long-read sequencing production. MG and NP assist the data analysis. XL and YHJ wrote the manuscript together with all co-authors.

Competing interests

YHJ is a scientific co-founder of Couragene. Inc but this study is unrelated to his role. The project was supported initially by sponsored research project by Taysha Gene Therapies. Taysha Gene Therapies did not have any direct tole for the conceptualization, design, data collection, analysis, decision to publish, or preparation of the manuscript.

STAR Methods

RESOURCE AVAILABILITY

Lead contact

- Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Yong-Hui Jiang (yong-hui.jiang@yale.edu).

Materials availability

Materials availability statement

- Oligonucleotide probe panels were synthesized by Integrated DNA Technologies (IDT). The probe coverage and design are provided in Supplementary Tables.

Data and code availability

- Both human and mouse raw sequencing data have been deposited at SRA under BioProject: PRJNA1066952 and are publicly available as of the date of publication. Accession numbers are listed in the key resources table. All UCSC tracks described in manuscript have been deposited at Mendeley and are publicly available as of the date of publication. The DOI is listed in the key resources table.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Human brain tissues

Adult human cortex tissues (n=4, 24-33 years old; frontal cortex, n=2; temporal cortex, n=2) were obtained from Mayo Clinic Florida Biospecimen Bank and processed at Yale University School of Medicine. Children cortex tissues (n=4, 5-12 years old; temporal cortex, n=3; amygdala, n=1) were obtained and processed from the Children's Hospital of Fudan University in Shanghai, followed the same RNA extraction, library preparation and sequencing protocols as Yale site. The IRB protocols were approved both at Mayo Clinic Florida and the Children's Hospital of Fudan University in Shanghai.

Mice

Wild type C57BL/6J mice were obtained from the Jackson Laboratory. *Shank3* mutant mice of Shank*3* exons 4-9 deletion (*Shank3*$^{\Delta e4\text{-}9}$) [34] and *Shank3* exons 4-22 (*Shank3*$^{\Delta e4\text{-}22}$)[19] were generated and maintained in Jiang's lab. *Shank3* exon 21 deletion (*Shank3*$^{\Delta e21}$) *was obtained from Jackson Laboratory (Shank3*$^{tm1.1Pfw}$/J and Strain #:018398)[101]**.** Mice were housed of 4-5 per cage in pathogen-free mouse facility with free access to food and water on a 12-hour light: dark cycle at the ambient temperature of 20-22^{o}C and humidity of 30-70%. An equal number of male and female mice were used for all experiments. All procedures were performed following the approved animal protocol by Yale University School of Medicine Animal Care and Use Committee.

METHOD DETAILS

RNA Isolation and Quality Control

Mouse brain tissues were snap-frozen in liquid nitrogen immediately after dissection. Human brain tissues were snap-frozen in liquid nitrogen within an hour after dissection. All tissues were stored in liquid nitrogen thereafter. Total RNA was isolated from 20 mg frozen tissues, using NucleoZOL™ (Takara Bio, 740404.200) and NucleoSpin® RNA set for NucleoZOL™ (Takara Bio, 740406.50) following the manufactures specifications, followed by rDNase Set (Takara Bio, 740963) to digest DNA, and NucleoSpin® RNA Clean-up XS (Takara Bio, 740903) for RNA repurification. RNA purity (260/280, 260/230) and concentration were measured on NanoDrop™ 2000/2000c Spectrophotometers. RNA integrity number (RIN) was assessed using Agilent 2100 Bioanalyzer system.

Generation of standard and captured Iso-seq libraries

The Iso-seq libraries were prepared by following the manufacturer's instructions for each step (Iso-Seq™ Express Template Preparation for Sequel® and Sequel II Systems for standard Iso-seq; Customer Collaboration – Iso-Seq® Express Capture Using IDT xGen® Lockdown® Probes for capture Iso-seq). The 600 ng of total RNA was used as input. Only the RNA with RIN higher than 7 of human samples, and 8 of mouse samples were processed for reverse transcription, amplification, enrichment, and library preparations.

Hybridization Capture Panel Design

Hybridization capture panel design was assisted by IDT (Integrated DAN Technologies). Briefly, after extracted as 120-base-length sequence of interested gene, xGen Lockdown probes were aligned to the genome and calculated the number of possible

837 enrichment sites. A “perfect” probe was considered as only has 1 hit (the target of
838 interest) with genome, but most of the sequences returned more than 1 hit. Following
839 IDT proprietary xGen Off-Target QC Method, any probes with more than 50 hits were
840 removed because of non-specific targets in genome. The specifics and details of each
841 probe panel are presented in supplementary table 3.
842
843 **Hybridization Protocol**
844 300 ng of total RNA in less than 5.4 µL of volume mixed with 2 µL of NEBNext Single
845 Cell RT Primer Mix. The final volume was brought up to 9 µL with nuclease-free water.
846 The reaction was placed in a thermocycler and run for 5 minutes at 70°C, followed by
847 holding at 4°C for primer annealing and first-strand synthesis. Reverse transcription
848 template switching reaction was then performed by adding 5 µL of NEBNext Single Cell
849 RT Buffer, 3 µL of nuclease-free water, and 2 µL of NEBNext Single cell RT Enzyme
850 Mix to the first-strand cDNA. The reaction was incubated in a thermocycler at 42°C with
851 the lid at 52°C for 75 minutes, followed by holding at 4°C. After adding 1 µL of Iso-Seq
852 Express Template Switching oligo to the 19 µL reaction for a final volume of 20 µL, the
853 reaction was incubated again in a thermocycler at 42°C with the lid at 52°C for 15
854 minutes, followed by holding at 4°C.
855
856 The Reverse Transcription and Template Switching reaction product was then purified
857 using ProNex Beads before proceeding with cDNA amplification. For amplification, 50
858 µL of NEBNext Single Cell cDNA PCR master Mix, 2 µL of NEBNext Single Cell cDNA
859 PCR Primer, 2 µL of Iso-Seq Express cDNA PCR primer, and 0.5 µL of NEBNext Cell

Lysis Buffer were added to the purified product. The reaction was incubated in a thermocycler and run for 45 seconds at 98°C, followed by 14 cycles of the following steps: 10 seconds at 98°C, 15 seconds at 62°C, and 3 minutes at 72°C. The reaction was then held for 5 minutes at 72°C, followed by holding at 4°C. Finally, the product was purified again using ProNex Beads before proceeding with either the library preparation for standard Iso-Seq (SIS) or the capture steps for capture-based Iso-Seq (CIS).

As for the capture steps, first concentrate a total of 500ng cDNA in a 1.5 mL LoBind tube along with 7.5 µL of Cot DNA. To this mixture, add 1.8X volume of ProNex beads and gently pipette mix 10 times, followed by incubation for 10 min at room temperature. Place the tube on a magnet stand and wait until supernatant is clear. Remove the supernatant and wash twice with 200µL of freshly prepared 80% ethanol while on the magnet stand. Spin the tube strip briefly after removing the second wash, return to magnetic stand, and remove residual ethanol. Next, immediately add the hybridization reaction mix (which comprises 2X Hybridization Buffer, Hybridization Buffer Enhancer, xGen Asym TSO block, xGen RT-primer-barcode block, and 1X xGen Lockdown Panel) to elute the cDNA. Gently pipette mix 10 times and incubate for 5 min at room temperature. Then, place the tube on the magnetic stand to separate the beads from the supernatant. Transfer 17 µL of the supernatant to a new 0.2 mL PCR tube and briefly centrifuge it. Ensure that the tube is tightly sealed to prevent evaporation. Finally, place the sample tube in the thermal cycler and start the hybridization program: HYB program (lid set at 100°C), 95°C for 30 sec, 65°C for 4 hr, and lastly hold at 65°C.

During the incubation, prepare 1X working buffers and beads for capture. Preheat the wash buffers to +65°C in a heat block or water bath. To prepare the capture beads, allow the Dynabeads M-270 Streptavidin to warm to room temperature for 30 minutes prior to use. Thoroughly vortex the beads for 15 seconds to mix them, then aliquot 50 µL of beads into a 0.2 mL PCR tube, followed by adding 100 µL of 1X Bead Wash Buffer per capture, and pipette the mixture 10 times. Place the PCR tube on a magnetic rack. When the supernatant is clear, carefully remove and discard it without disturbing the beads. Note: Allow the Dynabeads to settle for at least 1 minute before removing the supernatant. Thereafter, two washes are performed as follows: Add 100 µL of 1X Bead Wash Buffer, pipette 10 times to mix, then place the PCR tube on a magnetic rack, allowing the beads to fully separate from the supernatant. Carefully remove and discard the clear supernatant. Repeat this process for a total of two washes. Finally, resuspend the beads in 17 µL of Bead Resuspension Mix per capture. The Bead Resuspension Mix includes xGen 2X Hybridization Buffer (8.5 µL), xGen Hybridization Buffer Enhancer (2.7 µL), and Nuclease-Free Water (5.8 µL). By following these steps carefully, you can ensure that the buffers and beads are prepared correctly for the capture step and obtain reliable results.

Then Bind cDNA to the capture beads, by incubating the samples in a thermocycler set to +65C for 45 minutes. Then Wash the captured cDNA with 1X wash buffers and elute the cDNA with 46ul elution buffer. To amplify the captured DNA sample, NEBNext High-Fidelity 2X PCR Master Mix is recommended, and the NEBNext Single Cell cDNA PCR

Master Mix is alternative for post capture amplification. Assemble the following PCR reaction: 50 µL of NEBNext High-Fidelity 2X PCR Master Mix, 2 µL of NEBNext Single Cell cDNA PCR Primer, 2 µL of Iso-Seq Express cDNA PCR Primer, 0.5 µL of NEBNext Cell Lysis Buffer, and 45.5 µL of the captured library. Amplify the PCR reaction mix using the following PCR protocol: Denature the DNA at 98°C for 45 seconds. Perform 14 cycles of the following steps: a. Denature the DNA at 98°C for 10 seconds. b. Anneal the primers at 62°C for 15 seconds. c. Extend the DNA at 72°C for 3 minutes. Final extension at 72°C for 5 minutes, and hold at 4°C. Finally perform the post amplification clean up steps with ProNex brands and ethanol. Use 1 µL of sample to quantifiy with Qubit dsDNA HS kit and dilute 1 µL of sample to 1.5ng/µL and run 1 µL on an Agilent Bioanalyzer using the High Sensitivity DNA kit. We used 500ng cDNA for library construction as Sequel II sequence platform required. After DNA damage repair, end repair/A-Tailing, overhang adapter ligation, and purification with ProNex Beads, the cDNA library is ready for sequencing

Sequencing Platform

To load the cDNA library onto the PacBio Sequel II System, the diffusion method was applied and followed by a 24-hour movie time and a 2-hour pre-extension time. The samples were cleaned up using ProNex beads and loaded onto the plate at a concentration of 50-100 pM.

Sequence data filtering algorithm

The following pipeline was diagramed in Supplementary Fig.1. Sequencing reads were screened initially with Lima (v2.5.0) and IsoSeq (v3). A transcript with both cDNA primers and the poly(A) was identified and called Full-length reads[102]. The Full-length reads which had less than 100 base pairs 5' end overhang, less than 30 bases pairs 3' end overhang, and less than 10 base pairs gaps in the middle are considered as the same transcript. Clustering using hierarchical alignment, and iterative cluster merging, generate polished sequence, with quality scores. The output further filtered with SQANTI3 (v4.3) after cluster and collapse to generate unique transcripts. SQANTI3 filtered the transcripts as below: If a transcript is Full-Splice Match (FSM), then it was retained unless the 3' end was unreliable (intrapriming). If a transcript was not Full-Splice Match, then it was retained only if all below were met: (1) 3' end is reliable. (2) did not have a junction that was labeled as RT-Switching. (3) all intro-exon junctions were canonical [102]. Further criteria included a transcript had to include at least 2 exons, and in the sense orientation and predicted open reading frame (ORF) had longer than 100 amino acids for the given transcript.

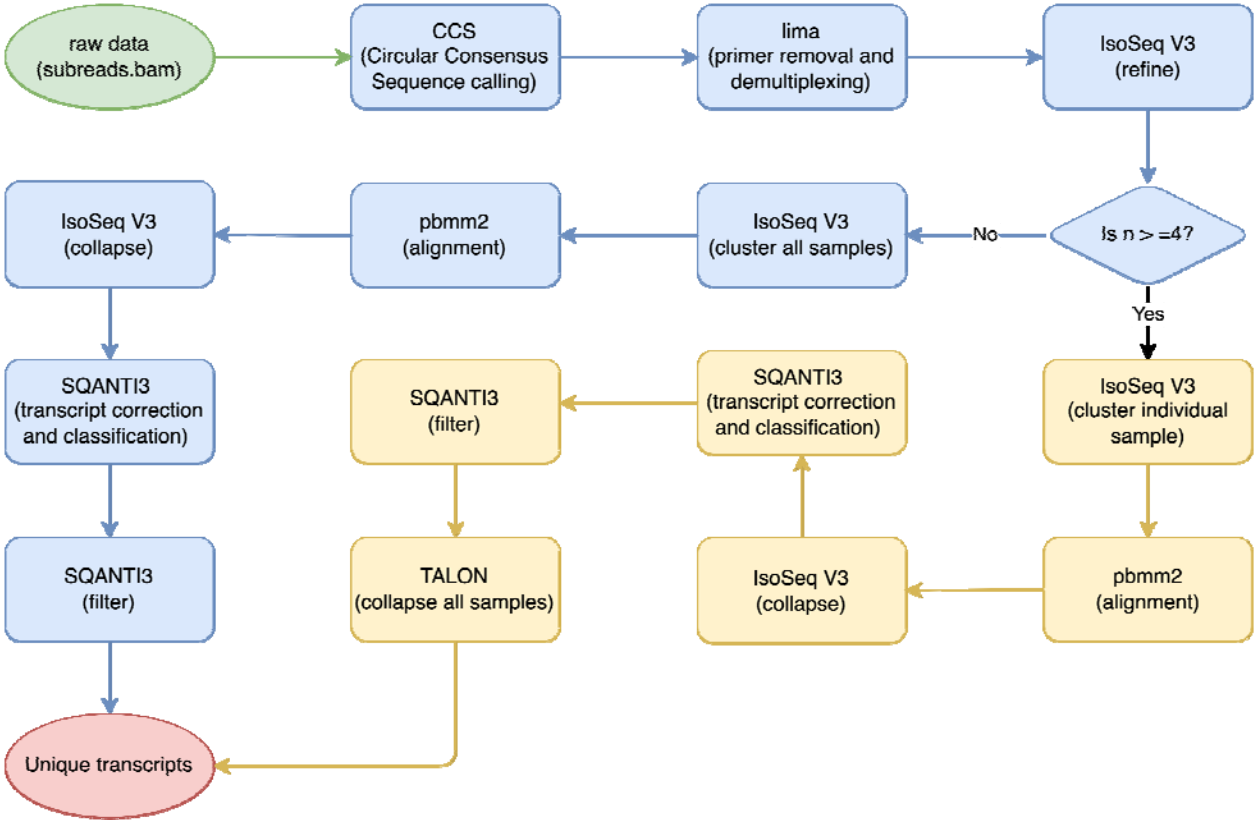
Transcript Confidence Score

To assess the quality of individual transcript, transcripts after filtering steps were scored by the following scoring metrics: (1) Score of 3 point: If the exons of transcript were presented in the sequences of by either Illumina short read methods of the bulk RNAseq (human dataset: UCLA-2022, BrainGVEX, CMC, CommonMind and LIDB) and SMART scRNAseq. (2) Score of 2 points: If a transcript had predicted ORF longer than 100AA. (3) If the abundance of a transcript were higher than 20 percentage of the rank

of the abundance of all transcripts. The summation of scores was confidence score to define each transcript: high confidence (≥ 4 points), moderate confidence (2-3 points), and low confidence (0-1 point).

Iso-seq data analysis pipeline.

The flow chart below described the analytic pipeline for ISO-Seq sequence dat. The subreads.bam file of an Iso-Seq SMRT cell was a raw input. The number of the SMRT cells, instead of the number of multiplex samples sequenced on a SMRT cell, regardless of the library preparation methods [Sta-Iso-Seq (SIS) or Cap-Iso-Seq (CIS)], dictated the direction of the analysis flow.



Real-time quantitative polymerase chain reaction (RT-qPCR)

964 Two µg of total RNA was reverse transcribed into cDNA templates using RNA to cDNA
965 EcoDry™ Premix kit including both random hexamer and oligo(dT)$_{18}$ primers (Takara
966 Bio, 639548). KAPA SYBR® FAST qPCR Master Mix (2X) Universal (Kapa Biosystems,
967 KK4602) was used for qPCR reactions with 18 ng of cDNA as template input. The
968 following program on CFX96 Touch Real-Time PCR Detection System (BIO-RAD) was
969 used: 3 minutes at 95^{o}C for enzyme activation, followed by 40 cycles of denaturation
970 (95^{o}C, 3 seconds) and annealing, extension, data acquisition (60^{o}C, 30 seconds),
971 followed by dissociation and holding at 4^{o}C. The PCR primers are shown in
972 supplementary Table 3.
973
974 **RNA-Seq Data Processing**
975 Illumina bulk RNAseq raw data in FASTQ format after quality control and filtering with
976 fastp [103], and SMART scRNAseq FASTQ data, were aligned to hg38 for human
977 sequences and mm39 for mouse sequences using HISAT 2.2.1 [104]. Aligned RNA-Seq
978 data (aligned to hg37/38) in BAM format were converted to FASTQ format using
979 SAMtools [105] when the raw FASTQ was not available, followed by the same process as
980 above. Gene expression counts and DEXSeq-counts were calculated using
981 FeatureCount [106] for further gene expression and exon usage analysis. Detailed
982 RNAseq datasets information summarized in supplementary table 4.
983
984 **Differential Transcript Usage**
985 Transcript-level quantification of the processed RNA-Seq data was performed using the
986 software Salmon 1.4.0 [107]. The transcriptome index used for quantification was built

from the reference genome annotation (in GTF format), along with the reference genome FASTA file. Transcript abundances were estimated using the quasi-mapping algorithm (--quasiMAP) mode, which performs a lightweight alignment-free estimation of abundances based on k-mer matching. The output files were generated in TPM (transcripts per million) format.

Differential Exon Usage (DEU)

DEXSeq-counts tables were imported into R, analysis with R package DEXSeq[108]. Normalization and filtering were performed to remove lowly expressed exons. DexSeq uses a binomial generalized linear model to estimate exon expression, accounting for the variability in exon-exon junction usage across samples. DEU was then tested using the DEXSeq function, which fits a statistical model to test for differences in exon usage between two or more groups of samples. Exons with an adjusted p-value ≤ 0.05 and a log2 fold change ≥ 1 or ≤ -1 were considered significantly differentially used, and visualized with built-in function of DEXSeq.

Whole Genome Sequencing and Exome Analysis

DNA variation data post variation calling in VCF format were downloaded from Autism Sequencing Consortium (ASC), Bipolar Exomes (BipEx), whole-exome sequencing case-control study of epilepsy (Epi25), Schizophrenia exome meta-analysis consortium (SCHEMA), and PsychENCODE. VCFs initially aligned to hg38 (BipEx and Epi25) and the datasets (ASC, SCHEMA and PsychENCODE) after alignment lift over from hg37 to hg38 with UCSC LiftOver tool and chain file, were subsetted to the region of interest

(SHANK3, chr22:50670000-50770000) using BCFtools (v 1.16) (Danecek et al. 2021). The data format was modified using HTSlib (v 1.16)[109] and TAB-delimited file InderXer (Tabix, v 0.2.5)[110]. Then the data were annotated with Ensembl Variant Effect Predictor (VET, release 107)(McLaren et al. 2016) and filtered with Genome Aggregation Database (gnomAD, v3.1.2)[72] by INFO/AF_popmax<=0.01. Filtered DNA variation were aligned to novel exons detected in SIS and CIS with SpliceAI [73] for splicing event analysis, and with SnpEff [74] to evaluate other deleterious SNV (stop lost, stop gain and frameshift).

Data Visualization

Visualization was performed using ggplot2 (version 3.3.2) in R (version 4.2.2) for plotting gene expression, transcript and exon usage profiles and heatmaps.

Spatial Transcriptional Analysis

An open access Visium dataset of mouse brain coronal section from 10x Genomics[111] in FASTQ format was analyzed using customized references and annotation generated from mouse *Shank3* CIS transcripts using Cell Ranger [112], followed by quantitation with customized probe-set (probe-transcripts relation spreadsheet) using 10x Genomics Space Ranger v2.0. The output cloupe file was visualized using 10x Genomics Loupe Visualization Software v6.5.

KEY RESOURCES TABLE

Submitted as a separate file

Reference

1. Park, E., Pan, Z., Zhang, Z., Lin, L., and Xing, Y. (2018). The Expanding Landscape of Alternative Splicing Variation in Human Populations. Am J Hum Genet *102*, 11-26. 10.1016/j.ajhg.2017.11.002.
2. Blencowe, B.J. (2017). The Relationship between Alternative Splicing and Proteomic Complexity. Trends Biochem Sci *42*, 407-408. 10.1016/j.tibs.2017.04.001.
3. Raj, B., and Blencowe, B.J. (2015). Alternative Splicing in the Mammalian Nervous System: Recent Insights into Mechanisms and Functional Roles. Neuron *87*, 14-27. 10.1016/j.neuron.2015.05.004.
4. Ray, T.A., Cochran, K., Kozlowski, C., Wang, J., Alexander, G., Cady, M.A., Spencer, W.J., Ruzycki, P.A., Clark, B.S., Laeremans, A., et al. (2020). Comprehensive identification of mRNA isoforms reveals the diversity of neural cell-surface molecules with roles in retinal development and disease. Nat Commun *11*, 3328. 10.1038/s41467-020-17009-7.
5. Gandal, M.J., Zhang, P., Hadjimichael, E., Walker, R.L., Chen, C., Liu, S., Won, H., van Bakel, H., Varghese, M., Wang, Y., et al. (2018). Transcriptome-wide isoform-level dysregulation in ASD, schizophrenia, and bipolar disorder. Science *362*. 10.1126/science.aat8127.
6. Patowary, A., Zhang, P., Jops, C., Vuong, C.K., Ge, X., Hou, K., Kim, M., Gong, N., Margolis, M., Vo, D., et al. (2023). Cell-type-specificity of isoform diversity in the developing human neocortex informs mechanisms of neurodevelopmental disorders. bioRxiv. 10.1101/2023.03.25.534016.
7. Ollà, I., Pardiñas, A.F., Parras, A., Hernández, I.H., Santos-Galindo, M., Picó, S., Callado, L.F., Elorza, A., Rodríguez-López, C., Fernández-Miranda, G., et al. (2023). Pathogenic mis-splicing of CPEB4 in schizophrenia. Biol Psychiatry. 10.1016/j.biopsych.2023.03.010.
8. Glinos, D.A., Garborcauskas, G., Hoffman, P., Ehsan, N., Jiang, L., Gokden, A., Dai, X., Aguet, F., Brown, K.L., Garimella, K., et al. (2022). Transcriptome variation in human tissues revealed by long-read sequencing. Nature *608*, 353-359. 10.1038/s41586-022-05035-y.
9. Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet *10*, 57-63. 10.1038/nrg2484.
10. Mercer, T.R., Clark, M.B., Crawford, J., Brunck, M.E., Gerhardt, D.J., Taft, R.J., Nielsen, L.K., Dinger, M.E., and Mattick, J.S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. Nat Protoc *9*, 989-1009. 10.1038/nprot.2014.058.
11. Clark, M.B., Amaral, P.P., Schlesinger, F.J., Dinger, M.E., Taft, R.J., Rinn, J.L., Ponting, C.P., Stadler, P.F., Morris, K.V., Morillon, A., et al. (2011). The reality of pervasive transcription. PLoS Biol *9*, e1000625; discussion e1001102. 10.1371/journal.pbio.1000625.
12. Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816. 10.1038/nature05874.

13. van Bakel, H., Nislow, C., Blencowe, B.J., and Hughes, T.R. (2010). Most "dark matter" transcripts are associated with known genes. PLoS Biol *8*, e1000371. 10.1371/journal.pbio.1000371.
14. Villa, T., and Porrua, O. (2023). Pervasive transcription: a controlled risk. Febs j *290*, 3723-3736. 10.1111/febs.16530.
15. Djebali, S., Davis, C.A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., et al. (2012). Landscape of transcription in human cells. Nature *489*, 101-108. 10.1038/nature11233.
16. Schmeisser, M.J., Ey, E., Wegener, S., Bockmann, J., Stempel, A.V., Kuebler, A., Janssen, A.L., Udvardi, P.T., Shiban, E., Spilker, C., et al. (2012). Autistic-like behaviours and hyperactivity in mice lacking ProSAP1/Shank2. Nature *486*, 256-260. 10.1038/nature11015.
17. Peça, J., Feliciano, C., Ting, J.T., Wang, W., Wells, M.F., Venkatraman, T.N., Lascola, C.D., Fu, Z., and Feng, G. (2011). Shank3 mutant mice display autistic-like behaviours and striatal dysfunction. Nature *472*, 437-442. 10.1038/nature09965.
18. Jiang, Y.H., and Ehlers, M.D. (2013). Modeling autism by SHANK gene mutations in mice. Neuron *78*, 8-27. 10.1016/j.neuron.2013.03.016.
19. Wang, X., Bey, A.L., Katz, B.M., Badea, A., Kim, N., David, L.K., Duffney, L.J., Kumar, S., Mague, S.D., Hulbert, S.W., et al. (2016). Altered mGluR5-Homer scaffolds and corticostriatal connectivity in a Shank3 complete knockout model of autism. Nat Commun *7*, 11459. 10.1038/ncomms11459.
20. Speed, H.E., Kouser, M., Xuan, Z., Reimers, J.M., Ochoa, C.F., Gupta, N., Liu, S., and Powell, C.M. (2015). Autism-Associated Insertion Mutation (InsG) of Shank3 Exon 21 Causes Impaired Synaptic Transmission and Behavioral Deficits. J Neurosci *35*, 9648-9665. 10.1523/jneurosci.3125-14.2015.
21. Jaramillo, T.C., Speed, H.E., Xuan, Z., Reimers, J.M., Escamilla, C.O., Weaver, T.P., Liu, S., Filonova, I., and Powell, C.M. (2017). Novel Shank3 mutant exhibits behaviors with face validity for autism and altered striatal and hippocampal function. Autism Res *10*, 42-65. 10.1002/aur.1664.
22. Duffney, L.J., Zhong, P., Wei, J., Matas, E., Cheng, J., Qin, L., Ma, K., Dietz, D.M., Kajiwara, Y., Buxbaum, J.D., and Yan, Z. (2015). Autism-like Deficits in Shank3-Deficient Mice Are Rescued by Targeting Actin Regulators. Cell Rep *11*, 1400-1413. 10.1016/j.celrep.2015.04.064.
23. Zhou, Y., Kaiser, T., Monteiro, P., Zhang, X., Van der Goes, M.S., Wang, D., Barak, B., Zeng, M., Li, C., Lu, C., et al. (2016). Mice with Shank3 Mutations Associated with ASD and Schizophrenia Display Both Shared and Distinct Defects. Neuron *89*, 147-162. 10.1016/j.neuron.2015.11.023.
24. Lee, J., Chung, C., Ha, S., Lee, D., Kim, D.Y., Kim, H., and Kim, E. (2015). Shank3-mutant mice lacking exon 9 show altered excitation/inhibition balance, enhanced rearing, and spatial memory deficit. Front Cell Neurosci *9*, 94. 10.3389/fncel.2015.00094.
25. Wang, X., Xu, Q., Bey, A.L., Lee, Y., and Jiang, Y.H. (2014). Transcriptional and functional complexity of Shank3 provides a molecular framework to understand the phenotypic heterogeneity of SHANK3 causing autism and Shank3 mutant mice. Mol Autism *5*, 30. 10.1186/2040-2392-5-30.
26. Bouquier, N., Sakkaki, S., Raynaud, F., Hemonnot-Girard, A.L., Seube, V., Compan, V., Bertaso, F., Perroy, J., and Moutin, E. (2022). The Shank3(Venus/Venus) knock in

mouse enables isoform-specific functional studies of Shank3a. Front Neurosci *16*, 1081010. 10.3389/fnins.2022.1081010.

27. Yoo, T., Yoo, Y.E., Kang, H., and Kim, E. (2022). Age, brain region, and gene dosage-differential transcriptomic changes in Shank3-mutant mice. Front Mol Neurosci *15*, 1017512. 10.3389/fnmol.2022.1017512.
28. Yoo, Y.E., Yoo, T., Kang, H., and Kim, E. (2022). Brain region and gene dosage-differential transcriptomic changes in Shank2-mutant mice. Front Mol Neurosci *15*, 977305. 10.3389/fnmol.2022.977305.
29. Lim, S., Naisbitt, S., Yoon, J., Hwang, J.I., Suh, P.G., Sheng, M., and Kim, E. (1999). Characterization of the Shank family of synaptic proteins. Multiple genes, alternative splicing, and differential expression in brain and development. J Biol Chem *274*, 29510-29518. 10.1074/jbc.274.41.29510.
30. Delling, J.P., and Boeckers, T.M. (2021). Comparison of SHANK3 deficiency in animal models: phenotypes, treatment strategies, and translational implications. J Neurodev Disord *13*, 55. 10.1186/s11689-021-09397-8.
31. Tian, R., Li, Y., Zhao, H., Lyu, W., Zhao, J., Wang, X., Lu, H., Xu, H., Ren, W., Tan, Q.Q., et al. (2023). Modeling SHANK3-associated autism spectrum disorder in Beagle dogs via CRISPR/Cas9 gene editing. Mol Psychiatry. 10.1038/s41380-023-02276-9.
32. Jaramillo, T.C., Speed, H.E., Xuan, Z., Reimers, J.M., Liu, S., and Powell, C.M. (2016). Altered Striatal Synaptic Function and Abnormal Behaviour in Shank3 Exon4-9 Deletion Mouse Model of Autism. Autism Res *9*, 350-375. 10.1002/aur.1529.
33. Drapeau, E., Dorr, N.P., Elder, G.A., and Buxbaum, J.D. (2014). Absence of strong strain effects in behavioral analyses of Shank3-deficient mice. Dis Model Mech *7*, 667-681. 10.1242/dmm.013821.
34. Wang, X., McCoy, P.A., Rodriguiz, R.M., Pan, Y., Je, H.S., Roberts, A.C., Kim, C.J., Berrios, J., Colvin, J.S., Bousquet-Moore, D., et al. (2011). Synaptic dysfunction and abnormal behaviors in mice lacking major isoforms of Shank3. Hum Mol Genet *20*, 3093-3108. 10.1093/hmg/ddr212.
35. Akbarian, S., Liu, C., Knowles, J.A., Vaccarino, F.M., Farnham, P.J., Crawford, G.E., Jaffe, A.E., Pinto, D., Dracheva, S., Geschwind, D.H., et al. (2015). The PsychENCODE project. Nat Neurosci *18*, 1707-1712. 10.1038/nn.4156.
36. Wang, D., Liu, S., Warrell, J., Won, H., Shi, X., Navarro, F.C.P., Clarke, D., Gu, M., Emani, P., Yang, Y.T., et al. (2018). Comprehensive functional genomic resource and integrative model for the human brain. Science *362*. 10.1126/science.aat8464.
37. Ramaker, R.C., Bowling, K.M., Lasseigne, B.N., Hagenauer, M.H., Hardigan, A.A., Davis, N.S., Gertz, J., Cartagena, P.M., Walsh, D.M., Vawter, M.P., et al. (2017). Post-mortem molecular profiling of three psychiatric disorders. Genome Med *9*, 72. 10.1186/s13073-017-0458-5.
38. Srinivasan, K., Friedman, B.A., Etxeberria, A., Huntley, M.A., van der Brug, M.P., Foreman, O., Paw, J.S., Modrusan, Z., Beach, T.G., Serrano, G.E., and Hansen, D.V. (2020). Alzheimer's Patient Microglia Exhibit Enhanced Aging and Unique Transcriptional Activation. Cell Rep *31*, 107843. 10.1016/j.celrep.2020.107843.
39. Yao, Z., van Velthoven, C.T.J., Nguyen, T.N., Goldy, J., Sedeno-Cortes, A.E., Baftizadeh, F., Bertagnolli, D., Casper, T., Chiang, M., Crichton, K., et al. (2021). A taxonomy of transcriptomic cell types across the isocortex and hippocampal formation. Cell *184*, 3222-3241.e3226. 10.1016/j.cell.2021.04.021.

40. Ihara, M., Yamasaki, N., Hagiwara, A., Tanigaki, A., Kitano, A., Hikawa, R., Tomimoto, H., Noda, M., Takanashi, M., Mori, H., et al. (2007). Sept4, a component of presynaptic scaffold and Lewy bodies, is required for the suppression of alpha-synuclein neurotoxicity. Neuron *53*, 519-533. 10.1016/j.neuron.2007.01.019.
41. Lin, W.H., Chiu, K.C., Chang, H.M., Lee, K.C., Tai, T.Y., and Chuang, L.M. (2001). Molecular scanning of the human sorbin and SH3-domain-containing-1 (SORBS1) gene: positive association of the T228A polymorphism with obesity and type 2 diabetes. Hum Mol Genet *10*, 1753-1760. 10.1093/hmg/10.17.1753.
42. Leung, S.K., Jeffries, A.R., Castanho, I., Jordan, B.T., Moore, K., Davies, J.P., Dempster, E.L., Bray, N.J., O'Neill, P., Tseng, E., et al. (2021). Full-length transcript sequencing of human and mouse cerebral cortex identifies widespread isoform diversity and alternative splicing. Cell Rep *37*, 110022. 10.1016/j.celrep.2021.110022.
43. Fu, J.M., Satterstrom, F.K., Peng, M., Brand, H., Collins, R.L., Dong, S., Wamsley, B., Klei, L., Wang, L., Hao, S.P., et al. (2022). Rare coding variation provides insight into the genetic architecture and phenotypic context of autism. Nat Genet *54*, 1320-1331. 10.1038/s41588-022-01104-0.
44. Zhou, X., Feliciano, P., Shu, C., Wang, T., Astrovskaya, I., Hall, J.B., Obiajulu, J.U., Wright, J.R., Murali, S.C., Xu, S.X., et al. (2022). Integrating de novo and inherited variants in 42,607 autism cases identifies mutations in new moderate-risk genes. Nat Genet *54*, 1305-1319. 10.1038/s41588-022-01148-2.
45. Satterstrom, F.K., Kosmicki, J.A., Wang, J., Breen, M.S., De Rubeis, S., An, J.Y., Peng, M., Collins, R., Grove, J., Klei, L., et al. (2020). Large-Scale Exome Sequencing Study Implicates Both Developmental and Functional Changes in the Neurobiology of Autism. Cell *180*, 568-584.e523. 10.1016/j.cell.2019.12.036.
46. Flörke-Gerloff, S., Töpfer-Petersen, E., Müller-Esterl, W., Schill, W.B., and Engel, W. (1983). Acrosin and the acrosome in human spermatogenesis. Hum Genet *65*, 61-67. 10.1007/bf00285030.
47. Monteiro, P., and Feng, G. (2017). SHANK proteins: roles at the synapse and in autism spectrum disorder. Nat Rev Neurosci *18*, 147-157. 10.1038/nrn.2016.183.
48. Tu, J.C., Xiao, B., Naisbitt, S., Yuan, J.P., Petralia, R.S., Brakeman, P., Doan, A., Aakalu, V.K., Lanahan, A.A., Sheng, M., and Worley, P.F. (1999). Coupling of mGluR/Homer and PSD-95 complexes by the Shank family of postsynaptic density proteins. Neuron *23*, 583-592. 10.1016/s0896-6273(00)80810-7.
49. Naisbitt, S., Kim, E., Tu, J.C., Xiao, B., Sala, C., Valtschanoff, J., Weinberg, R.J., Worley, P.F., and Sheng, M. (1999). Shank, a novel family of postsynaptic density proteins that binds to the NMDA receptor/PSD-95/GKAP complex and cortactin. Neuron *23*, 569-582. 10.1016/s0896-6273(00)80809-0.
50. Tang, S., Lomsadze, A., and Borodovsky, M. (2015). Identification of protein coding regions in RNA transcripts. Nucleic Acids Res *43*, e78. 10.1093/nar/gkv227.
51. Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nature Protocols *7*, 562-578. 10.1038/nprot.2012.016.
52. Cooper, G.M., Stone, E.A., Asimenos, G., Green, E.D., Batzoglou, S., and Sidow, A. (2005). Distribution and intensity of constraint in mammalian genomic sequence. Genome Res *15*, 901-913. 10.1101/gr.3577405.

53. Davydov, E.V., Goode, D.L., Sirota, M., Cooper, G.M., Sidow, A., and Batzoglou, S. (2010). Identifying a high fraction of the human genome to be under selective constraint using GERP++. PLoS Comput Biol *6*, e1001025. 10.1371/journal.pcbi.1001025.

54. Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050. 10.1101/gr.3715005.

55. Marcel, V., Dichtel-Danjoy, M.L., Sagne, C., Hafsi, H., Ma, D., Ortiz-Cuaran, S., Olivier, M., Hall, J., Mollereau, B., Hainaut, P., and Bourdon, J.C. (2011). Biological functions of p53 isoforms through evolution: lessons from animal and cellular models. Cell Death Differ *18*, 1815-1824. 10.1038/cdd.2011.120.

56. Khoury, M.P., and Bourdon, J.C. (2010). The isoforms of the p53 protein. Cold Spring Harb Perspect Biol *2*, a000927. 10.1101/cshperspect.a000927.

57. Ayoub, A.E., Oh, S., Xie, Y., Leng, J., Cotney, J., Dominguez, M.H., Noonan, J.P., and Rakic, P. (2011). Transcriptional programs in transient embryonic zones of the cerebral cortex defined by high-resolution mRNA sequencing. Proc Natl Acad Sci U S A *108*, 14950-14955. 10.1073/pnas.1112213108.

58. Belgard, T.G., Marques, A.C., Oliver, P.L., Abaan, H.O., Sirey, T.M., Hoerder-Suabedissen, A., García-Moreno, F., Molnár, Z., Margulies, E.H., and Ponting, C.P. (2011). A transcriptomic atlas of mouse neocortical layers. Neuron *71*, 605-616. 10.1016/j.neuron.2011.06.039.

59. Fertuzinhos, S., Li, M., Kawasawa, Y.I., Ivic, V., Franjic, D., Singh, D., Crair, M., and Sestan, N. (2014). Laminar and temporal expression dynamics of coding and noncoding RNAs in the mouse neocortex. Cell Rep *6*, 938-950. 10.1016/j.celrep.2014.01.036.

60. Genomics, x. (2022). FFPE Mouse Brain Coronal Section 1 (FFPE), Spatial Gene Expression Dataset by Space Ranger 2.0.0.

61. Niu, M., Cao, W., Wang, Y., Zhu, Q., Luo, J., Wang, B., Zheng, H., Weitz, D.A., and Zong, C. (2023). Droplet-based transcriptome profiling of individual synapses. Nat Biotechnol. 10.1038/s41587-022-01635-1.

62. Gauthier, J., Champagne, N., Lafrenière, R.G., Xiong, L., Spiegelman, D., Brustein, E., Lapointe, M., Peng, H., Côté, M., Noreau, A., et al. (2010). De novo mutations in the gene encoding the synaptic scaffolding protein SHANK3 in patients ascertained for schizophrenia. Proc Natl Acad Sci U S A *107*, 7863-7868. 10.1073/pnas.0906232107.

63. Vucurovic, K., Landais, E., Delahaigue, C., Eutrope, J., Schneider, A., Leroy, C., Kabbaj, H., Motte, J., Gaillard, D., Rolland, A.C., and Doco-Fenzy, M. (2012). Bipolar affective disorder and early dementia onset in a male patient with SHANK3 deletion. Eur J Med Genet *55*, 625-629. 10.1016/j.ejmg.2012.07.009.

64. Levy, T., Foss-Feig, J.H., Betancur, C., Siper, P.M., Trelles-Thorne, M.D.P., Halpern, D., Frank, Y., Lozano, R., Layton, C., Britvan, B., et al. (2022). Strong evidence for genotype-phenotype correlations in Phelan-McDermid syndrome: results from the developmental synaptopathies consortium. Hum Mol Genet *31*, 625-637. 10.1093/hmg/ddab280.

65. Moessner, R., Marshall, C.R., Sutcliffe, J.S., Skaug, J., Pinto, D., Vincent, J., Zwaigenbaum, L., Fernandez, B., Roberts, W., Szatmari, P., and Scherer, S.W. (2007). Contribution of SHANK3 mutations to autism spectrum disorder. Am J Hum Genet *81*, 1289-1297. 10.1086/522590.

66. Leblond, C.S., Nava, C., Polge, A., Gauthier, J., Huguet, G., Lumbroso, S., Giuliano, F., Stordeur, C., Depienne, C., Mouzat, K., et al. (2014). Meta-analysis of SHANK Mutations in Autism Spectrum Disorders: a gradient of severity in cognitive impairments. PLoS Genet *10*, e1004580. 10.1371/journal.pgen.1004580.
67. Singh, T., Poterba, T., Curtis, D., Akil, H., Al Eissa, M., Barchas, J.D., Bass, N., Bigdeli, T.B., Breen, G., Bromet, E.J., et al. (2022). Rare coding variants in ten genes confer substantial risk for schizophrenia. Nature *604*, 509-516. 10.1038/s41586-022-04556-w.
68. Palmer, D.S., Howrigan, D.P., Chapman, S.B., Adolfsson, R., Bass, N., Blackwood, D., Boks, M.P.M., Chen, C.Y., Churchhouse, C., Corvin, A.P., et al. (2022). Exome sequencing in bipolar disorder identifies AKAP11 as a risk gene shared with schizophrenia. Nat Genet *54*, 541-547. 10.1038/s41588-022-01034-x.
69. Werling, D.M., Pochareddy, S., Choi, J., An, J.Y., Sheppard, B., Peng, M., Li, Z., Dastmalchi, C., Santpere, G., Sousa, A.M.M., et al. (2020). Whole-Genome and RNA Sequencing Reveal Variation and Transcriptomic Coordination in the Developing Human Prefrontal Cortex. Cell Rep *31*, 107489. 10.1016/j.celrep.2020.03.053.
70. Bryois, J., Garrett, M.E., Song, L., Safi, A., Giusti-Rodriguez, P., Johnson, G.D., Shieh, A.W., Buil, A., Fullard, J.F., Roussos, P., et al. (2018). Evaluation of chromatin accessibility in prefrontal cortex of individuals with schizophrenia. Nat Commun *9*, 3121. 10.1038/s41467-018-05379-y.
71. McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biol *17*, 122. 10.1186/s13059-016-0974-4.
72. Koch, L. (2020). Exploring human genomic diversity with gnomAD. Nature Reviews Genetics *21*, 448-448. 10.1038/s41576-020-0255-7.
73. Jaganathan, K., Kyriazopoulou Panagiotopoulou, S., McRae, J.F., Darbandi, S.F., Knowles, D., Li, Y.I., Kosmicki, J.A., Arbelaez, J., Cui, W., Schwartz, G.B., et al. (2019). Predicting Splicing from Primary Sequence with Deep Learning. Cell *176*, 535-548.e524. 10.1016/j.cell.2018.12.015.
74. Cingolani, P., Platts, A., Wang le, L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80-92. 10.4161/fly.19695.
75. Nurk, S., Koren, S., Rhie, A., Rautiainen, M., Bzikadze, A.V., Mikheenko, A., Vollger, M.R., Altemose, N., Uralsky, L., Gershman, A., et al. (2022). The complete sequence of a human genome. Science *376*, 44-53. 10.1126/science.abj6987.
76. Chau, K.K., Zhang, P., Urresti, J., Amar, M., Pramod, A.B., Chen, J., Thomas, A., Corominas, R., Lin, G.N., and Iakoucheva, L.M. (2021). Full-length isoform transcriptome of the developing human brain provides further insights into autism. Cell Rep *36*, 109631. 10.1016/j.celrep.2021.109631.
77. Mertens, F., Johansson, B., Fioretos, T., and Mitelman, F. (2015). The emerging complexity of gene fusions in cancer. Nat Rev Cancer *15*, 371-381. 10.1038/nrc3947.
78. Dorney, R., Dhungel, B.P., Rasko, J.E.J., Hebbard, L., and Schmitz, U. (2023). Recent advances in cancer fusion transcript detection. Brief Bioinform *24*. 10.1093/bib/bbac519.
79. Mehani, B., Narta, K., Paul, D., Raj, A., Kumar, D., Sharma, A., Kaurani, L., Nayak, S., Dash, D., Suri, A., et al. (2020). Fusion transcripts in normal human cortex increase with

age and show distinct genomic features for single cells and tissues. Sci Rep *10*, 1368. 10.1038/s41598-020-58165-6.

80. Gandal, M.J., Haney, J.R., Wamsley, B., Yap, C.X., Parhami, S., Emani, P.S., Chang, N., Chen, G.T., Hoftman, G.D., de Alba, D., et al. (2022). Broad transcriptomic dysregulation occurs across the cerebral cortex in ASD. Nature *611*, 532-539. 10.1038/s41586-022-05377-7.

81. Beri, S., Tonna, N., Menozzi, G., Bonaglia, M.C., Sala, C., and Giorda, R. (2007). DNA methylation regulates tissue-specific expression of Shank3. J Neurochem *101*, 1380-1391. 10.1111/j.1471-4159.2007.04539.x.

82. Zhu, L., Wang, X., Li, X.L., Towers, A., Cao, X., Wang, P., Bowman, R., Yang, H., Goldstein, J., Li, Y.J., and Jiang, Y.H. (2014). Epigenetic dysregulation of SHANK3 in brain tissues from individuals with autism spectrum disorders. Hum Mol Genet *23*, 1563-1578. 10.1093/hmg/ddt547.

83. Bey, A.L., Wang, X., Yan, H., Kim, N., Passman, R.L., Yang, Y., Cao, X., Towers, A.J., Hulbert, S.W., Duffney, L.J., et al. (2018). Brain region-specific disruption of Shank3 in mice reveals a dissociation for cortical and striatal circuits in autism-related behaviors. Transl Psychiatry *8*, 94. 10.1038/s41398-018-0142-6.

84. Won, H., Lee, H.R., Gee, H.Y., Mah, W., Kim, J.I., Lee, J., Ha, S., Chung, C., Jung, E.S., Cho, Y.S., et al. (2012). Autistic-like social behaviour in Shank2-mutant mice improved by restoring NMDA receptor function. Nature *486*, 261-265. 10.1038/nature11208.

85. Han, K., Holder, J.L., Jr., Schaaf, C.P., Lu, H., Chen, H., Kang, H., Tang, J., Wu, Z., Hao, S., Cheung, S.W., et al. (2013). SHANK3 overexpression causes manic-like behaviour with unique pharmacogenetic properties. Nature *503*, 72-77. 10.1038/nature12630.

86. Qin, L., Ma, K., Wang, Z.J., Hu, Z., Matas, E., Wei, J., and Yan, Z. (2018). Social deficits in Shank3-deficient mouse models of autism are rescued by histone deacetylase (HDAC) inhibition. Nat Neurosci *21*, 564-575. 10.1038/s41593-018-0110-8.

87. Ecker, J.R., Geschwind, D.H., Kriegstein, A.R., Ngai, J., Osten, P., Polioudakis, D., Regev, A., Sestan, N., Wickersham, I.R., and Zeng, H. (2017). The BRAIN Initiative Cell Census Consortium: Lessons Learned toward Generating a Comprehensive Brain Cell Atlas. Neuron *96*, 542-557. 10.1016/j.neuron.2017.10.007.

88. Raj, A., and van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. Cell *135*, 216-226. 10.1016/j.cell.2008.09.050.

89. Gupta, A., Martin-Rufino, J.D., Jones, T.R., Subramanian, V., Qiu, X., Grody, E.I., Bloemendal, A., Weng, C., Niu, S.Y., Min, K.H., et al. (2022). Inferring gene regulation from stochastic transcriptional variation across single cells at steady state. Proc Natl Acad Sci U S A *119*, e2207392119. 10.1073/pnas.2207392119.

90. Thattai, M., and van Oudenaarden, A. (2001). Intrinsic noise in gene regulatory networks. Proc Natl Acad Sci U S A *98*, 8614-8619. 10.1073/pnas.151588598.

91. Bohrer, C.H., and Larson, D.R. (2021). The Stochastic Genome and Its Role in Gene Expression. Cold Spring Harb Perspect Biol *13*. 10.1101/cshperspect.a040386.

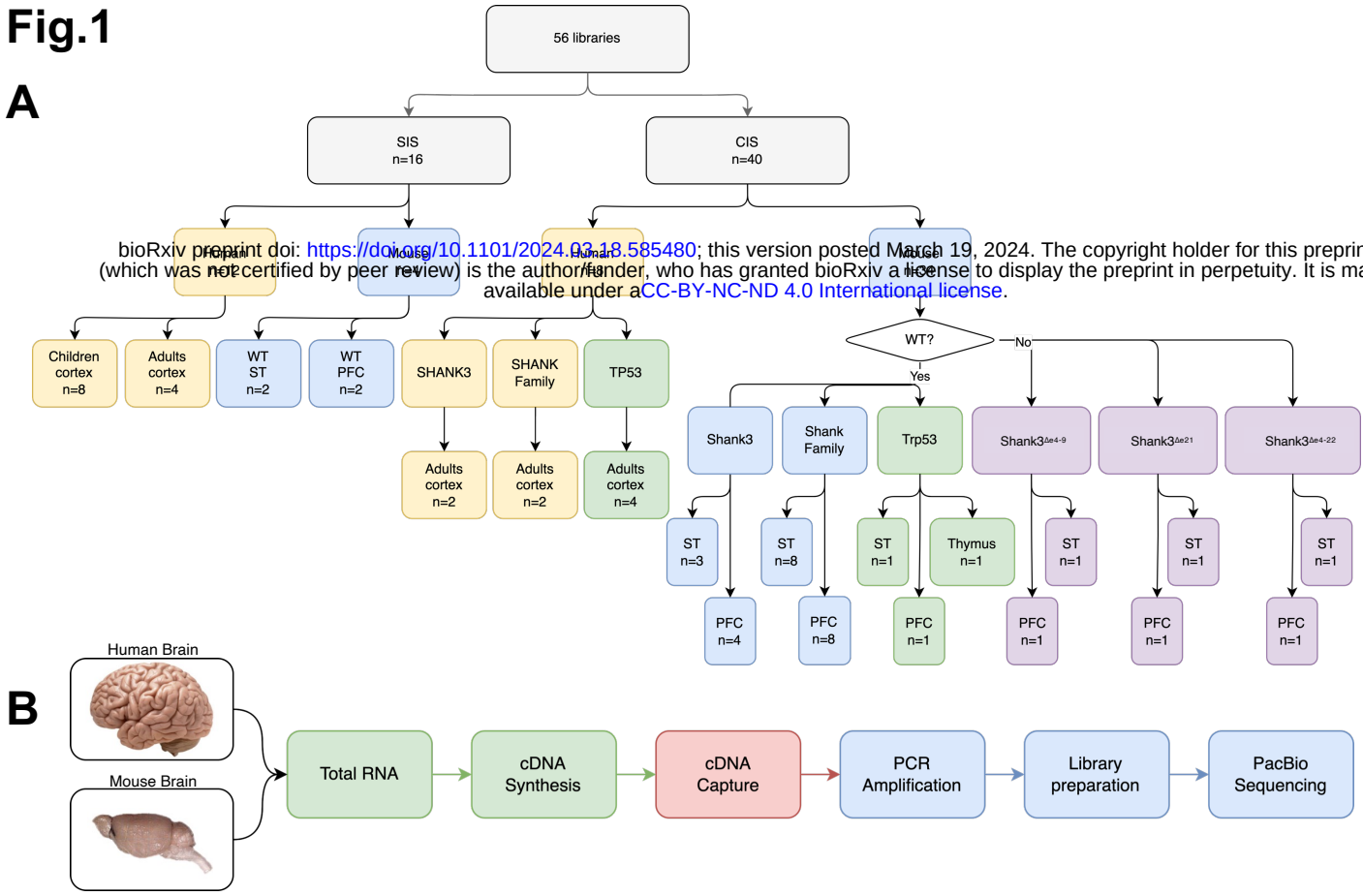
92. Raser, J.M., and O'Shea, E.K. (2005). Noise in gene expression: origins, consequences, and control. Science *309*, 2010-2013. 10.1126/science.1105891.

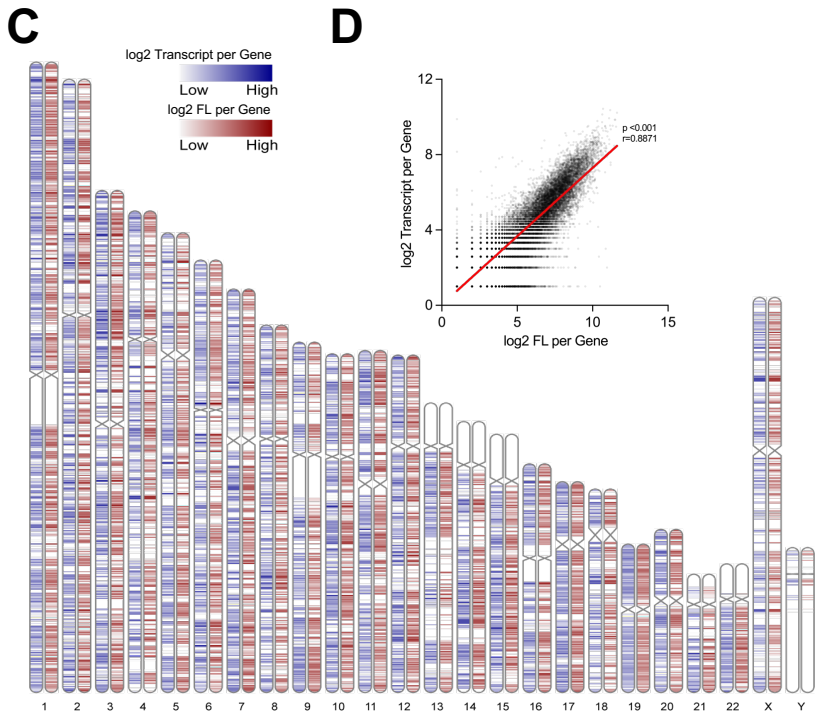
93. Girbig, M., Misiaszek, A.D., and Müller, C.W. (2022). Structural insights into nuclear transcription by eukaryotic DNA-dependent RNA polymerases. Nat Rev Mol Cell Biol *23*, 603-622. 10.1038/s41580-022-00476-9.

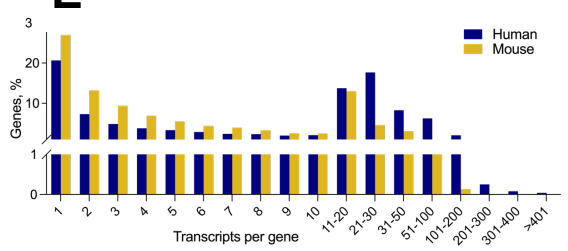
94. Agapov, A., Olina, A., and Kulbachinskiy, A. (2022). RNA polymerase pausing, stalling and bypass during transcription of damaged DNA: from molecular basis to functional consequences. Nucleic Acids Res *50*, 3018-3041. 10.1093/nar/gkac174.
95. Vassylyev, D.G., Vassylyeva, M.N., Perederina, A., Tahirov, T.H., and Artsimovitch, I. (2007). Structural basis for transcription elongation by bacterial RNA polymerase. Nature *448*, 157-162. 10.1038/nature05932.
96. Landick, R. (2001). RNA polymerase clamps down. Cell *105*, 567-570. 10.1016/s0092-8674(01)00381-6.
97. McDowell, J.C., Roberts, J.W., Jin, D.J., and Gross, C. (1994). Determination of intrinsic transcription termination efficiency by RNA polymerase elongation rate. Science *266*, 822-825. 10.1126/science.7526463.
98. Snyder, M.P., Gingeras, T.R., Moore, J.E., Weng, Z., Gerstein, M.B., Ren, B., Hardison, R.C., Stamatoyannopoulos, J.A., Graveley, B.R., Feingold, E.A., et al. (2020). Perspectives on ENCODE. Nature *583*, 693-698. 10.1038/s41586-020-2449-8.
99. Jensen, T.H., Jacquier, A., and Libri, D. (2013). Dealing with pervasive transcription. Mol Cell *52*, 473-484. 10.1016/j.molcel.2013.10.032.
100. Robinson, R. (2010). Dark matter transcripts: sound and fury, signifying nothing? PLoS Biol *8*, e1000370. 10.1371/journal.pbio.1000370.
101. Bangash, M.A., Park, J.M., Melnikova, T., Wang, D., Jeon, S.K., Lee, D., Syeda, S., Kim, J., Kouser, M., Schwartz, J., et al. (2011). Enhanced polyubiquitination of Shank3 and NMDA receptor in a mouse model of autism. Cell *145*, 758-772. 10.1016/j.cell.2011.03.052.
102. Tardaguila, M., de la Fuente, L., Marti, C., Pereira, C., Pardo-Palacios, F.J., Del Risco, H., Ferrell, M., Mellado, M., Macchietto, M., Verheggen, K., et al. (2018). SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification. Genome Res *28*, 396-411. 10.1101/gr.222976.117.
103. Chen, S., Zhou, Y., Chen, Y., and Gu, J. (2018). fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics *34*, i884-i890. 10.1093/bioinformatics/bty560.
104. Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol *37*, 907-915. 10.1038/s41587-019-0201-4.
105. Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., and Li, H. (2021). Twelve years of SAMtools and BCFtools. GigaScience *10*. 10.1093/gigascience/giab008.
106. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923-930. 10.1093/bioinformatics/btt656.
107. Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., and Kingsford, C. (2017). Salmon provides fast and bias-aware quantification of transcript expression. Nature Methods *14*, 417-419. 10.1038/nmeth.4197.
108. Anders, S., Reyes, A., and Huber, W. (2012). Detecting differential usage of exons from RNA-seq data. Genome Res *22*, 2008-2017. 10.1101/gr.133744.111.
109. Bonfield, J.K., Marshall, J., Danecek, P., Li, H., Ohan, V., Whitwham, A., Keane, T., and Davies, R.M. (2021). HTSlib: C library for reading/writing high-throughput sequencing data. GigaScience *10*. 10.1093/gigascience/giab007.

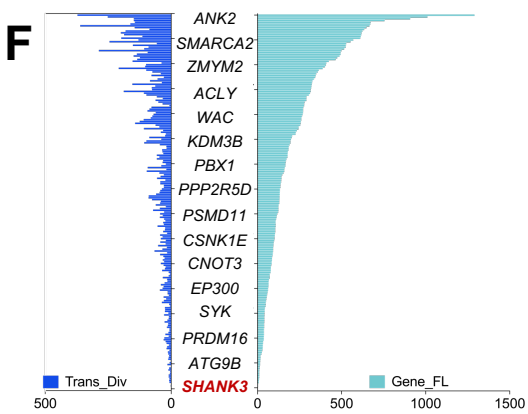
110. Li, H. (2011). Tabix: fast retrieval of sequence features from generic TAB-delimited files. Bioinformatics *27*, 718-719. 10.1093/bioinformatics/btq671.
111. Genomics, x. FFPE Mouse Brain Coronal Section 1 (FFPE), Spatial Gene Expression Dataset by Space Ranger 2.0.0.
112. Zheng, G.X.Y., Terry, J.M., Belgrader, P., Ryvkin, P., Bent, Z.W., Wilson, R., Ziraldo, S.B., Wheeler, T.D., McDermott, G.P., Zhu, J., et al. (2017). Massively parallel digital transcriptional profiling of single cells. Nature Communications *8*, 14049. 10.1038/ncomms14049.

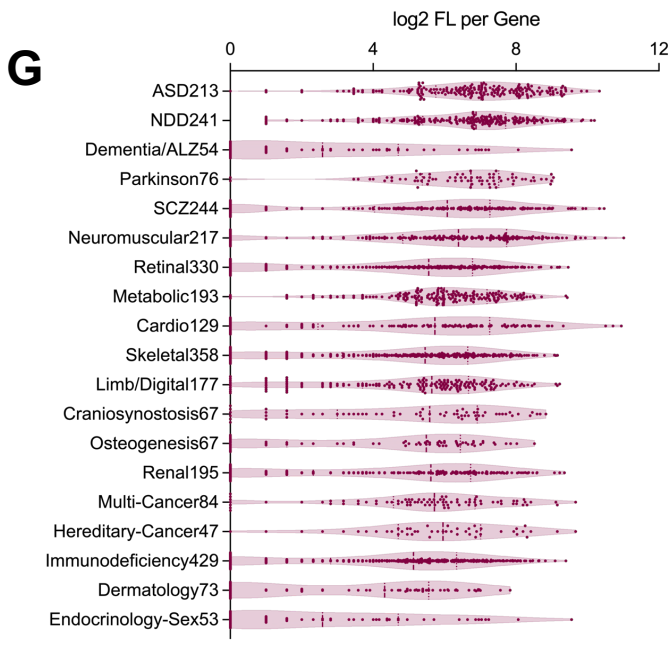
Fig.1











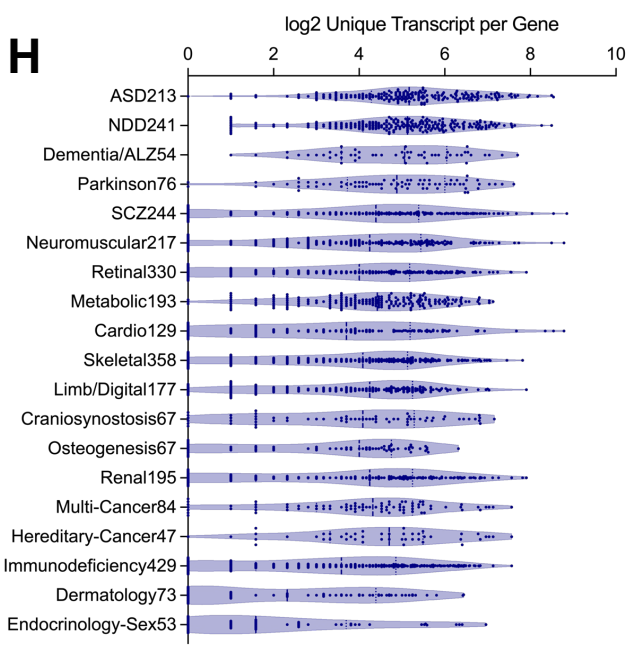
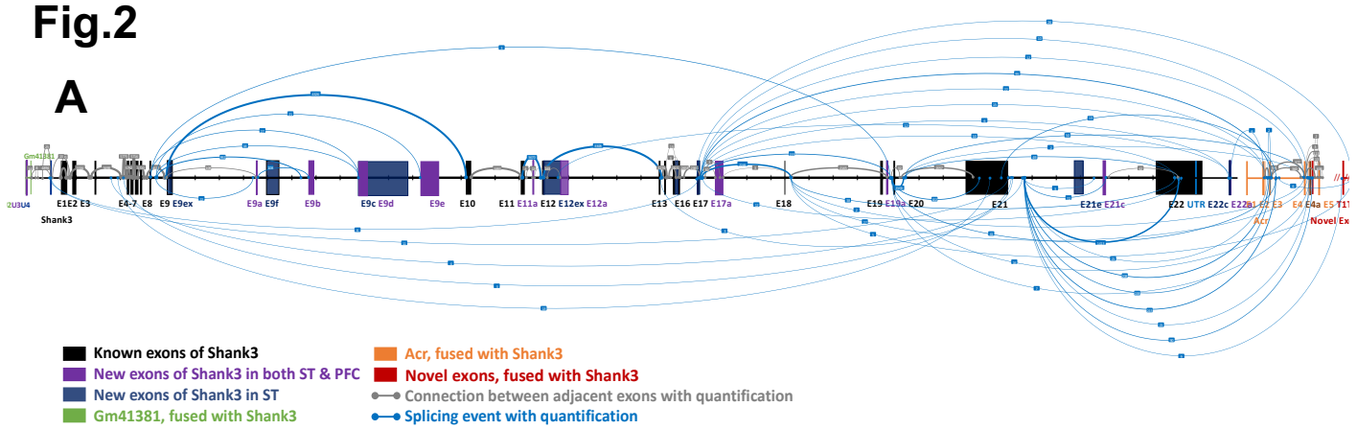
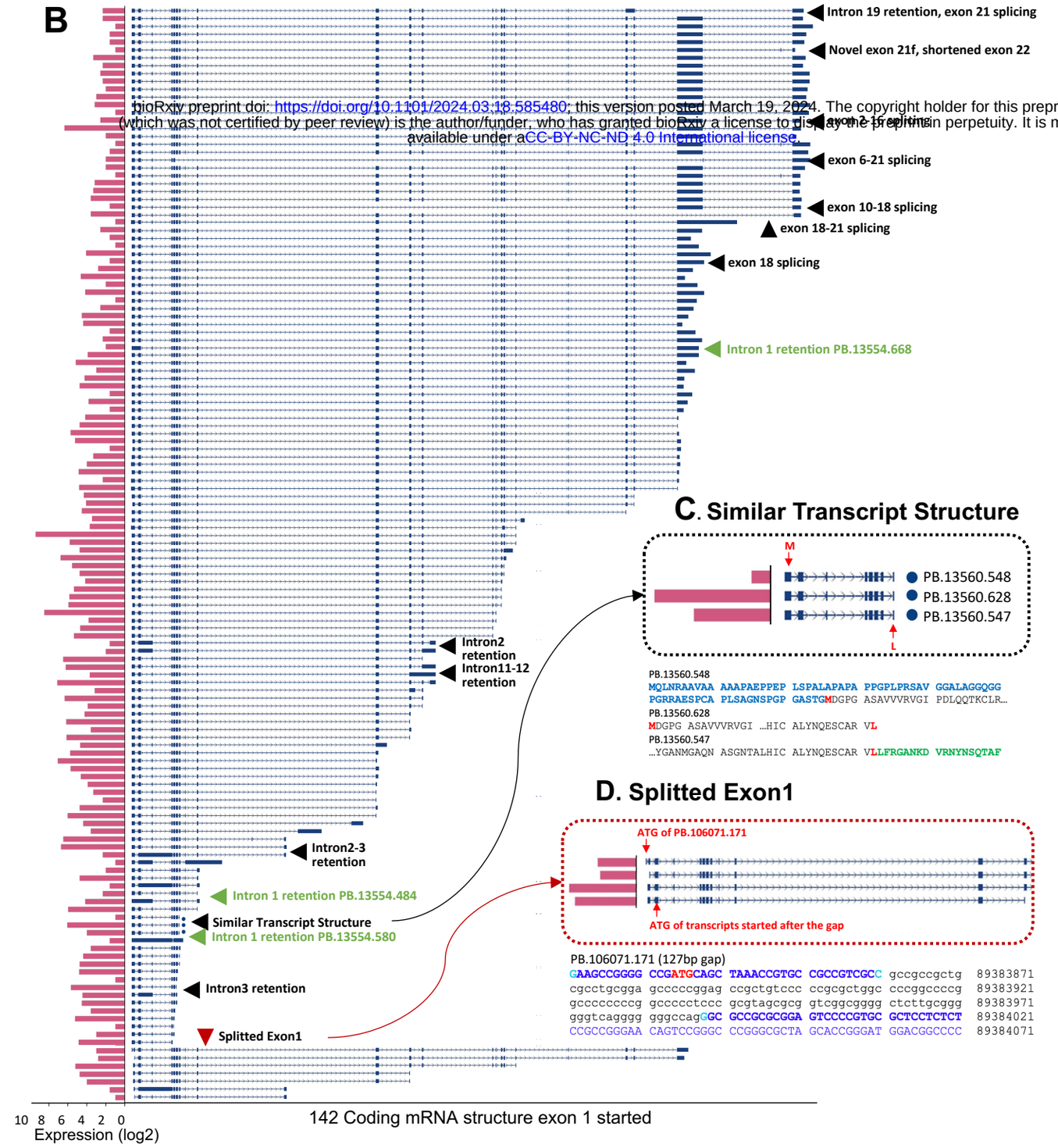
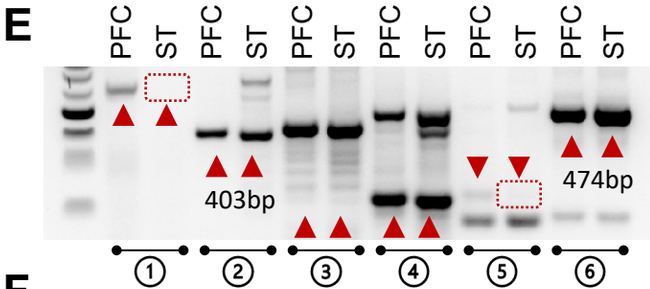
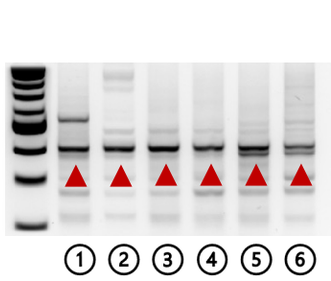


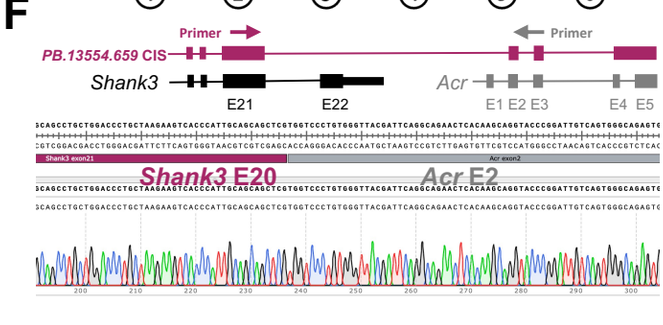
Fig.2











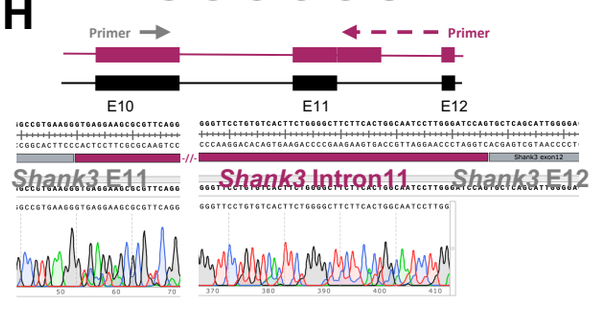
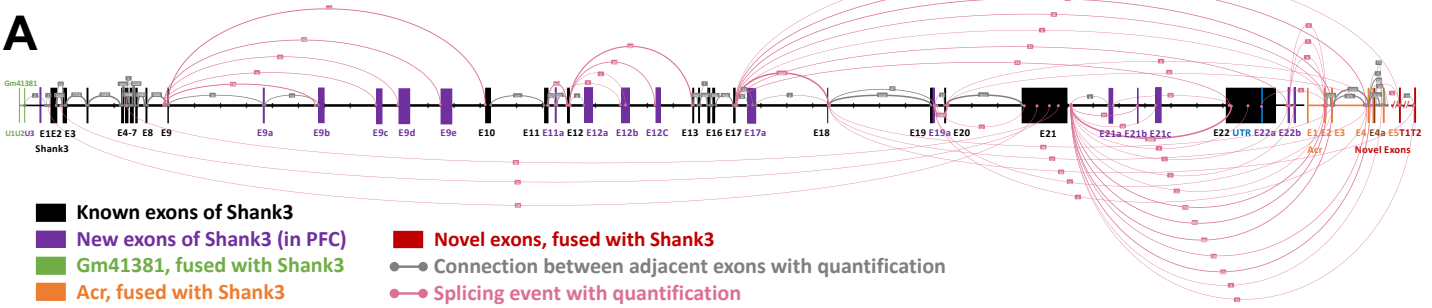
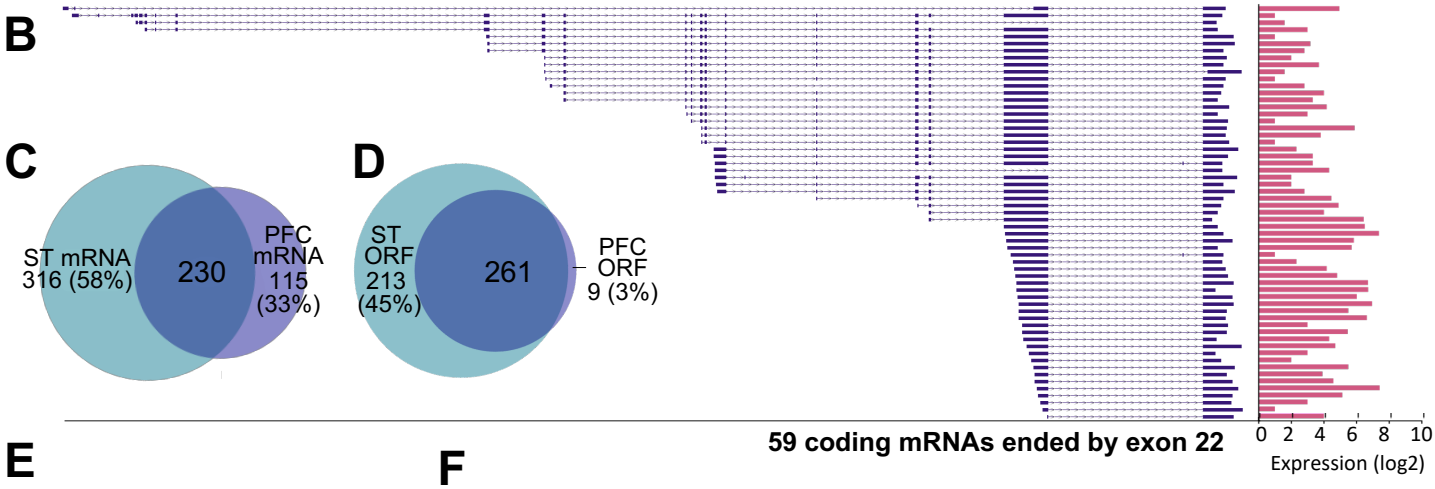
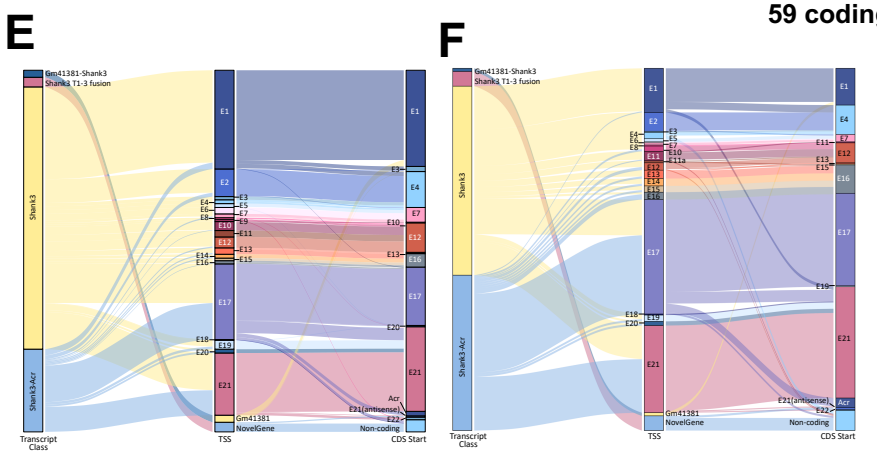
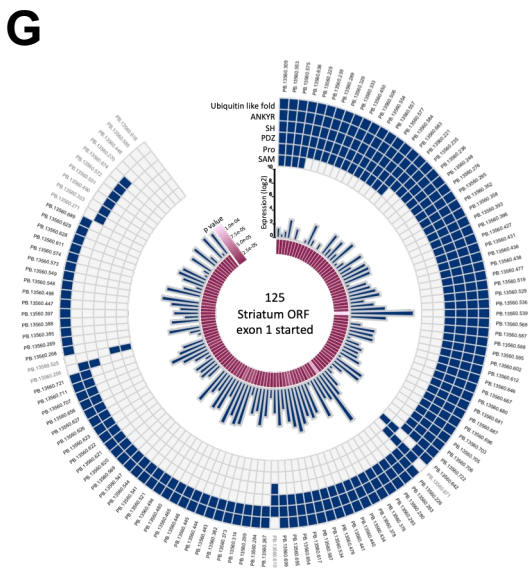


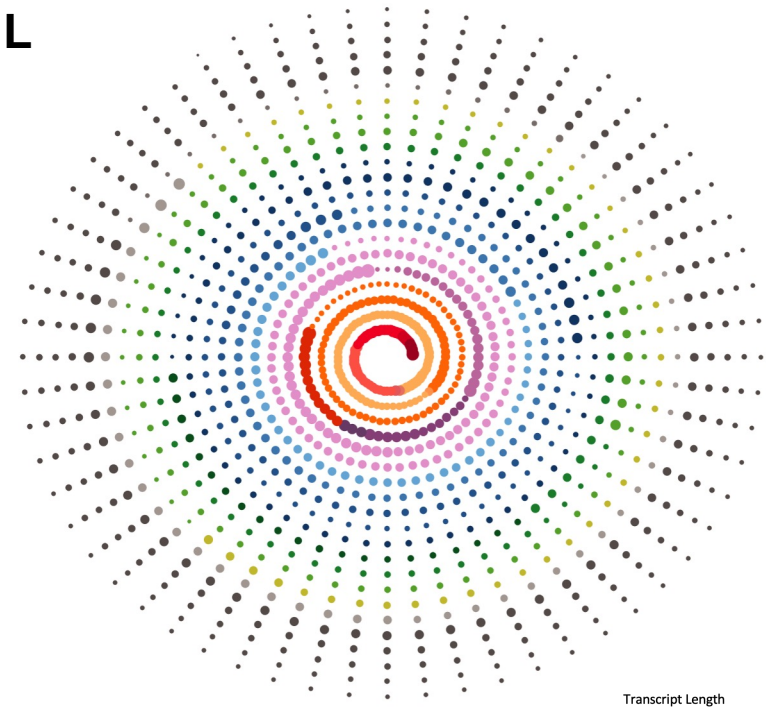
Fig.3

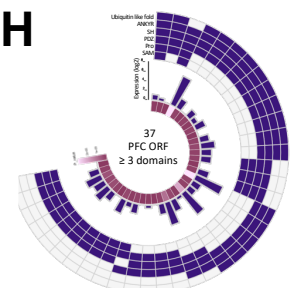


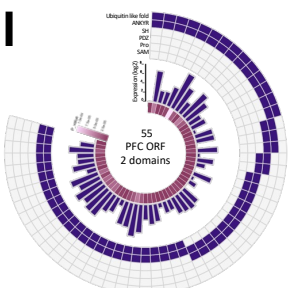


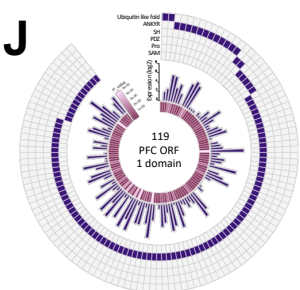












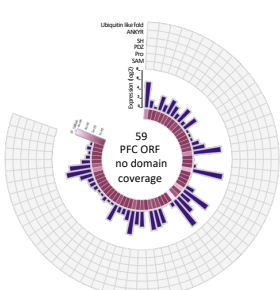


Fig.4

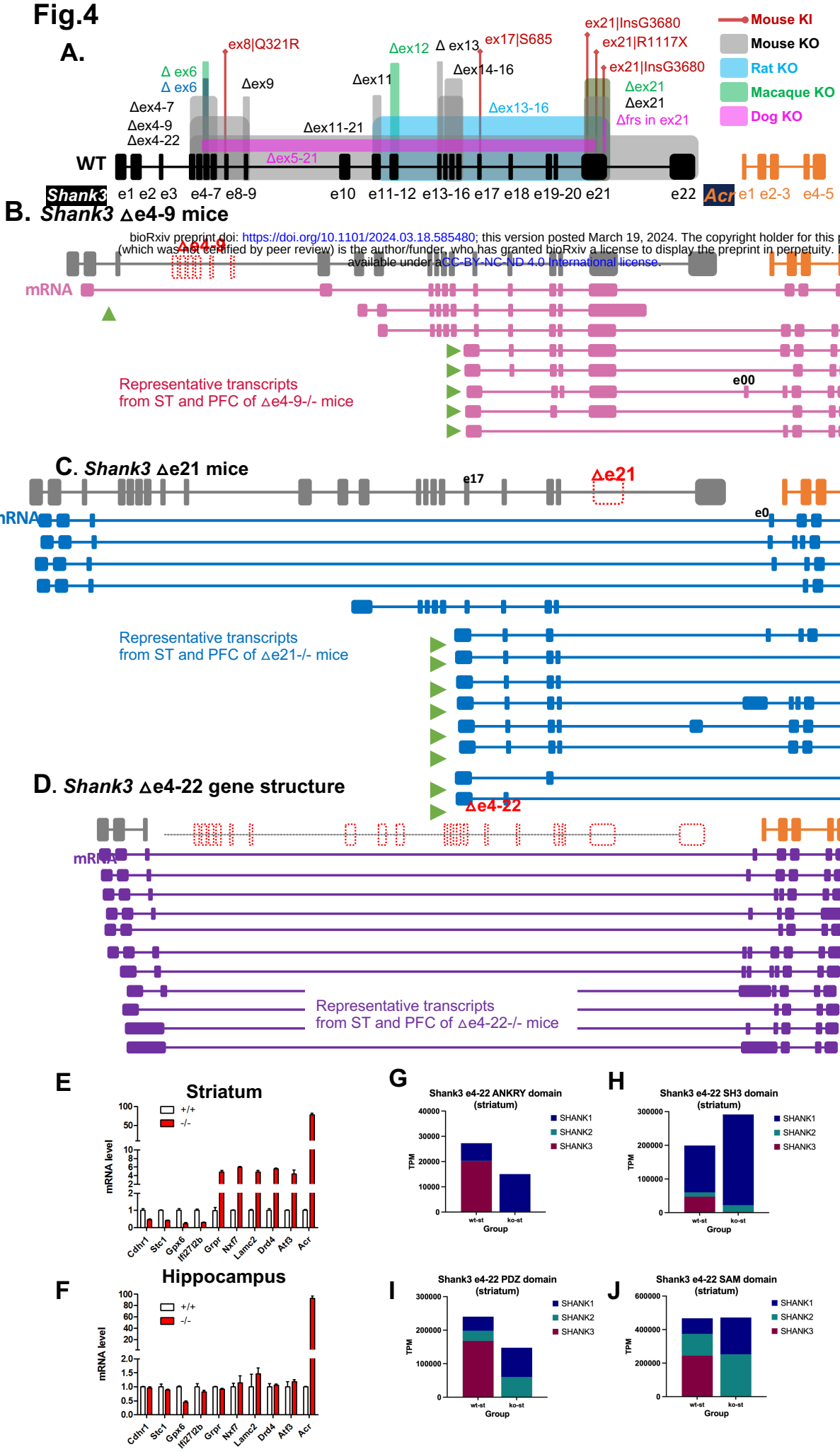
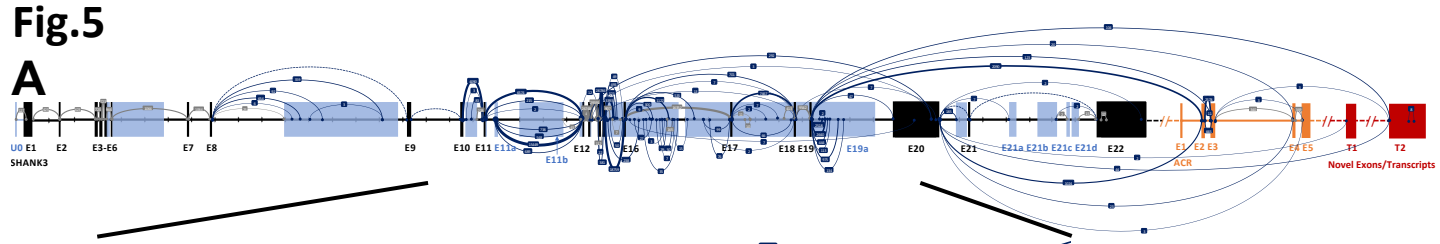
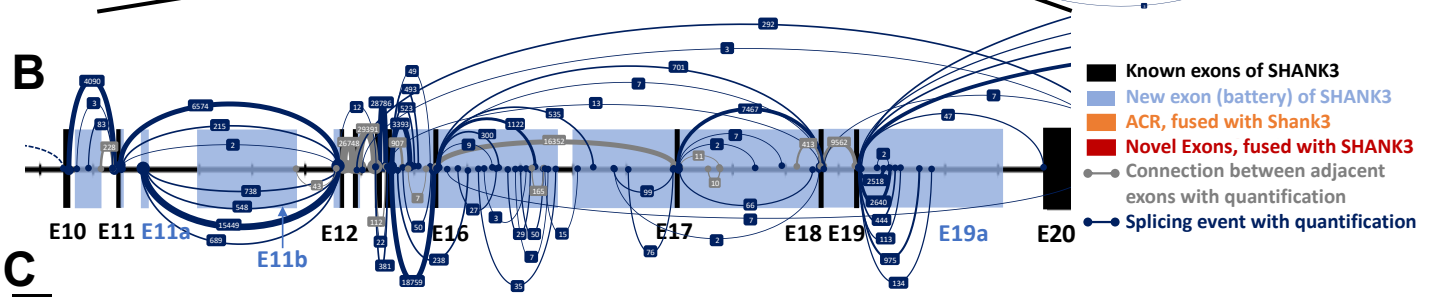
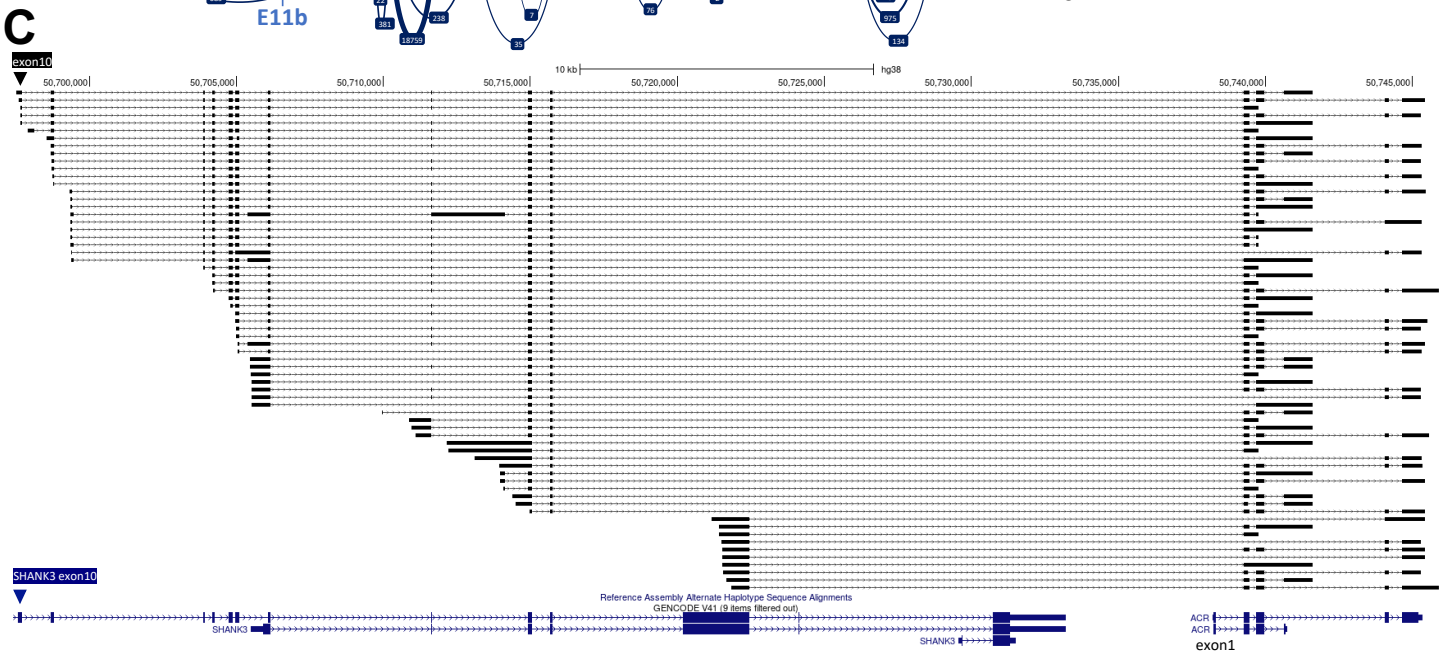
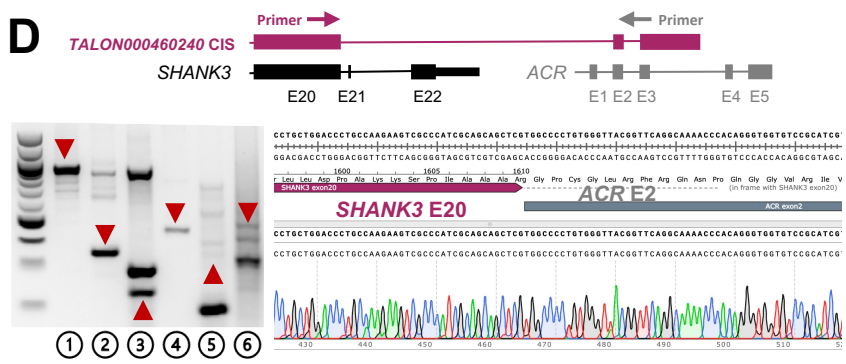


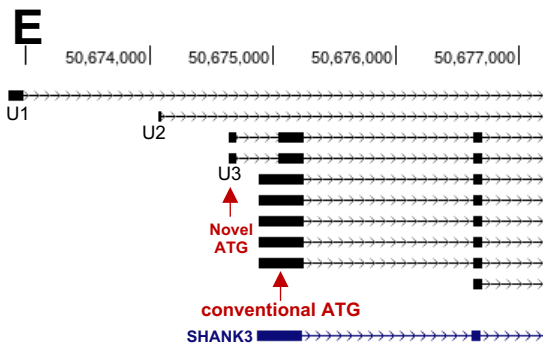
Fig.5

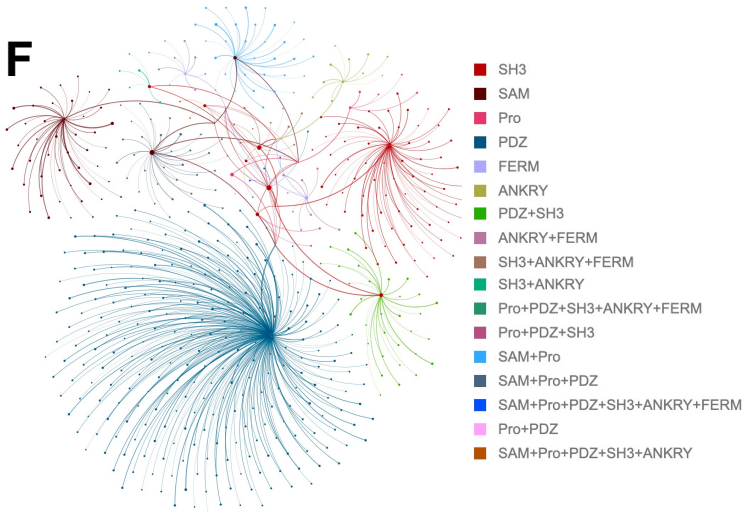

A
U0 E1
SHANK3
E2
E3-E6
E7
E8
E9
E10
E11
E11a
E11b
E12
E16
E17
E18 E19
E19a
E20
E21
E21a E21b E21c E21d
E22
E1 E2 E3
ACR
E4 E5
T1
T2
Novel Exons/Transcripts

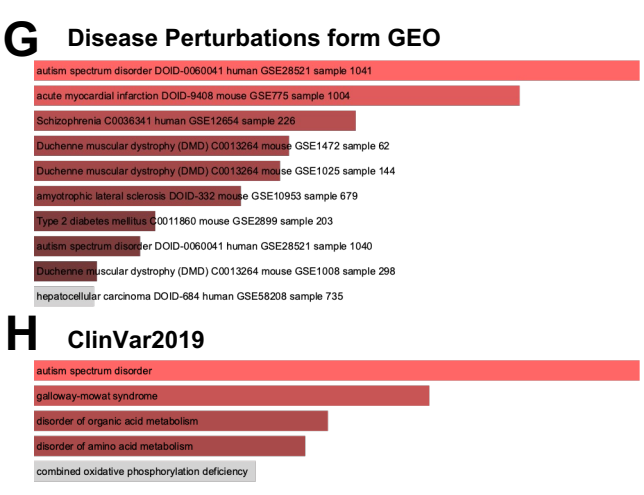

B
E10 E11 E11a
E11b
E12
E16
E17
E18 E19
E19a
E20
Known exons of SHANK3
New exon (battery) of SHANK3
ACR, fused with Shank3
Novel Exons, fused with SHANK3
Connection between adjacent exons with quantification
Splicing event with quantification


C
exon10
10 kb
hg38
50,700,000
50,705,000
50,710,000
50,715,000
50,720,000
50,725,000
50,730,000
50,735,000
50,740,000
50,745,000
SHANK3 exon10
Reference Assembly Alternate Haplotype Sequence Alignments
GENCODE V41 (9 items filtered out)
SHANK3
ACR
exon1

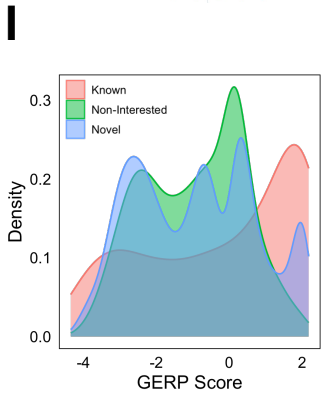

D
Primer
Primer
TALON000460240 CIS
SHANK3
ACR
E20 E21 E22
E1 E2 E3
E4 E5
SHANK3 E20
ACR E2
1 2 3 4 5 6

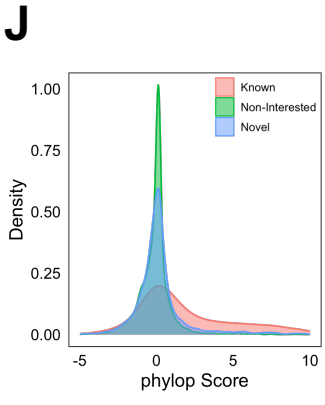

E
50,674,000
50,675,000
50,676,000
50,677,000
U1
U2
U3
Novel ATG
conventional ATG
SHANK3

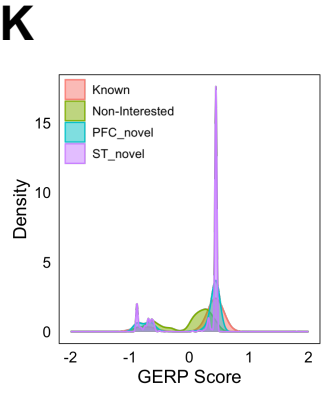

F
SH3
SAM
Pro
PDZ
FERM
ANKRY
PDZ+SH3
ANKRY+FERM
SH3+ANKRY+FERM
SH3+ANKRY
Pro+PDZ+SH3+ANKRY+FERM
Pro+PDZ+SH3
SAM+Pro
SAM+Pro+PDZ
SAM+Pro+PDZ+SH3+ANKRY+FERM
Pro+PDZ
SAM+Pro+PDZ+SH3+ANKRY


G
Disease Perturbations form GEO
autism spectrum disorder DOID-0060041 human GSE28521 sample 1041
acute myocardial infarction DOID-9408 mouse GSE775 sample 1004
Schizophrenia C0036341 human GSE12654 sample 226
Duchenne muscular dystrophy (DMD) C0013264 mouse GSE1472 sample 62
Duchenne muscular dystrophy (DMD) C0013264 mouse GSE1025 sample 144
amyotrophic lateral sclerosis DOID-332 mouse GSE10953 sample 679
Type 2 diabetes mellitus C0011860 mouse GSE2899 sample 203
autism spectrum disorder DOID-0060041 human GSE28521 sample 1040
Duchenne muscular dystrophy (DMD) C0013264 mouse GSE1008 sample 298
hepatocellular carcinoma DOID-684 human GSE58208 sample 735

H
ClinVar2019
autism spectrum disorder
galloway-mowat syndrome
disorder of organic acid metabolism
disorder of amino acid metabolism
combined oxidative phosphorylation deficiency


I
Known
Non-Interested
Novel
Density
GERP Score


J
Known
Non-Interested
Novel
Density
phylop Score


K
Known
Non-Interested
PFC_novel
ST_novel
Density
GERP Score

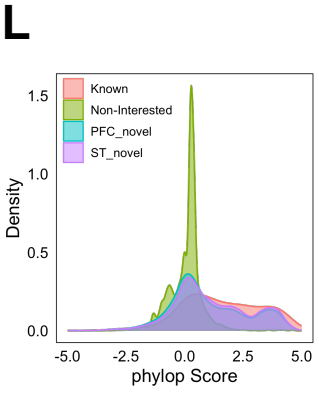

L
Known
Non-Interested
PFC_novel
ST_novel
Density
phylop Score

Fig.6

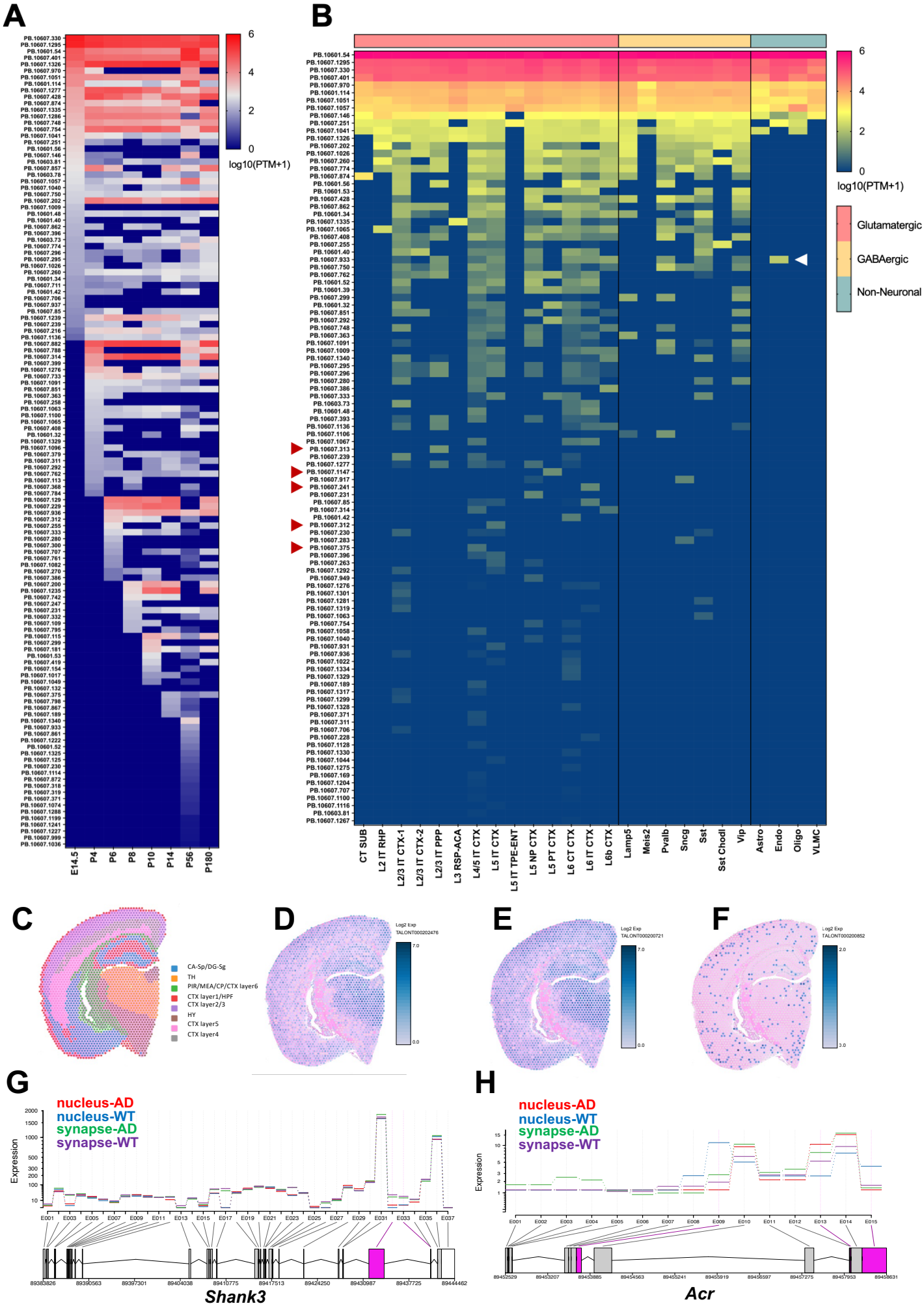
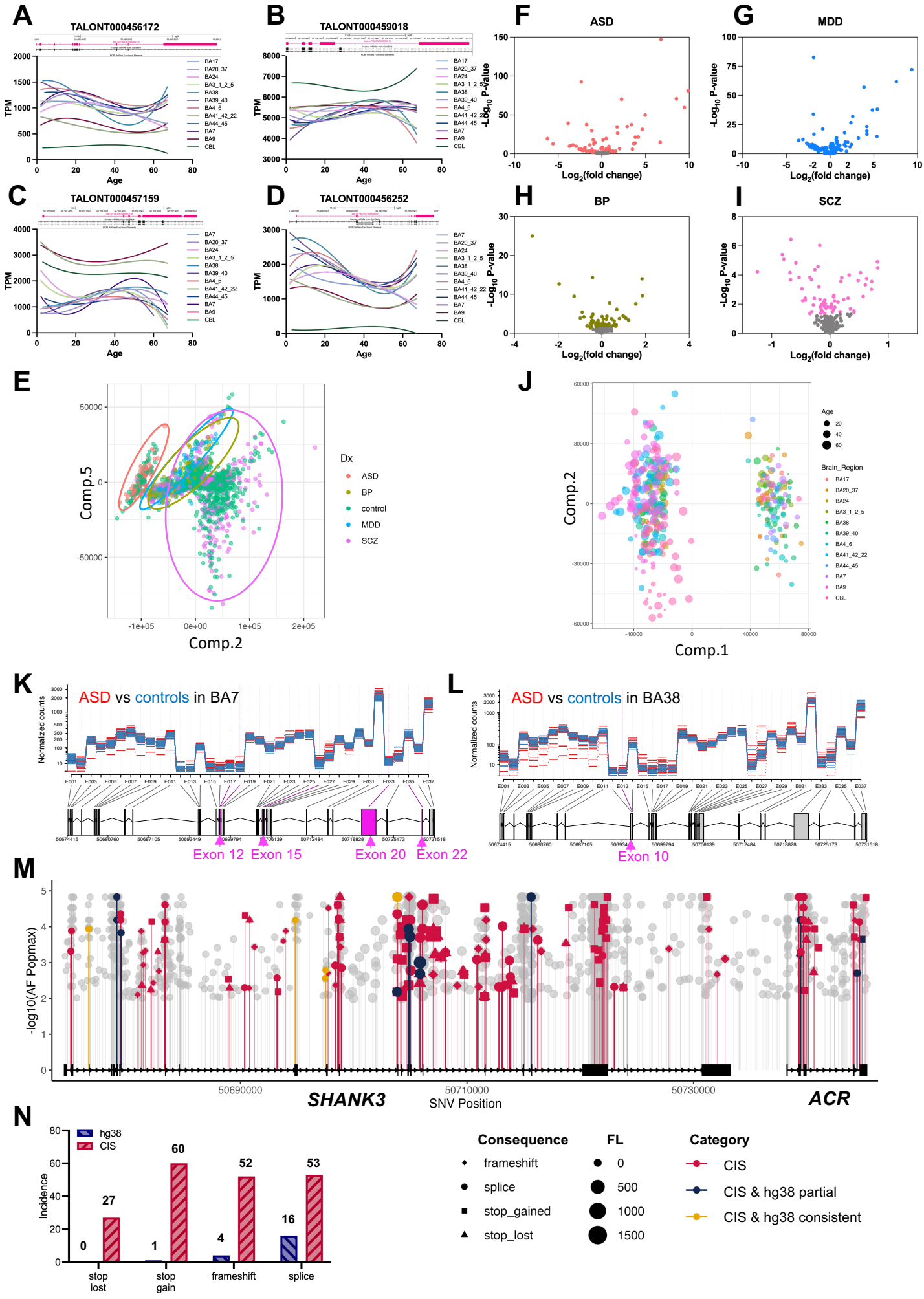

A
B
log10(PTM+1)
E14.5
P4
P6
P8
P10
P14
P56
P180
CT SUB
L2 IT RHP
L2/3 IT CTX-1
L2/3 IT CTX-2
L2/3 IT PPP
L3 RSP-ACA
L4/5 IT CTX
L5 IT CTX
L5 IT TPE-ENT
L5 NP CTX
L5 PT CTX
L6 CT CTX
L6 IT CTX
L6b CTX
Lamp5
Meis2
Pvalb
Sncg
Sst
Sst Chodl
Vip
Astro
Endo
Oligo
VLMC
Glutamatergic
GABAergic
Non-Neuronal
C
CA-Sp/DG-Sg
TH
PIR/MEA/CP/CTX layer6
CTX layer1/HPF
CTX layer2/3
HY
CTX layer5
CTX layer4
D
E
F
Log2 Exp
TALONT000202476
TALONT000200721
TALONT000200852
G
H
nucleus-AD
nucleus-WT
synapse-AD
synapse-WT
Expression
Shank3
Acr

Fig.7


A
TALONT000456172
B
TALONT000459018
C
TALONT000457159
D
TALONT000456252
TPM
Age
BA17
BA20_37
BA24
BA3_1_2_5
BA38
BA39_40
BA4_6
BA41_42_22
BA44_45
BA7
BA9
CBL
F
ASD
G
MDD
H
BP
I
SCZ
-Log10 P-value
Log2(fold change)
E
Comp.5
Comp.2
Dx
ASD
BP
control
MDD
SCZ
J
Comp.2
Comp.1
Age
Brain_Region
K
ASD vs controls in BA7
Normalized counts
Exon 12
Exon 15
Exon 20
Exon 22
L
ASD vs controls in BA38
Exon 10
M
-log10(AF Popmax)
SHANK3
SNV Position
ACR
N
hg38
CIS
Incidence
0
27
1
60
4
52
16
53
stop lost
stop gain
frameshift
splice
Consequence
frameshift
splice
stop_gained
stop_lost
FL
0
500
1000
1500
Category
CIS
CIS & hg38 partial
CIS & hg38 consistent