Long-read sequencing reveals the RNA isoform repertoire of neuropsychiatric risk genes in human brain

Ricardo De Paoli-Iseppi[1*], Shweta Joshi[1], Josie Gleeson[1], Yair David Joseph Prawer[1], Yupei Yu[1], Ria Agarwal[1], Anran Li[1], Anthea Hull[1], Eloise Marie Whitehead[1], Yoonji Seo[1], Rhea Kujawa[1], Raphael Chang[1], Mriga Dutt[1], Catriona McLean[2,3], Benjamin Leo Parker[1], Michael Ben Clark[1*]

[1]Department of Anatomy and Physiology, The University of Melbourne, Parkville, VIC, Australia

[2]Department of Anatomical Pathology, Alfred Health, Melbourne, Victoria, Australia

[3]Victorian Brain Bank, The Florey, Parkville, Victoria, Australia

*Corresponding authors

Abstract

Neuropsychiatric disorders are highly complex conditions and the risk of developing a disorder has been tied to hundreds of genomic variants that alter the expression and/or products (isoforms) made by risk genes. However, how these genes contribute to disease risk and onset through altered expression and RNA splicing is not well understood. Here we show our current understanding of gene isoforms is far from complete and reveal the precise splicing profiles of neuropsychiatric disorder risk genes. Combining our new bioinformatic pipeline IsoLamp with nanopore long-read amplicon sequencing, we deeply profiled the RNA isoform repertoire of 31 high-confidence neuropsychiatric disorder risk genes in human brain. We show most risk genes are more complex than previously reported, identifying 440 novel isoforms and 28 novel exons, including isoforms which alter protein domains, and genes such as *ATG13* and *GATAD2A* where most expression was from previously undiscovered isoforms. The greatest isoform diversity was present in the schizophrenia risk gene *ITIH4*. Mass spectrometry of brain protein isolates confirmed translation of a novel exon skipping event in ITIH4, suggesting a new regulatory mechanism for this gene in brain. Our results emphasize the widespread presence of previously undetected RNA and protein isoforms in brain and provide an effective approach to address this knowledge gap. Uncovering the isoform repertoire of neuropsychiatric risk genes will underpin future analyses of the functional impact these isoforms have on neuropsychiatric disorders, enabling the translation of genomic findings into a pathophysiological understanding of disease.

Introduction

Over 90% of multi-exonic human genes undergo alternative splicing (AS), a process that enables genes to produce multiple mRNA products (RNA isoforms) [1]. Common AS events include exon skipping, intron retention and alternative 5' and 3' exonic splice sites [2]. These mRNA alterations can impact the open reading frame (ORF) and/or alter post-transcription regulation and translation of an RNA, increasing both transcriptomic and proteomic diversity [1, 3, 4]. AS has been established as an important regulator of organ development and physiological functions and is highly regulated under normal conditions [5, 6]. Conversely, aberrant RNA splicing has been linked to the development of cancer, autoimmune and neurodevelopmental disorders [7-11]. AS plays an especially important role in the brain, which has a distinct splicing program including the largest number of tissue specific exons and frequent use of microexons [12]. Numerous studies have reported crucial roles for AS in brain development and dysregulation in disease [13, 14].

Neuropsychiatric or mental health disorders (MHDs) including schizophrenia (SZ), major depressive disorder (MDD), autism spectrum disorder (ASD) and bipolar disorder (BD) can carry significant morbidity for affected individuals [15]. Comorbidities, delayed diagnoses and stigma surrounding MHDs also present a significant challenge to individuals and their families [16]. However, treatment options remain limited or are not well tolerated or effective in some individuals and the underlying aetiology of disease and risk remains poorly understood [17-19]. Recently, large genome wide association studies (GWAS) have revealed hundreds of common single nucleotide polymorphisms (SNPs) that are associated with risk of developing neuropsychiatric disease [20-25]. The vast majority of these variants are found in non-coding parts of the genome and are expected to be regulatory, impacting gene expression levels or which RNA isoforms are produced. For example, risk variants could impact splicing factor binding potentially resulting in novel RNA isoforms, gross transcript alterations or changes to isoform splicing ratios [8]. Confirmatory studies including transcriptome wide association studies (TWAS), summary data–based Mendelian randomization (SMR) [26], multimarker analysis of genomic annotation (MAGMA) and variants (H-MAGMA [27], nMAGMA [28]) and functional genomics have helped to identify the risk genes at these loci and also showed a considerable number of risk loci are shared between disorders [29]. However, there is a current lack of understanding about how risk gene expression and splicing are altered by the risk variants and therefore profiling both their expression and RNA isoforms is essential to link genetic changes to disease pathophysiology.

Current sequencing technologies including Illumina short-reads perform well at detecting novel AS, however the lack of long-range exon connectivity information inherent in short-reads means these approaches are limited in their ability to identify and quantify full-length isoforms and this issue is exacerbated in longer, more complex genes [30, 31]. In contrast, long-read technologies including nanopore sequencing from Oxford Nanopore Technologies (ONT) and Pacific Biosciences (PacBio) single-molecule, real-time sequencing can sequence entire isoforms in a single read enabling more accurate isoform profiling [7, 32]. Such technologies now make it feasible to comprehensively examine gene isoform profiles. Initial investigations of *SNX19* and *CACNA1C* demonstrated the incomplete knowledge of isoform profiles in humans and the likely importance of novel MHD risk gene isoforms in disease [33, 34].

In this study we addressed the lack of knowledge surrounding MHD risk gene isoform expression using nanopore amplicon sequencing. We developed a new bioinformatic tool, IsoLamp, to identify known and novel RNA isoforms from long-read data. Analysis of the RNA splicing profiles of 31 MHD risk genes identified 360 novel RNA isoforms and 28 novel exons. We identified several genes where most expression is from novel isoforms, including *ATG13* and *GATAD2A*, where the most highly expressed isoforms were novel. Our results show the transcript structure for most risk genes is more complex than current annotations, containing additional exon skipping events, retained introns, novel splice sites and novel exons, including novel isoforms that alter the protein and potentially its function. This work lays the foundation for a better understanding of how risk gene isoforms may play a role in disease pathophysiology.

Methods

Sample preparation and QC

Healthy control post-mortem human brain samples were obtained from six individuals collected through the Victorian Brain Bank (VBB) under HREC approval #12457. Brain samples from two further individuals with a non-control phenotype were also obtained from the VBB for protocol testing and validation purposes. Age, sex and additional details including the post-mortem interval (PMI), pH and tissue weight are shown in Supplementary Table 1. Briefly, samples comprised 5 males and 1 female, age range: 51 – 72 yrs, PMI range: 31 – 64 hrs and pH range: 5.7 – 6.7. Frozen

tissue (weight range: 57 – 135 mg) was cut from seven brain regions including Brodmann areas (BA), BA9 (dorsolateral prefrontal cortex (DLPFC)), BA46 (medial prefrontal cortex (MPFC)), BA10 (fronto-parietal cortex (FPC)), Brodmann Area 24 (dorsal anterior cingulate cortex (dACC)), caudate, cerebellum and temporal cortex. Total RNA was extracted from bulk tissue in eight randomised batches of 3 – 6 samples. First, frozen brain tissue was homogenised on ice, using a manual tissue grinder (Potter-Elvehjem, PTFE), whilst immersed in 1 mL QIAzol Lysis Reagent (QIAGEN). Lysate was then processed using a RNeasy Lipid Tissue Kit (QIAGEN, 74804), according to the manufacturers' instructions. Isolated RNA quality and quantity was checked using a Qubit 4 Fluorometer (2 µL), TapeStation 4200 (RNA integrity number equivalent (RINe), cut-off = 6) and Nanodrop 2000.

Database curation and risk gene selection

MHD risk genes were selected for long-read amplicon sequencing using an internal database that aimed to collate evidence from the literature of gene involvement in disease risk from the original GWAS, meta-analyses including MAGMA (and variants including eMAGMA, hMAGMA, nMAGMA), TWAS, SMR and follow-up studies including fine mapping, protein-protein interaction (PPI), epigenetic (DNA methylation) and targeted experimental validation (Supplementary Figure 1). The foundation of this database was a list of significant GWAS SNPs for SZ, BD, MDD and ASD. Association data was downloaded from the NHGRI-EBI GWAS Catalog [35]. MHD GWAS associations were filtered on the 'Disease/Trait' column to exclude effects of treatments including pharmaceutical, mixed disorder studies and associations with behavioural traits like smoking or alcohol intake. Associations were excluded if both the 'reported gene' column was 'not reported (NR)' and the 'mapped gene' column was blank. Date data was downloaded, filters applied, and percentage associations retained are detailed in Supplementary Table 2.

Follow-up studies and experiments were then identified in the literature and the reported genes were manually collated and assigned to an 'evidence' category as described above. Each evidence category had the following information headers including the PubMed ID of the reporting manuscript, the first author and the reported SNP and gene. An R script (Supplementary File 1) was used to curate risk genes and appearances in unique studies. Counts of reported risk genes and unique studies, for each evidence category, were then combined with the original GWAS table. Risk genes were then sorted by evidence (high to low), separately for each MHD. A multi-trait evidence list was also made by combining each MHD table together and again sorting by descending evidence. This gave us flexibility to focus on risk genes that appeared to be specific to a single MHD or those with shared risk across disorders.

Primer design, cDNA synthesis and long-range PCR

Thirty-one (31) MHD risk genes were selected from our database and the full coding sequence (CDS) from the canonical isoform was downloaded from the UCSC Genome Browser [36]. Primers, located in the 5' and 3' UTRs, were designed to amplify the CDS using Primer3 Plus [37]. Additional primers were made to amplify alternative start or end sites that were not captured by a single primer pair. Additional UCSC track sources including expressed sequence tags (EST), transcript support level (TSL), APPRIS designation, human mRNA support, cap-analysis of gene expression (CAGE) peaks, CpG islands and H3K4Me3 marks were examined to ensure there was enough evidence that alternative start or end sites were real before a primer was designed [38]. All primers, primer combinations and modified Primer3 Plus settings are listed in Supplementary Table 3. Risk gene primers from Primer3 Plus were aligned to tracks on the UCSC Genome Browser using BLAT for visualisation and tested using the In-Silico PCR [39].

To amplify risk gene CDS, 1 µg of total RNA was used as template for cDNA synthesis using Maxima H Minus Reverse Transcriptase (Thermo Fisher Scientific, EP0752, 200 U/µL) according to the manufacturers' instructions. Two duplicate cDNA plates were generated simultaneously to reduce

variability and provide enough template for multiple risk gene PCRs. Risk genes were amplified using one of following DNA polymerases; LongAmp® Taq 2X Master Mix (NEB, M0287S), Platinum™ SuperFi II PCR Master Mix (Thermo Fisher Scientific, 12368010) or PrimeSTAR GXL (TakaraBio, R050B). Each set of gene primers were individually optimised by adjusting PCR cycling conditions (Supplementary Table 3) until sufficient pure template (~1 – 10 ng) could be produced for input to barcoding. Short-fragments and primer-dimer were removed prior to barcoding using AMPure XP beads (Beckmann Coulter) at 0.5 – 0.8x ratios to PCR volume. An overview of the experimental protocol is shown in Figure 1A.

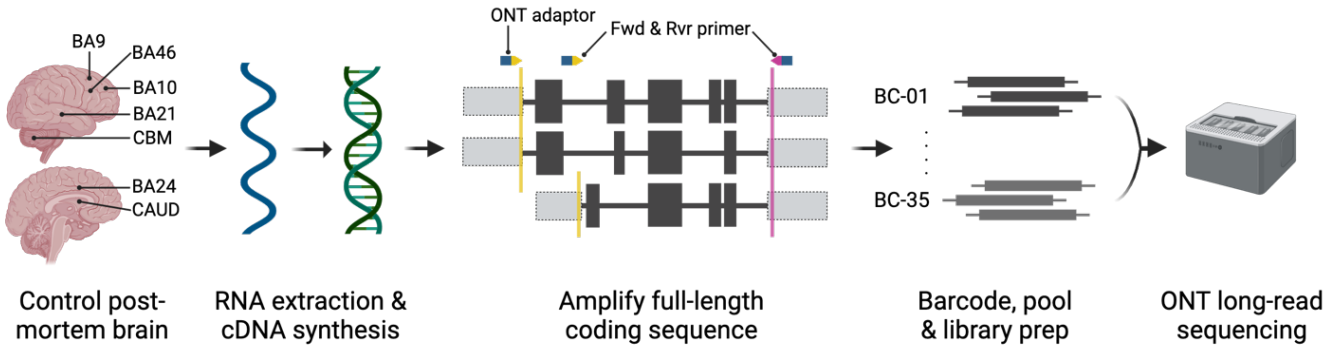
Long-read amplicon sequencing

Barcoding conditions for sample multiplexing (N=35, EXP-PBC096, ONT) and library preparation for long-read sequencing followed the recommended ligation sequencing protocol (Figure 1A) (SQK-LSK109/110, ONT). All barcoding PCR was done using LongAmp® Taq 2X Master Mix with an amplicon specific extension time (approximately 1 min/Kb) and 10 – 15x cycles. AMPure clean-up following adaptor ligation was adjusted from the default ratio of 0.4X depending on the length of the target amplicon. Adaptor ligated libraries were loaded (25 – 35 fmol) onto MinION (FLO-MIN106) flow cells and a minimum of 10,000 reads per sample were targeted before flushing and storing the flow cell. All runs were re-basecalled using the super-accurate (SUP) basecalling model (Guppy v6.0.17, 2022) and minimum qscore = 10.

Isoform discovery from long-read amplicon sequencing with IsoLamp

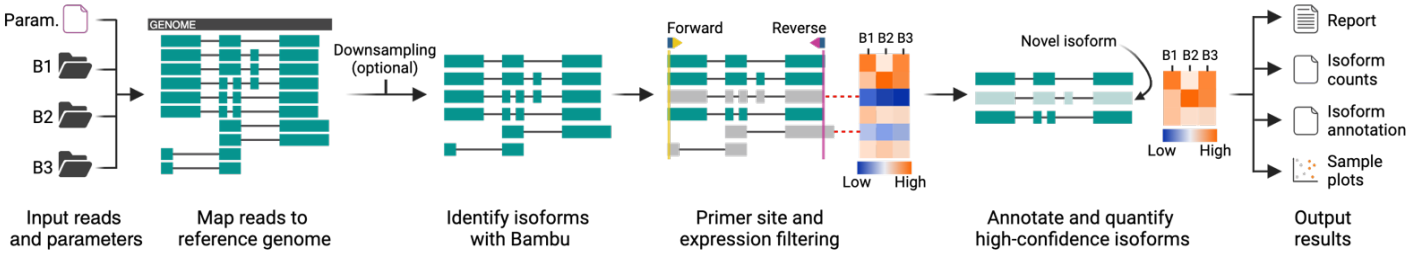
We developed a new bioinformatic pipeline, IsoLamp, for the analysis of long-read amplicon data (Figure 1B). First, pass reads were down sampled [40] to a consistent number (8000) per barcode and mapped to the reference genome with minimap2 (v2.24) [41]. Then, low accuracy reads (< 0.95) were removed and samples were merged prior to isoform identification. Read accuracy was calculated using CIGAR strings in the BAM files and is defined as: (‘X’+’=’+’I’+’D’-‘NM’)/(‘X’+’=’+’I’+’D’). Next, the merged BAM file of high accuracy reads was used as input for isoform discovery with Bambu (v3.2.4) using the following parameters: novel discovery rate (NDR) = 1 and min.fractionByGene = 0.001 [42]. Next, isoforms identified with Bambu were filtered to remove any with zero expression and to retain only isoforms with start and end positions overlapping the known primer coordinates using bedtools intersect (v2.30) [43] to retain only isoforms with start and end positions overlapping the known primer coordinates. Finally, the remaining isoforms were used to create an updated transcriptome with GffRead (v0.12.7) [44], and reads from each barcode were quantified with salmon in alignment-based mode (v0.14.1) [45, 46] and isoforms were annotated with GffCompare (v0.12.6) [44]. The pipeline outputs a list of isoform annotations (.gtf), isoform expression as transcripts per million (TPM) and proportion of overall gene expression, as well as a report summarising the results. An optional TPM filter was also applied to further filter isoforms which required a minimum TPM of 5000. If the user specifies a grouping variable for their input samples, a t-test is performed between isoform proportions between groups and p-values are adjusted for the false discovery rate (0.05).

175 **A**



176

177 **B**



178

179 **Figure 1A. RNA isoform sequencing of human post-mortem brain.** RNA was isolated from frontal cortical
180 regions, caudate (CAUD) and cerebellum (CBM) and converted to cDNA. The coding sequence (black boxes)
181 was amplified using specific forward (Fwd, yellow arrows) and reverse (Rvr, pink arrow) primers generally
182 designed in the 5' and 3' UTR regions (grey boxes) to capture as many isoforms as possible. An Oxford
183 Nanopore Technologies (ONT) adaptor sequence (blue box) was incorporated into each primer for sample
184 multiplexing. Samples were then barcoded and pooled to create a single library for long-read sequencing on a
185 GridION. Key: Brodmann Area (BA), barcode (BC), Oxford Nanopore Technologies (ONT). **B. Isoform discovery**
186 **with long-read amplicon sequencing (IsoLamp) workflow.** A gene specific parameters file (containing
187 chromosome and primer coordinates) was used to align long-reads for each sample (B1-3) against the
188 reference (black box) using Minimap2. Known and novel RNA isoforms were identified using Bambu. Identified
189 isoforms are then filtered (grey isoforms) on correct forward (yellow line) and reverse (pink line) primer
190 position ensuring full-length isoform discovery. Low expression (blue on heatmap) isoforms were also filtered
191 out as indicated by dashed red lines. Filtered known and novel isoforms can then be annotated, quantified
192 using IsoLamp output files and visualised using IsoVis.

193

We benchmarked the performance of the IsoLamp pipeline using Spike-in RNA Variant (SIRV) Set 1 synthetic RNA controls (Lexogen). SIRV isoforms are present in three mixes (E0, E1, E2) that contain isoforms in varying known concentrations. Primers were designed to amplify from the first to the last exon (as described above) of SIRV5 and SIRV6 genes from cDNA generated in triplicate from each SIRV mix (N=27) (Supplementary Figure 2). PCR amplification conditions for SIRV amplicons are shown in Supplementary Table 4. Samples were barcoded and sequenced as described above, and subsequent basecalling and demultiplexing was performed with Guppy (v6.0.17, SUP, 2022). The IsoLamp pipeline was compared against three other isoform discovery tools: StringTie2 [47], FLAIR [48] and Bambu (using both default parameters and the optimised parameters used in IsoLamp). The sensitivity, specificity, accuracy and correlations (expected versus observed counts) of the programs were compared using three SIRV reference annotations: complete (C), incomplete (I) (missing isoforms, to test ability to recover unannotated true positive isoforms) and over (O) (extra isoforms, to test ability to minimise false positive annotated isoforms). Novel isoforms were categorised using SQANTI (v4.2) against the human reference (GENCODE release 41, GRCh38.p13). Finally principal component analyses (PCA) were done on a combined dataset of expression values for each known and novel isoform and its associated metadata including brain region, gene, RINe, pH, individual, PMI and age.

Novel exon validation

Nanopore long-read supported novel exons were validated by RT-PCR. Amplicons were designed from the known 5' flanking exon into the novel exon and from the novel exon into the known 3' flanking exon. An amplicon spanning the known 5' and 3' flanking exons was used as a positive control. Primers were designed using Primer3 [37] and checked using Primer BLAST [49] and are listed in Supplementary Table 5. In some cases, the primer design space was restricted by the novel exon sequence length and/or nucleotide composition. Novel exon amplifications were done using *Taq* 2X MasterMix (NEB, M0270L) and cycling conditions can also be found in Supplementary Table 5. PCR products were visualised via gel electrophoresis using GelGreen® Nucleic Acid Stain (Biotium, 41005) and GeneRuler 100 bp ladder (TFS, SMN0243). PCR products in the expected size range were cleaned up using AMPure XP Reagent (Beckmann Coulter, A63881) at a 1.8X ratio to remove fragments <100 bp and sent for Sanger sequencing (100 – 200 bp, AGRF).

Protein isolation and novel sequence detection using targeted mass spectrophotometry (MS)

Post-mortem frontal cortex and cerebellum brain samples (mean weight = 53.25 mg) were used for protein isolation. Samples were from a healthy 59 yo female and 74 yo male with PMIs of 30 and 22 hrs respectively who had no known neurological or neuropsychiatric conditions. Briefly, samples were lysed in 500 µL of guanidinium-HCl buffer using tip-probe sonication, heated briefly to 95 °C and diluted 1:1 with LC-MS water before 4 mL of ice-cold acetone was added to precipitate protein overnight at -30 °C. Supernatant following a wash (3 mL 80% cold acetone) and incubation (-30 °C, 1 hr) was discarded and protein air-dried (RT, 30 mins). The protein pellet was resuspended in 500 µL 10% TFE in 100 mM HEPES (pH 7.5) and sonicated. Protein concentration was estimated with BCA (1 µL sample + 9 µL 2% SDS). Normalised protein (10 µg/10 µL) was then digested using a combination of LysC/trypsin or GluC for all samples. Finally, digested peptides were analysed using an OrbiTrap Eclipse mass spectrophotometer using known peptide targets informed by long-reads. Protein structure prediction was done using AlphaFold accessed through UCSF ChimeraX (v1.5) [50-52].

Data visualisation

The publicly available web-tool IsoVis (v1.1, https://isomix.org/isovis/) was used to visualise RNA isoforms and associated expression data (Wan et al., 2024, *under review*). Known and novel RNA isoforms are represented as a stack to compare alternative splicing events between different isoforms. Read counts assigned to each RNA isoform for each of the 35 samples were visualised as a heatmap.

Results

Experimental overview

To identify the RNA isoforms expressed from genes of interest we aimed to perform long-read amplicon sequencing, which provides a highly sensitive means for comprehensive isoform discovery and relative quantification (Figure 1A) [33]. We selected seven regions of post-mortem human brain, encompassing both transcriptionally divergent regions as well as those highly implicated in MHDs. Amplicons were designed cover the full coding region of target genes and, where possible, run from the first to the last exon. Multiple set of primers were used for genes with alternative transcriptional initiation and termination exons and/or alternative coding sequence initiation and termination sites.

IsoLamp: a tool for RNA isoform discovery from long-read amplicon sequencing

While there are several long-read isoform discovery and quantification tools, these are not generally optimised for amplicon sequencing of single genes at high depth. Therefore, we created ISOform discovery with Long-read AMPlicon sequencing **(**IsoLamp), a custom pipeline designed for isoform profiling from amplicon sequencing (Figure 1B). In contrast to previous tools [33] IsoLamp can be applied to any gene, provides flexible filtering options and provides a simpler, unified, output of isoforms.

We benchmarked the performance of IsoLamp using synthetic Spike-in RNA variants (SIRVs) that provide a known ground truth for isoform exonic structures and abundances. We performed long-read amplicon sequencing on SIRV5 and SIRV6, targeting five isoforms per gene, as these SIRVs allowed targeting of the largest number of isoforms with a single primer pair and so best recapitulated human genes. The SIRV dataset comprised nine replicates from each of the three SIRV mixes (E0, E1, E2) for each gene. 99% of reads mapped to the SIRV genome with minimap2 [41], confirming on-target amplification. We benchmarked the performance of IsoLamp with Bambu [42], FLAIR [48], FLAMES [53] and Stringtie2 (-L) [47]. We assessed the precision, recall and expression correlations of the five tools using three different reference annotations (Figure 2A-C, Supplementary Figure 3, Supplementary Table 6): **1.** Complete - contains all SIRV isoforms, **2.** Insufficient - missing SIRV isoforms known to be present, and **3.** Over – contains additional isoforms that are not present in the SIRV mixes.

Our benchmarking results demonstrated IsoLamp had the highest precision and recall values, consistently outperforming other isoform discovery tools by correctly identifying true isoforms and minimising false positives (Figure 2A, Supplementary Table 6). This included maintaining high performance with the more challenging, but also more realistic, insufficient and over annotation references. IsoLamp expression quantification was also consistently accurate and maintained performance irrespective of the annotation provided (Figure 2B, C).

Bambu, which is also utilised within the IsoLamp pipeline, was the next best performing tool although it identified more false positives and had poorer recall and quantification results using the insufficient annotation (Figure 2A-C). IsoLamp utilises Bambu parameters optimised for amplicon-sequencing, including a novel discovery rate (NDR) of 1. Adjusting the Bambu 'NDR' to 1 improved its recall but didn't improve precision (Figure 2A-C, Supplementary Table 6). These results demonstrate how IsoLamp outperforms tools designed for whole-transcriptome analysis, including when Bambu is provided with optimised isoform discovery parameters for amplicon sequencing.

FLAIR had the highest number of isoforms of all tools tested identifying 261, 181, and 278 novel transcripts in the complete, insufficient, and over-annotated reference-based analyses, respectively. This high level of false-positive novel transcripts led to inaccuracies in transcript abundance assignments, resulting in low correlations compared to other tools (Figure 2A-C). FLAMES exhibited 100% recall for SIRV5 across all annotations, however, its performance with SIRV6 was suboptimal, indicating a higher degree of variability in the FLAMES isoform discovery pipeline. FLAMES also performed

poorly for isoform quantification. Lastly, while Stringtie2 did not introduce large numbers of false positives, it had the highest number of false negatives, including when provided with complete annotations (Figure 2A-C, Supplementary Table 6).

IsoLamp employs an optimised expression-based filter to remove lowly expressed isoforms that are likely to be false positive detections. Applying this filter to Bambu, FLAIR, and FLAMES substantially reduced false positive novel isoforms and enhanced overall precision (Supplementary Figure 3, Supplementary Table 6), though IsoLamp was still the top performing tool. Beyond synthetic benchmarking data, reference annotations are typically a combination of insufficient and over annotations. In such scenarios, IsoLamp demonstrated better or comparable correlations with all other tools (Figure 1A-C, Supplementary Figure 3), suggesting its superiority for amplicon-sequencing based isoform discovery and quantification from real biological data.

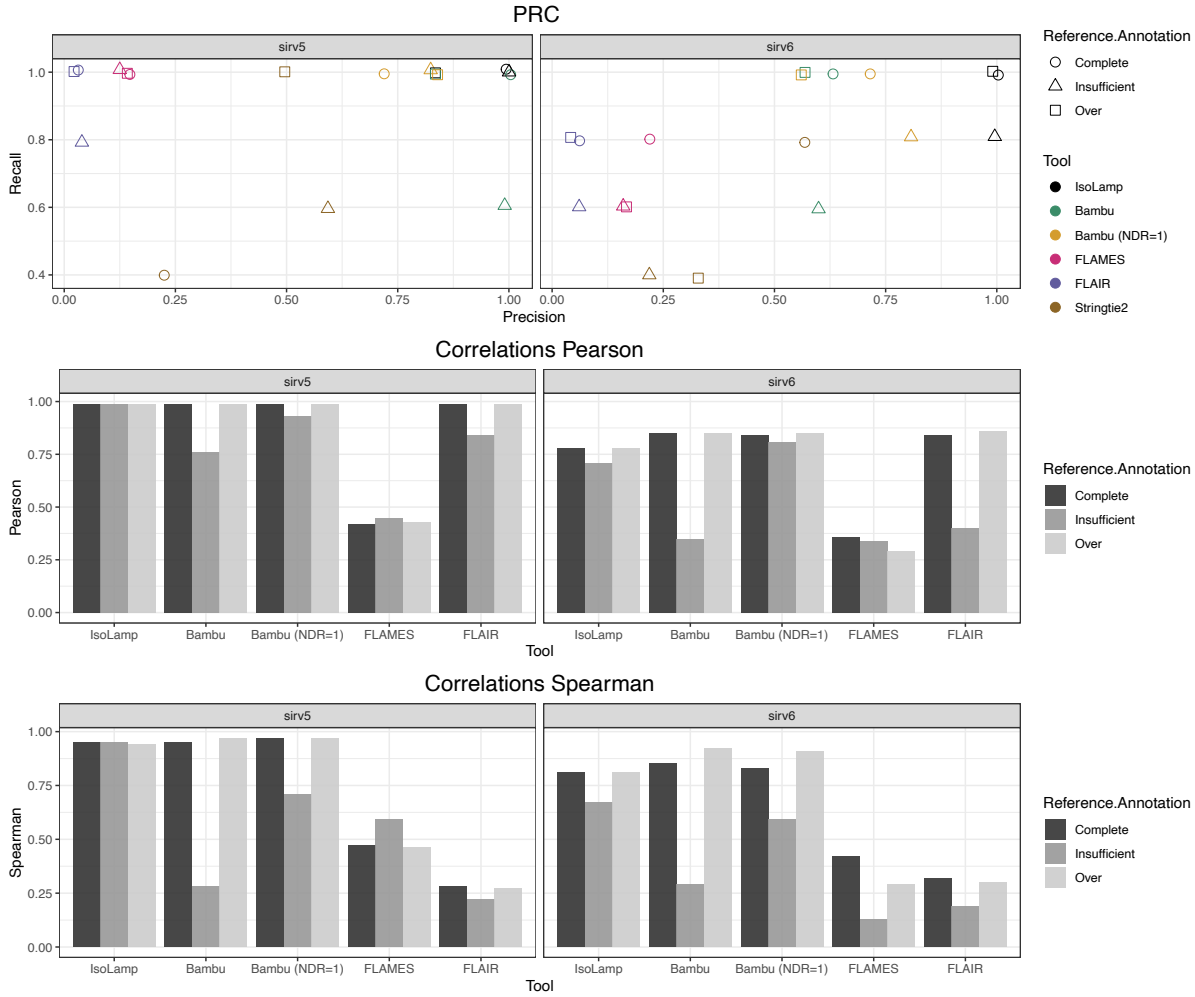


Figure 2. Benchmarking IsoLamp using spike-in SIRVs. A. Precision recall of each tested pipeline with the complete, insufficient or over annotated SIRV reference. IsoLamp (black) returned high quality isoforms from amplicon data of both SIRV5 and 6. Pearson **(B)** and Spearman **(C)** correlations for each pipeline between known and observed expression values for SIRV 5 and 6 mixes.

Post-mortem human brain RNA quantity and quality

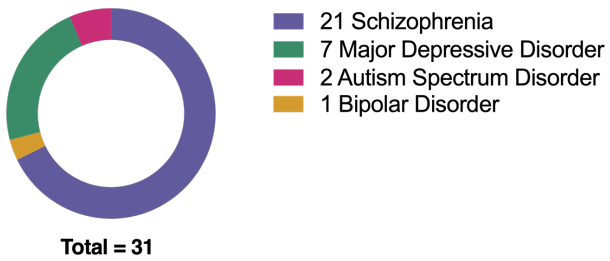
Total RNA for long-range amplicon sequencing was extracted from 7 brain regions from 5 healthy individuals (Ind01 - 05) and subject to sample QC (Supplementary Figure 4A-D). RINe (mean = 7.4, range = 6 - 8.1) did not differ by brain region, however Ind04 had significantly lower RINe scores (Supplementary Figure 4B). No trend between the PMI (mean = 44.25 hrs) and RINe was observed (Supplementary Figure 4C). RINe appeared to worsen with decreasing brain tissue pH levels (Supplementary Figure 4D). A principal component analysis (PCA) showed separation of Ind04

(likely driven by lower sample pH and RINe) and of cerebellum and caudate samples from cortical regions in PC1 and PC2 (Supplementary Figure 5AB). A relatively small proportion of variance (4.7%) was attributed to control donor age in PC5 (Supplementary Figure 5C).

Long read sequencing identifies 360 novel RNA isoforms

A total of 31 risk genes were selected for amplicon sequencing based on the accumulated evidence for their involvement in neuropsychiatric disorder risk. A custom database of risk genes and their evidence levels was created and genes ranked (Methods). In a reflection of current GWAS cohort sizes, 21 of the selected genes had the highest evidence for involvement in risk for SZ, 7 for MDD, 2 for ASD and 1 for BPD (Figure 3A). Evidence from GWAS, TWAS and other studies show that some genes appear to be risk factors for multiple disorders including *KLC1* for SZ, MDD and ASD (Figure 3B).

A



B

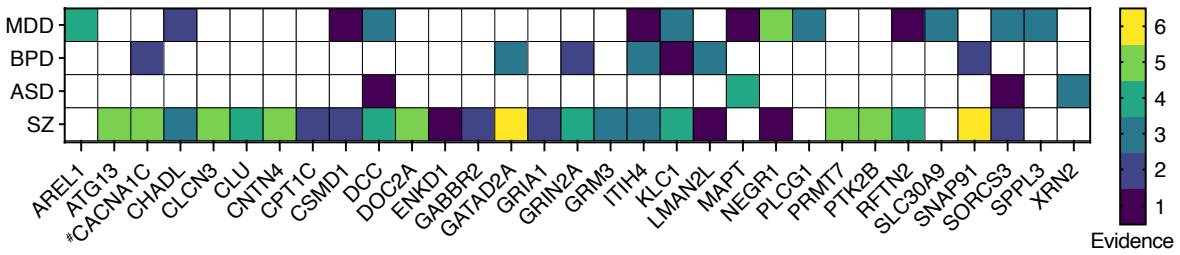
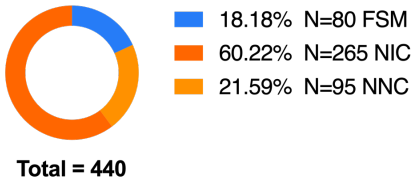


Figure 3. Selection of high-confidence MHD risk genes for amplicon sequencing. A. Risk genes included in this study classified by the disorder for which they have the highest evidence of association. **B.** Sequenced genes and their evidence levels for each MHD. The evidence count was calculated as the sum of independent analysis types for example, GWAS, MAGMA, TWAS, SMR, DNA methylation, fine mapping, protein-protein interaction and targeted validation studies, that supported gene involvement in risk for a particular disorder. #Indicates re-sequencing of a gene from a previous study (Clark *et al.* 2019).

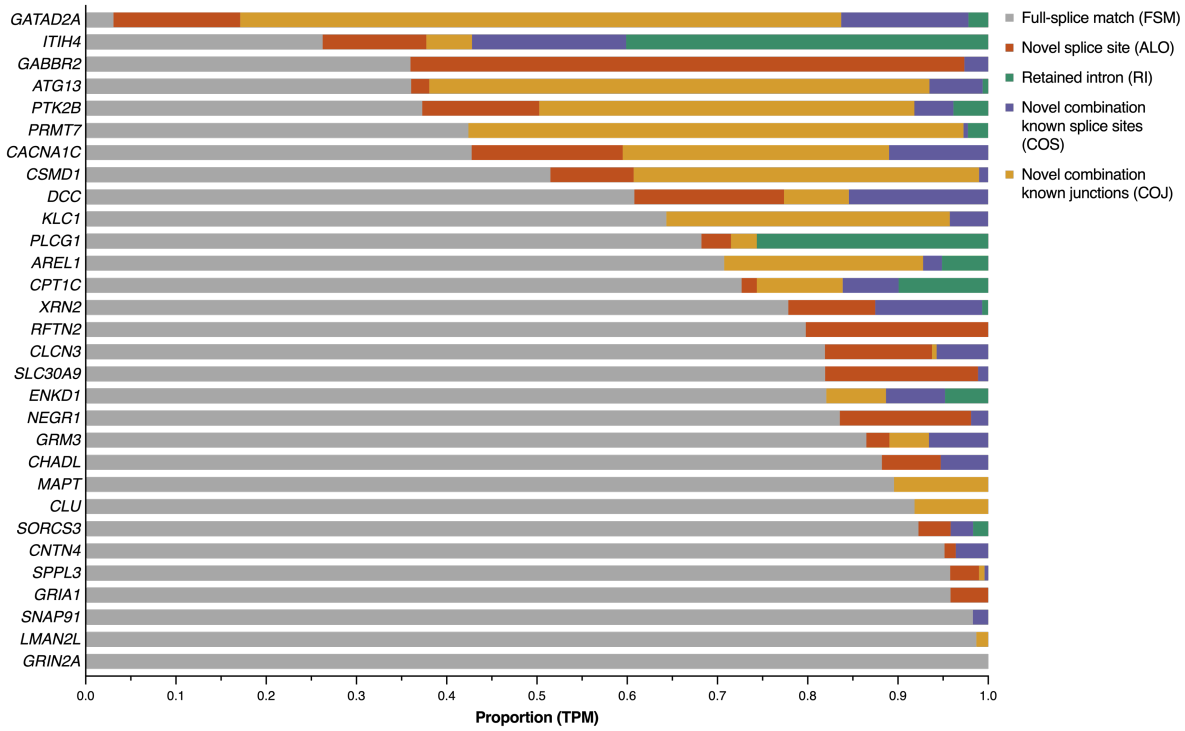
The full RNA isoform profile for each gene was sequenced using nanopore long-read amplicon sequencing. Mapping accuracy ranged from 93.7% (*CLCN3*) to 97.5% (*SORCS3*) (Supplementary Figure 6A, B). Each novel isoform and its predicted impact on known protein domains, open reading frame (ORF) and associated instability index was recorded (Supplementary File 2) and visualised using IsoVis (Supplementary File 3) (Wan et al., 2024, *under review*). With no TPM filter set in IsoLamp we identified 872 known and novel isoforms across all 31 neuropsychiatric disorder risk genes. To filter this list for more highly expressed novel isoforms we applied a TPM filter which resulted in 440 known and novel isoforms across all genes (Figure 4A). Of these, SQANTI [54] classified 80 as known (full splice match (FSM)), 266 as novel but using only known splice sites or junctions (novel in catalogue (NIC)), and 95 as containing at least one novel splice site (novel not in catalogue (NNC)) (Figure 4A).

We next asked what proportion of reads for each gene were assigned to novel isoforms (Figure 4B). This ranged widely from approximately 96.9% for *GATAD2A* to 0% for *GRIN2A*, which was the only gene for which no novel RNA isoforms were detected. Approximately one quarter (7/31) of genes investigated had most of their gene expression assigned to novel isoforms, demonstrating how isoforms and their expression profiles for many genes are still poorly understood. As our amplicon-sequencing does not encompass all variations in transcriptional initiation and termination sites, these results can be seen as a lower bound for novel isoforms and their expression proportions. Linear regression of gene isoform counts (Supplementary Figure 7) and novel isoform proportion did not reveal a significant relationship with amplicon length or canonical exon count, indicating that detection of novel isoforms is largely gene dependent (Supplementary Figure 8A-D). To determine what was different about the splicing pattern of each novel isoform we further sub-classified them using SQANTI, based on the use of a combination of known exon junctions (COJ) or splice sites (COS), retained intron (RI) or containing at least one novel splice site (ALO) (donor, acceptor or pair) [54]. Overall, the most reads were assigned to “novel combination of known junctions”, where all individual exon combination were known but the entire chain of exons was novel. The type and proportion of novel isoforms from each category was highly gene specific, demonstrating a wide variety of novel RNA types missing from current gene annotations.

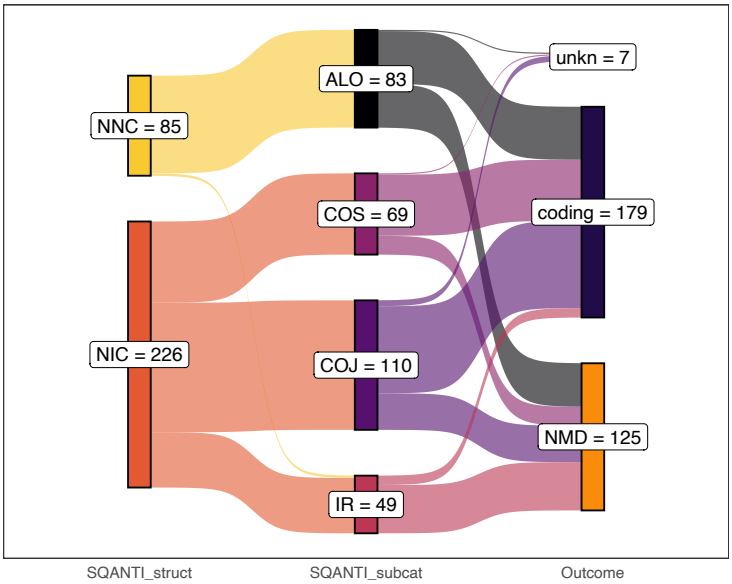
A


18.18% N=80 FSM
60.22% N=265 NIC
21.59% N=95 NNC
Total = 440

B


GATAD2A
ITIH4
GABBR2
ATG13
PTK2B
PRMT7
CACNA1C
CSMD1
DCC
KLC1
PLCG1
AREL1
CPT1C
XRN2
RFTN2
CLCN3
SLC30A9
ENKD1
NEGR1
GRM3
CHADL
MAPT
CLU
SORCS3
CNTN4
SPPL3
GRIA1
SNAP91
LMAN2L
GRIN2A
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Proportion (TPM)
Full-splice match (FSM)
Novel splice site (ALO)
Retained intron (RI)
Novel combination known splice sites (COS)
Novel combination known junctions (COJ)

C


NNC = 85
NIC = 226
ALO = 83
COS = 69
COJ = 110
IR = 49
unkn = 7
coding = 179
NMD = 125
SQANTI_struct
SQANTI_subcat
Outcome

372 **Figure 4. A.** The total number of known and novel isoforms identified across all 31 risk genes. SQANTI
373 structural categories are known/full splice match (FSM), novel in catalogue (NIC) and novel not in catalogue
374 (NNC). **B.** Proportion of reads for each gene as classified by the SQANTI sub-category. **C.** Count of predicted
375 outcomes for novel isoform subcategories. Expasy (Gasteiger *et al*., 2003) was used to examine the open
376 reading frame (ORF) of novel isoforms (SQANTI structural category: novel not in catalogue (NNC) or novel in
377 catalogue (NIC)) using the canonical start and stop as a reference. Predictions were categorised as coding if the
378 ORF was retained, nonsense mediated decay (NMD) if a premature termination codon was present and not
379 within 50 nt of the final exon junction or unknown (unkn) if there was not enough information. Novel isoform
380 SQANTI subcategories (subcat) are, at least one novel splice site (ALO), intron retention (IR) and combination
381 of known junctions (COJ) or splice sites (COS).

382

383 The impact of each novel isoform on the encoded ORF was examined using Expasy [55] and recorded
384 as retaining the canonical or other known reading frame (coding), likely-NMD or unknown. Novel
385 isoforms were classified as coding for 54.2%, 67.3% and 75.4% for ALO, COJ and COS
386 subcategories respectively. We identified 49 novel isoforms that contained retained introns, 39 (83%)
387 of which were predicted to lead to NMD (Figure 4C).

388 Isoforms that contained 'at least one novel splice site' (ALO) generally contained a novel deletion
389 within a known exon or had novel donors and/or acceptors. All novel junctions in ALO isoforms were
390 canonical GT-AG, GC-AG or AT-AC junction pairs, though often only the splice donor (GT) or
391 acceptor (AG) was novel (Supplementary Figure 9A). We found that ~47% of ALO isoforms
392 contained either a single novel splice donor or acceptor. Novel GC-AG pairing was detected in two
393 SZ risk genes, within the 5'UTR of *GABBR2* and the donor site of a novel exon in *RFTN2*. These
394 results show a clear advantage of using long-read sequencing to contextualise novel splice sites which
395 aids in predicting the outcome on the isoform and ORF.

396 **Detection of highly expressed novel isoforms**

397 A key question regarding novel isoforms is whether they are expressed at high enough level to impact
398 the biological function of a gene. This is a complex question, because a novel isoform could be low at
399 the tissue level but highly abundant in a specific cell type, or multiple expressed novel isoforms can
400 be significant cumulatively, especially if they all encode the same change to a protein. Therefore, we
401 focused on genes with significant individual or cumulative expression of novel isoforms (analysis on
402 all gene isoforms is available in Supplementary File 2).

403 We identified 22 novel isoforms for the schizophrenia risk gene autophagy-related protein 13
404 (*ATG13*). Novel isoforms represented 64% of gene expression, compared to 36% for full-splice
405 matches. The most abundant class of novel isoforms (15/22) were COJ, which made up 55.4% of gene
406 expression (Figure 4B). *ATG13* had two alternative splicing hotspots, firstly with the 5'UTR and
407 secondly around a predicted disordered region involving exons 12 and 13 in the canonical isoform.
408 Across all brain regions the most highly expressed isoform was the novel COJ transcript 26 (Tx26),
409 which represented 23% of total reads, surpassing the canonical transcript ENST00000683050
410 (12.8%). Tx26 differs from the canonical transcript by skipping of exon 12. It contains the same CDS
411 as ENST00000359513 but includes an additional exon in the 5'UTR (exon 3). Novel COJ transcripts
412 6 and 8 also had high read counts and together accounted for 16.6% of expression. These isoforms
413 were novel due to a combination of 5'UTR exons not previously seen within full-length GENCODE
414 annotations (Figure 5A, B).

415 The schizophrenia risk gene CUB and sushi multiple domains 1 (*CSMD1*) was the longest CDS we
416 amplified at approximately 10,838 nt encompassing 70 coding exons. In total 8/9 detected isoforms
417 were classified as novel. Following the canonical isoform (ENST00000635120, 51.5% of reads),
418 novel transcripts 26 (COJ) and 33 (ALO) accounted for 38.3% and 7.1% of assigned reads,

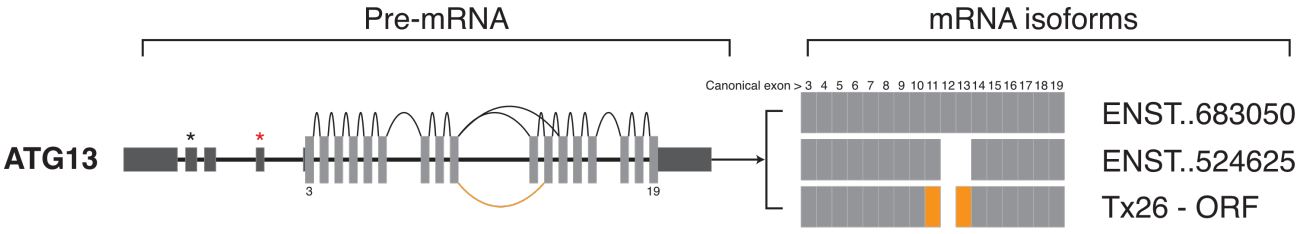
respectively (Figure 5C, D). Novel Tx26 skipped known exon 65 which encodes a sushi 28 extracellular domain and glycosylation site. The ORF of Tx26 retained the reading frame encoding a 3549 amino acid (aa) protein. Novel Tx33 contained a novel splice donor (GT, -8 nt) in canonical exon 21 predicted to lead to a PTC in canonical exon 22. The full Tx33 mRNA also skipped canonical exon 65. *CSMD1* also provides a useful example of the benefit of long read for profiling isoforms. GTEx isoform expression data (https://www.gtexportal.org/home/gene/CSMD1) for *CSMD1* in brain is almost exclusively assigned to isoforms with downstream transcriptional initiation sites (including the two-exon fragment ENST00000521646), despite splice junction level expression largely supporting expression from the canonical start site. The emphasises the difficulty of assembling and quantification expression of full-length isoforms from long, complex genes, which can be achieved using long isoform spanning reads.

The chromatin remodelling subunit and shared SZ and BPD risk gene GATA zinc finger domain containing 2A (*GATAD2A*) had one of the highest proportions of reads (96.9%) assigned to novel isoforms. Most novel isoforms were predicted to be coding COJs (10/24) and these also accounted for the majority of novel read assignment (66.6%). Novel Tx17 had the highest expression level (22.7%) of any *GATAD2A* isoform and skipped canonical exon 10 which overlaps a CpG island (212 nt, 21.7% CpG) and contains a disordered, polar residue biased region and a phosphorylation site (Figure 5E). Two additional novel isoforms (Txs 8 and 12), together accounting for 19.9% of expression, incorporated a known 89 nt 5'UTR exon (ENST00000494516) into full-length isoforms for the first time, clarifying the isoforms expressed from this gene.

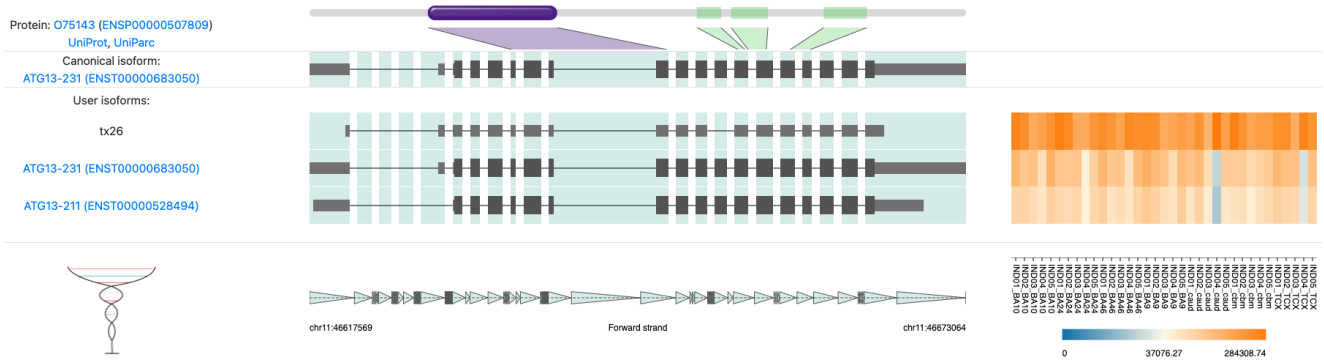
Our results were also useful for several genes in identifying the probable isoforms represented by GENCODE transcript fragments. For the SZ risk gene clusterin (*CLU*), the novel Tx1 (COJ) extended ENST00000520796 to the canonical stop codon and suggested this isoform is moderately abundant (8.2% of TPM) across all brain tissues. The ASD risk gene microtubule associated protein tau (*MAPT*) novel Tx5 extended ENST00000703977 and further demonstrated that inclusion of canonical exon 7 (chr17:45,989,878-45,990,075) does not always exhibit coordinated splicing with canonical exon 5. This isoform had a moderate read count comprising 3.2% of *MAPT* expression.

The full RNA isoform profile for the SZ risk gene calcium voltage-gated channel subunit alpha1 C (*CACNA1C*) has previously been reported and was repeated in this study [33]. In total, we identified 5 annotated and 22 novel isoforms. The most highly expressed novel isoform identified in our previous study, novel 2199, now known as ENST00000682835.1 was also identified in our samples and exhibited similar cerebellar specific expression. Importantly, 10 novel isoforms replicated one of two alternative splicing events in a hotspot identified previously in canonical exon 7 [33]. This hotspot contains the canonical splice site and two alternative 3'SS acceptors over only 12 nucleotides (chr12: 2,493,190-2,493,201) (Supplementary Figure 10).

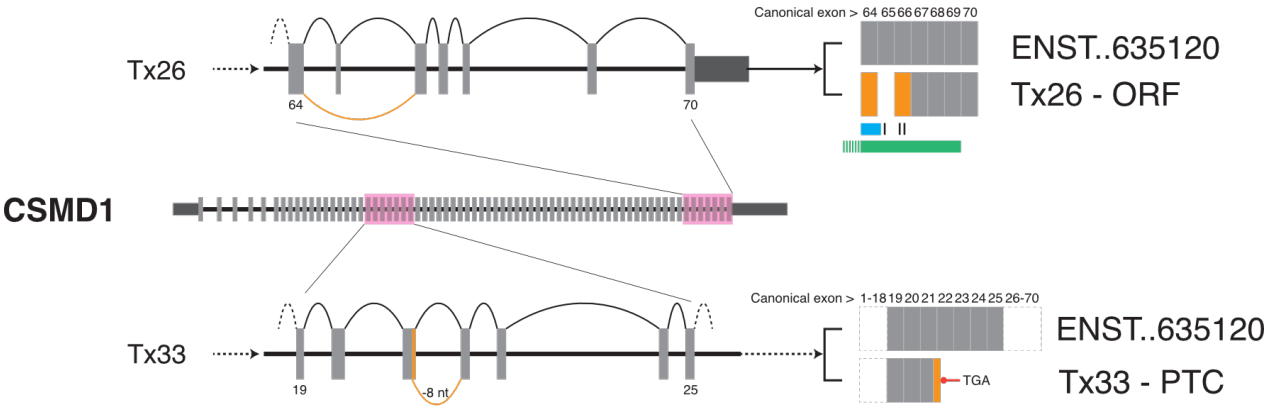
A


Pre-mRNA
mRNA isoforms
ATG13
Canonical exon > 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
3
19
ENST..683050
ENST..524625
Tx26 - ORF

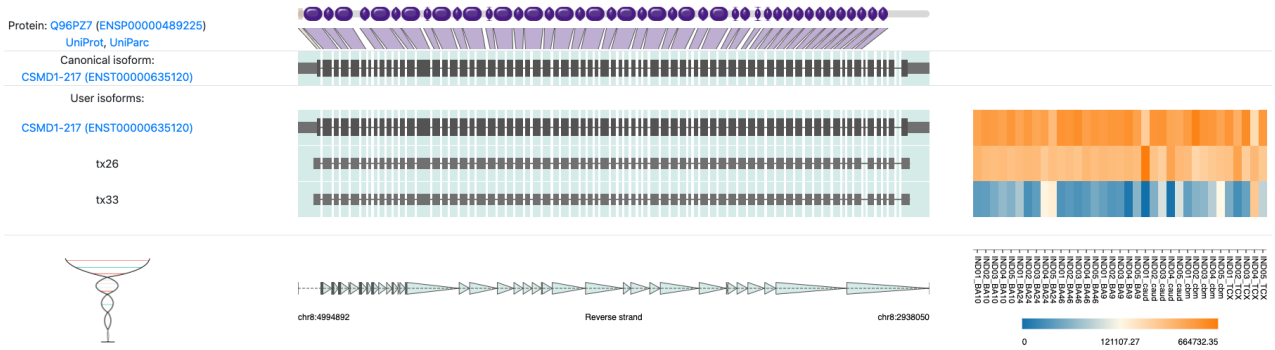
B


Protein: O75143 (ENSP00000507809)
UniProt, UniParc
Canonical isoform:
ATG13-231 (ENST00000683050)
User isoforms:
tx26
ATG13-231 (ENST00000683050)
ATG13-211 (ENST00000528494)
chr11:46617569
Forward strand
chr11:46673064
0
37076.27
284308.74

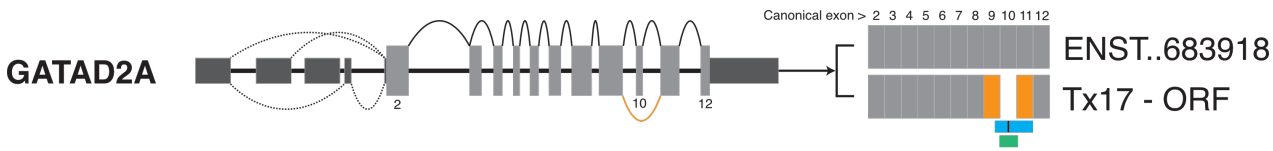
C


Tx26
Canonical exon > 64 65 66 67 68 69 70
64
70
ENST..635120
Tx26 - ORF
CSMD1
Tx33
Canonical exon > 1-18 19 20 21 22 23 24 25 26-70
19
-8 nt
25
TGA
ENST..635120
Tx33 - PTC

D


Protein: Q96PZ7 (ENSP00000489225)
UniProt, UniParc
Canonical isoform:
CSMD1-217 (ENST00000635120)
User isoforms:
CSMD1-217 (ENST00000635120)
tx26
tx33
chr8:4994892
Reverse strand
chr8:2938050
0
121107.27
664732.35

468 **E**



469
470 **Figure 5. Highly abundant novel isoforms and the predicted mRNA outcome. (A,C,E)** mRNA splice graphs.
471 Dark and light grey boxes indicate 5' and 3' UTR and coding exons respectively. Numbers indicate the coding
472 exon. Orange lines (pre-mRNA) and boxes (mRNA) indicate novel splicing events. mRNA isoforms depict known
473 isoforms (ENST) against novel (Tx) isoforms, ".." indicates abbreviated zeroes. **(B,D)** IsoVis visualisation of
474 Isoform structures (centre stack) and expression levels (heatmap). Canonical isoform shown at top of stack
475 including exonic mapping of protein domains (purple) and disordered regions (green) **A.** Splice graph of *ATG13*
476 highlights the open reading frame (ORF) preserving skipping event of canonical exon 12. **B.** High expression of
477 *ATG13* novel transcript 26 (Tx26). **C.** Splice graphs of *CSMD1* novel transcript 26 (Tx26) and 33 (Tx33) changes
478 located within highlighted pink regions for clarity. The ORF retaining skipping event of canonical exon 65 may
479 disrupt a known glycosylation site (black bar), a sushi domain extending from exon 64 (blue) and part of an
480 extracellular domain (green). Tx33 contains a novel splice donor (-8 nt) within exon 21 leading to a premature
481 termination codon (PTC) in exon 22. Dashed lines indicate continuation of the transcript to 5' or 3' coding
482 exons. **D.** Relatively high expression of *CSMD1* novel transcripts 26 and 33. **E.** *GATAD2A* novel transcript Tx17
483 contained a novel, ORF retaining, skipping event of canonical exon 10 which contains a phosphorylation site
484 (black bar), part of a polar biased region (blue) and overlaps a CpG island (<300 bp, green). Dashed lines
485 indicate alternative splicing of 5'UTR exons.

486

487 **Novel isoforms alter predicted protein structures**

488 Novel isoforms have the potential to affect post-transcriptional regulation, protein sequence, structure
489 and function or both. We next investigated a selection of isoforms that would be predicted to lead to
490 protein changes to understand their possible impact.

491 Several novel isoforms (including 5 of the top 20 by expression, Supplementary File 3R) predicted a
492 novel exon 22 skipping event in the SZ risk gene *ITIH4*. Targeted mass spectrophotometry confirmed
493 a novel junction between exons 21 and 23 (ETLFSVMPG//PVLPGGALGISSSIR) created due to
494 skipping of exon 22 (Figure 6A). This event was predicted to encode a PTC <50 nt from the final
495 exon junction, indicating it may not be directed to NMD. Protein structure prediction of the canonical
496 (ENST00000266041.9) and a representative novel isoform (Tx71) indicated a loss of 106 aa (~44%)
497 of the 35 kDa heavy chain domain but retention of three O-glycosylation sites (Thr:719, 720, 722)
498 (Figure 6B-D). Novel transcript 71 accounted for ~3.7% of *ITIH4* TPM and this skipping event was
499 found in an additional 24/68 (35.2%) novel isoforms which together accounted for 23.4% of reads.
500 Tx71 also skipped canonical exons 15 and 16, which contain a protease susceptibility region (residues
501 633 - 713) and a MASP-1 cleavage site (645 - 646: RR) [56]. Cleavage at this site and subsequent
502 formation of an ITIH4-MASP complex can inhibit complement activation via the lectin pathway [56].
503 However, skipping of canonical exon 22 was not mutually inclusive with skipping of exons 15/16 as
504 other novel isoforms with exon 22 skipping retained exons 15/16. The absence of much of the 35 kDa
505 heavy chain domain is likely to impact on ITIH4 protein function and further studies will be required
506 to examine if it plays a role in neuronal phenotypes.

507 Ten novel isoforms were identified for the SZ risk gene glutamate metabotropic receptor 3 (*GRM3*).
508 Three novel isoforms (Txs 6, 7, 9) skip exon 2 which contains the canonical translation start site and
509 instead could use an alternative, frame retaining, translation initiation site in exon 1, extending the
510 truncated reference isoform ENST00000454217.1 which is also supported by human amygdala
511 mRNA (AK294178) (Supplementary File 3Q). Translation of these isoforms would likely cause

512 significant disruption to the resultant protein with removal of the signal peptide, transmembrane
513 domain and disulphide bonds. Cumulatively these novel isoforms accounted for a relatively low 8.8%
514 of expression when compared to the canonical isoform (86.5%).

515 Both novel isoforms and exons were identified for the shared MDD and ASD risk gene neuronal
516 growth regulator 1 (*NEGR1*). Most reads (83.5%) were assigned to the canonical *NEGR1* isoform
517 (ENST00000357731, 354 aa). Three novel isoforms were identified, two of which (Txs 1 and 2)
518 contained novel exons (Supplementary Figure 11A, B). These transcripts accounted for 9.4% (Tx2)
519 and 5% (Tx1) of read counts. Both novel exons were located between cassette exons 6 - 7 and were
520 validated using Sanger sequencing. The novel exon within Tx1 was 42 nt (14 aa) in length, had high
521 100 vertebrate conservation (UCSC) and was predicted to be frame retaining (Supplementary Figure
522 11C). Protein structure prediction of the 368 aa Tx1 using AlphaFold [50] showed a 14 aa extension
523 near the C-terminal prior between the GPI anchor (G:324 aa) and the three immunoglobulin-like
524 domains (Supplementary Figure 11D). In contrast, the 58 nt novel exon within Tx2 encoded a PTC
525 (TAG) only 35 nt distant to the final exon junction complex, suggesting it might not trigger NMD.
526 Truncation of the protein at this position (313 + 7 novel aa) would remove the GPI anchor potentially
527 creating a near complete protein (320 aa) that is unable to attach to the cell membrane (Supplementary
528 Figure 11E).

529

530 **Brain region specific expression of novel isoforms**

531 Many isoforms have brain region enriched or specific expression [57, 58]. Our amplicon sequencing
532 approach identifies the presence and relative expression proportion of different isoforms. We next
533 asked if any novel risk genes isoforms showed expression differences between brain regions. Overall,
534 cerebellum exhibited most differences in isoform expression, consistent with previous whole
535 transcriptome results [12].

536 Depression risk gene DCC netrin 1 receptor (*DCC*) novel isoform Tx9 had significantly higher TPM
537 in cerebellum (Figure 7A, Supplementary File 3J). TPMs of Tx9 in CBM were approximately 10x
538 higher than the average for cortical regions and 3x higher than in caudate. This isoform, classified as a
539 COJ and predicted to encode a 1425 aa protein, accounted for ~5% of *DCC* expression. Tx9 uses an
540 alternative 3'SS (-60 nt) in cassette exon 17 and the skipped nucleotides cover an extracellular region
541 and fibronectin type-III domain (UniProt). The SZ risk gene double C2 domain alpha (*DOC2A*) had
542 two novel isoforms with significant variation in brain specific expression including Tx8 in cerebellum
543 and Tx41 in caudate (Figure 7B, C, Supplementary File 3K). Novel Tx8 used a novel splice donor in
544 canonical 5'UTR exon 1 (GT, +158 nt) and was predicted to encode a 400 aa protein unchanged from
545 the canonical transcript. Tx41 was the only novel transcript that showed moderate but specific
546 expression in caudate samples or any tissue other than cerebellum. Tx41 extends the known isoform
547 ENST00000574405 to the canonical stop and is predicted to encode a 400 aa protein. Overall, 28
548 novel isoforms in 11 risk genes were found to have variable expression amongst brain tissues
549 supporting a role for these isoforms within specific brain regions or potentially in a subset of cells.

550

551 **A**

552

553 **B** **C**

554

555

556

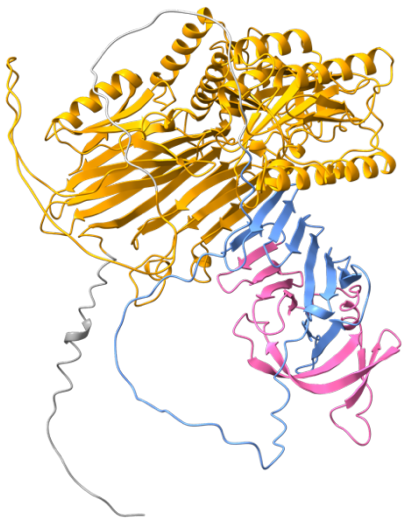
557

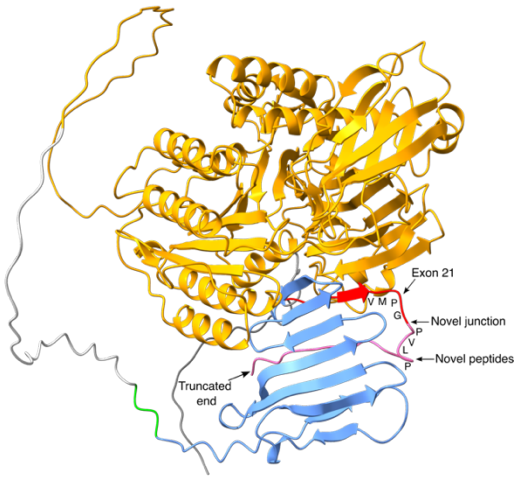
558

559

560

561





562 **D**

563

564

565

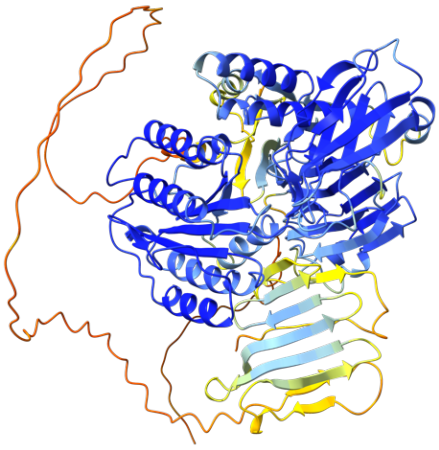
566

567

568

569

570



571

572 **Figure 6. *ITIH4* canonical and novel isoform protein structure predictions. A.** IsoVis stack of the top seven
573 *ITIH4* isoforms sorted by expression. Several novel isoforms contained the novel exon 22 skipping event (red
574 box) including Txs71 and 94. **B.** Canonical isoform (ENST00000266041, UniProt:Q14624) structure prediction
575 indicating 70 kDa (orange) and 30 kDA (blue) chains **C.** Novel isoform (Tx71) structure prediction indicating 70
576 kDa chain (orange), truncated 30 kDA chain (blue), O-glycosylation sites (green), novel splice junction peptide
577 detected using mass spectrophotometry (red) and novel peptides (pink). Black arrow indicates termination <50

578 nt from the final exon junction complex. **D.** AlphaFold per-residue confidence scores (pLDDT) (0-100) for *ITIH4*
579 novel transcript 71: very high (>90, blue), confident (90-70, light-blue), low (70>50, yellow) and very low (<50,
580 orange).

581

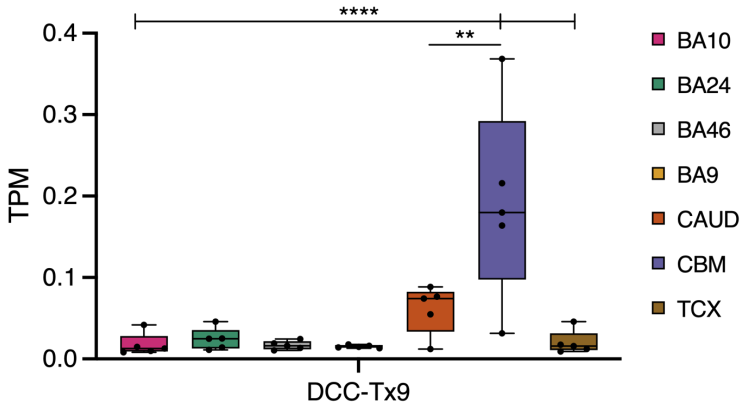
582 **Sequencing and validation of novel exons**

583 Our amplicon sequencing approach detected a total of 28 novel exons in 13 MHD risk genes. Using
584 RT-PCR followed by Sanger sequencing we validated 21/21 targeted novel exons (Table 1). The SZ
585 risk gene chloride voltage-gated channel 3 (*CLCN3*) contained four novel exons within six novel
586 isoforms and an example of PCR validation is shown in Supplementary Figure 12A. Validated novel
587 exon mean length was 99 nt, ranging from 41 nt (*CLCN3*) to 231 nt (*GRM3*). 16 (76%) of validated
588 novel exons were classified as 'poison exons' as they encoded a PTC (Supplementary Figure 12B),
589 although two of these poison exons, within *NEGR1* and *XRN2*, were <50 nt from the final exon
590 junction and therefore may not undergo NMD. The novel exon contained within Tx3 for *XRN2* had
591 the second highest isoform read count for the gene, following the canonical transcript
592 (ENST0000037191), with 4.7% of assigned reads. If translated, this transcript would omit an omega-
593 N-methylarginine modification site (ARG:946) within a disordered region at the C-terminus
594 (Supplementary Figure 13, Supplementary File 3AE).

595 Three novel exons were in untranslated regions and two were predicted to retain the ORF, including
596 the 42 nt exon in *NEGR1* mentioned previously and a 60 nt exon within *SORCS3*. *SORCS3* is a
597 member of the VPS10 transmembrane protein family and assists with neuronal protein trafficking and
598 sorting and a lack of *SORCS3* in the hippocampus in mice has been associated with impaired learning
599 and fear memory in mice [59-61]. The novel exon in Tx1, encoding 20 aa
600 (AMCGRAQWFTPVILALWETE), falls within the *SORCS3* lumenal region (position:
601 LYS:956/PRO:957) and did not appear to disrupt the transmembrane or cytoplasmic domains.
602 Comparison of protein prediction models of the canonical (ENST00000369701.8) and novel isoform
603 (Tx1) showed the addition of an unstructured loop with a partial alpha helix within the second
604 polycystic kidney disease (PKD2) domain (Figure 8A-D), though the prediction confidence was low,
605 so the structural impact on the PKD2 domain remains uncertain [52].

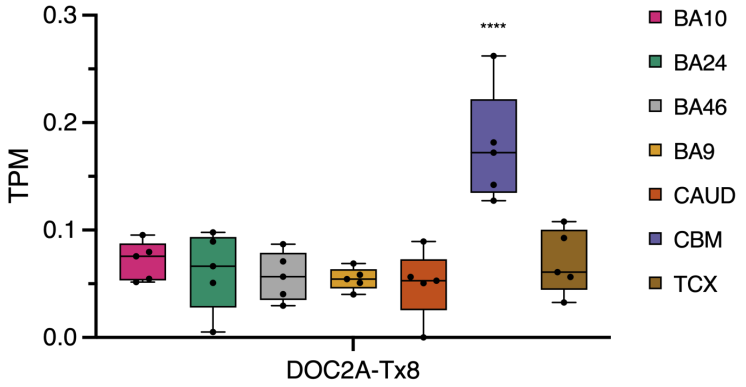
606

607 **A**



608

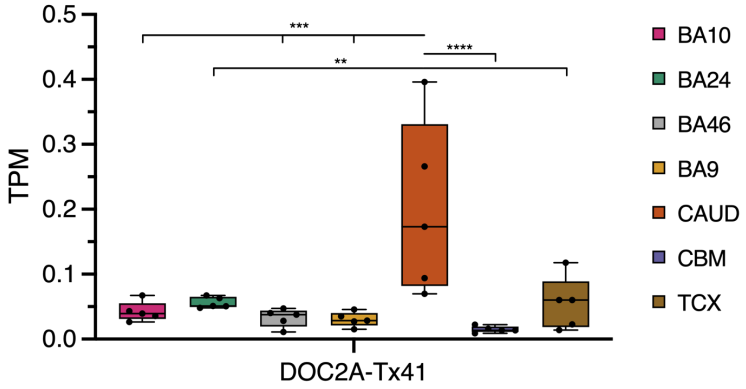
609 **B**



610

611

612 **C**



613

614 **Figure 7. Brain region specific expression of novel isoforms. A.** *DCC* novel transcript 9 (Tx9) uses a known
615 (ENST00000581580) alternative 3'SS (-60 nt) in cassette exon 17 and had significantly higher read counts in
616 CBM. **B-C.** *DOC2A* novel transcripts. Tx8 used a novel splice donor in canonical 5'UTR exon 1 (GT, +158 nt).
617 Tx41 extended the known reference isoform ENST00000574405 to the canonical translational stop and had
618 significantly higher reads in caudate. Brodmann's Area (BA), caudate (CAUD), cerebellum (CBM) and temporal
619 cortex (TCX). Adjusted P value: **= P ≤ 0.01, ***=P ≤ 0.001, ****=P ≤ 0.0001.

620

621

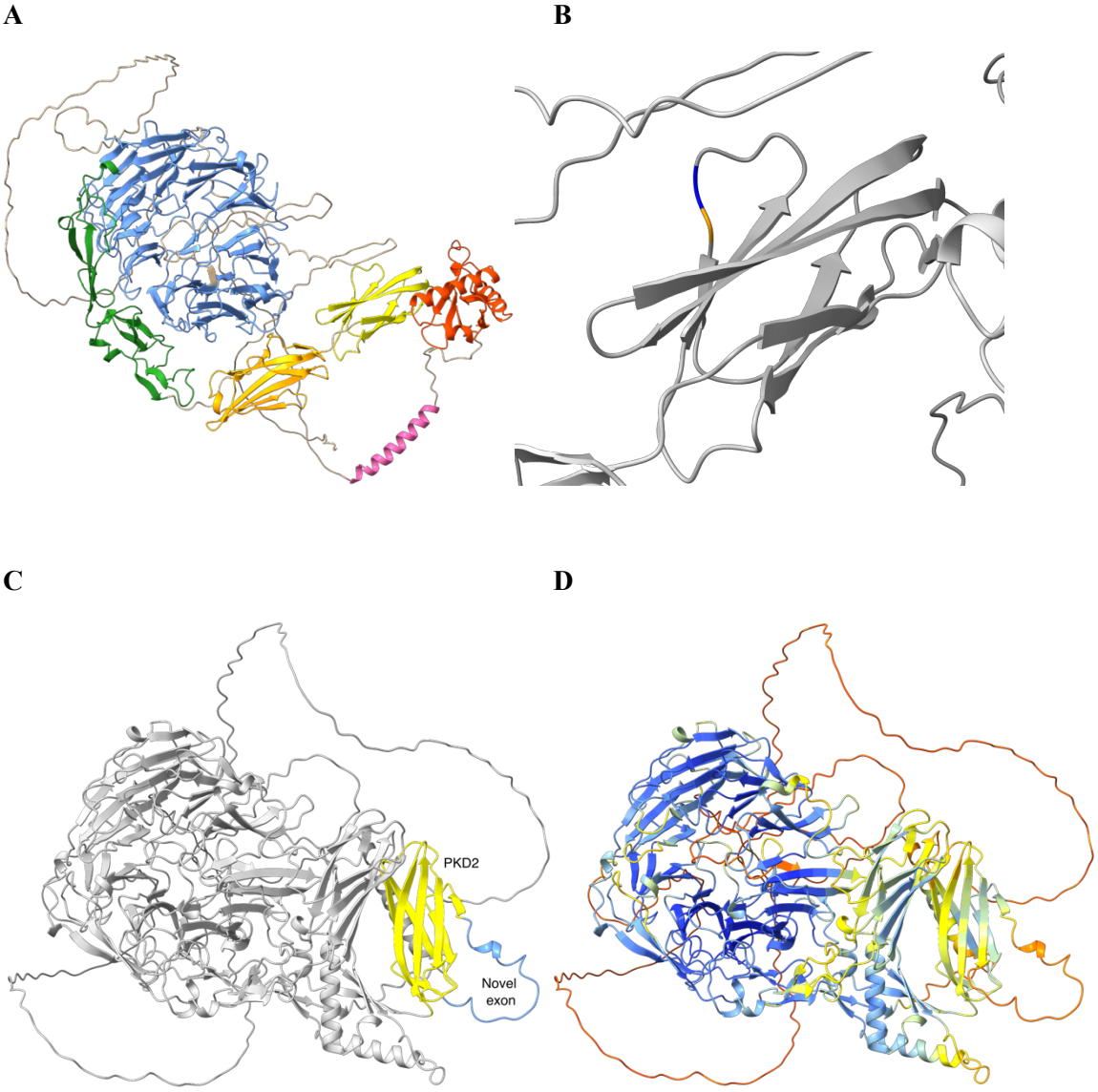


Figure 8. *SORCS3* novel exon and protein structure predictions. A. AlphaFold prediction of *SORCS3* canonical isoform (ENST00000369701.8: Q17R88) coloured by domain: β-propeller (blue), 10CC domain domains (green), polycystic kidney disease (PKD) domains PKD1 (orange) and PKD2 (yellow), SorCS membrane proximal (SoMP) (red) and transmembrane domain (pink). **B.** Zoomed view of the PKD2 domain indicating LYS:956 (blue) and PRO:957 (orange) where frame-retaining novel exon 20a (60 nt) is inserted. **C.** Protein structure prediction of novel transcript 1 (Tx1) containing the novel exon (blue) within the PKD2 domain (yellow). **D.** AlphaFold per-residue confidence scores (pLDDT) (0-100) for novel transcript 1: very high (>90, blue), confident (90-70, light-blue), low (70>50, yellow) and very low (<50, orange).

Table 1. Neuropsychiatric disorder risk gene novel exon validation.

Definitions: chromosome (Chr), nucleotide (nt), open reading frame (ORF), premature termination codon (PTC), untranslated region (UTR).

Gene	Novel exon	Chr	Genomic coordinates *Start*	*End*	Size (nt)	Classification	UniProt/Pfam domain
SORCS3	20a	10	105244778	105244837	60	ORF	Lumenal
GRIA1	2a	5	153509750	153509806	57	PTC	Extracellular
XRN2	16a	20	21345834	21345887	54	PTC	None
XRN2	29a	20	21387308	21387371	64	PTC (<50nt)	Disordered
SLC30A9	6a	4	42028140	42028260	121	PTC	Helix
SLC30A9	9a	4	42059964	42060036	73	PTC	Cation efflux family
GRM3	2a	7	86776784	86777014	231	PTC	None
GRM3	3a	7	86832296	86832411	116	PTC	None
NEGR1	6a	1	71532866	71532907	42	ORF	None
NEGR1	6b	1	71587343	71587400	58	PTC (<50nt)	Transmembrane
RFTN2	1a	2	197654257	197654391	135	PTC	None
RFTN2	1b	2	197671542	197671664	123	PTC	None
RFTN2	6a	2	197616952	197617073	122	PTC	None
CNTN4	1a	3	2656485	2656587	103	PTC	None
PTK2B	3a	8	27318676	27318793	118	UTR	None
PTK2B	3a (short)	8	27318696	27318793	98	UTR	None
PTK2B	3b	8	27319918	27319985	68	UTR	None
CLCN3	1a	4	169630165	169630335	171	UTR	None
CLCN3	2a	4	169638597	169638742	146	PTC	Cytoplasmic
CLCN3	2b	4	169640168	169640208	41	PTC	Cytoplasmic
CLCN3	2c	4	169663527	169663617	91	PTC	Cytoplasmic

637

Discussion

In this study we used long-read sequencing to profile 31 neuropsychiatric disorder risk genes identifying 360 novel RNA isoforms. We also present a new bioinformatic tool, IsoLamp, that can accurately identify and quantify novel RNA isoforms from long-read amplicon data. The recent proliferation of GWAS studies examining increasingly large population-wide data has identified hundreds of genomic variants associated with risk of developing a mental health disorder [23, 62]. Evidence suggests that some risk variants, specifically those that are non-coding, play a role in pre-mRNA splicing and our current understanding of the transcriptomic profile for these risk genes is limited [33, 63]. A key finding of our research is both the high number of novel expressed RNA isoforms and, for some risk genes, the high expression of novel isoforms both individually and collectively. This finding reflects both the known complexity of alternative splicing in the human brain [64] and the current incompleteness of the reference transcriptome. As a result of the relatively deep sequencing afforded by this long-read approach, we have shown that there is a much higher level of RNA isoform diversity for these genes than reported in the current reference annotations. These findings provide new insight into the repertoire of RNA isoforms expressed in brain that could be important for understanding the risk and onset of neuropsychiatric disorders.

RNA isoform discovery, classification and visualisation

We generated a set of high-confidence RNA isoforms from nanopore long-read data using IsoLamp. IsoLamp optimises and streamlines transcript identification, quantification and annotation from long-read amplicon data and outperformed other methods. IsoLamp improves upon our previously published pipeline TAQLoRe [33] by expanding analysis capabilities to any gene and identifying all isoforms within a single pipeline. Our overall approach also overcomes the significant challenge of re-assembling and classifying RNA isoforms using short reads [65-67]. The primary outputs from IsoLamp, filtered transcripts (GTF) and transcript expression (TPM) is designed to be compatible with multiple downstream tools, including our visualisation tool IsoVis (https://isomix.org/isovis) (Wan et al., 2024, *under review*).

Taken together, our benchmarking results highlight that IsoLamp's optimised isoform discovery parameters, coupled with its specific filters (expression, overlapping primers, and full-length reads), yield significant improvements in both precision and recall compared to Bambu, FLAIR, FLAMES and StringTie2. The TPM filter applied to the data presented in this study is conservative and for long and complex genes may need to be tested to yield a balance of novel isoform detection and acceptable expression levels. IsoLamp also output consistent expression quantification that was robust to the quality of the annotations provided.

Novel RNA isoforms in neuropsychiatric disorders

The results presented in this study confirm our current limited understanding of RNA isoform profiles in the human brain and demonstrates how long-read sequencing technologies are a powerful tool to address this issue [32, 33, 68].

Several novel isoforms and exons were identified for ion homeostasis and channel genes *CLCN3*, *CACNA1C* and *SLC30A9* which have shared risk for SZ, BPD and ASD [21, 23, 62]. Voltage-gated ion channels are widely distributed in the brain and regulate neuronal firing. Mutations to these genes have been associated with disease and the emerging role of these channels in neuropsychiatric disorders has been previously reviewed [69]. A study of SZ, BPD and MDD patients versus controls reported gene expression changes in human striatum for potassium, calcium and chloride channel genes including *CACNA1C* and *CLCN3* [70]. *CLCN3* belongs to the CLC family of anion channels and transporters and has an established role in human neurodevelopment [71, 72]. We identified and validated four novel exons in *CLCN3*, three of which were predicted to encode a PTC which could lead to NMD. The fourth was located within the 5'UTR, an area known to impact translation

685 regulation in humans potentially through structural interference with the ribosome [73]. Splice
686 variants of *CLCNC3* have been shown to impact intracellular localisation and our results confirmed
687 these splice variants and further add to them, in particular a novel RNA isoform (Tx9) which is
688 similar to ENST0000613795, but includes the 76 bp exon 12 [71]. Twenty-two novel isoforms were
689 identified for calcium channel gene *CACNA1C* and supporting previous findings [33] the top ten
690 novel isoforms, by assigned read count, were classified as frame retaining, supporting their potential
691 to generate functional proteins. *SLC30A9* (first known as *HUEL*) encodes the zinc transporter protein
692 9 (ZnT9) which is involved in zinc transport and homeostasis in the endoplasmic reticulum and also
693 localises to the cytoplasm and nucleus [74]. Whilst the function of the protein is not fully understood,
694 a 3 nt familial deletion (c.1047_1049delGCA) in the highly conserved cation efflux domain (CED)
695 has been recorded to result in changes to protein structure, intracellular zinc levels and intellectual
696 disability [75, 76]. Critically, we identified and validated a novel exon (Tx3-9a:73 nt) within this CED
697 providing evidence that this region may be alternatively spliced more commonly than previously
698 understood, potentially impacting protein function [74].

699 Several novel exons were detected in the long reads for post-mortem brains used in this study.
700 Previous studies have identified approximately 250 conserved neuronal micro-exons (3-27 nt) that
701 typically preserve the ORF and are involved in neuronal differentiation [77, 78]. Although not
702 classified as micro-exons, a third (N=7) of the validated novel exons found in post-mortem brain in
703 our study were <65 nt, similar to previously reported exon lengths that are difficult for the splicing
704 machinery to process, indicating that these exons may also be inefficiently spliced [79]. Micro-exons
705 are known to be disrupted, usually through increased skipping events, in the brains of individuals with
706 ASD [78, 80]. Importantly, we identified novel exons in two ASD risk genes, *SORCS3* and *XRN2*.
707 *SORCS3*, which has remarkable shared neuropsychiatric disorder risk, contained a novel exon (22 aa)
708 predicted to preserve the reading frame [59, 81]. How these novel exons impact expression or
709 function of a mature protein remains unknown. Further studies with a focus on the transcriptional
710 landscape in the developing brain will be crucial in furthering our understanding of these splicing
711 events [80].

712 **Limitations and future directions**

713 The results of our study are limited by the sample size of available control post-mortem brain tissue.
714 The nanopore long-read data for each risk gene was generated from five elderly, male control
715 individuals, with a single female sample removed from further analyses due to quality constraints.
716 These results may not be representative of all populations, particularly when it comes to novel exons
717 and lowly expressed novel RNA isoforms. Individual transcriptome-wide variation has been observed
718 between age, sex and pathology [82, 83], despite this we did not detect high levels of individual
719 isoform expression variation and novel isoforms were almost always present in all individuals. The
720 small number of available individuals means this dataset was not powered to investigate genotype
721 impacts on isoform expression, though this will be an important area of investigation to determine
722 which risk genotypes act through changes in isoform structure and/or expression.

723 Sample and RNA quality, as measured by RINe, is critical to high-quality sequencing and this is
724 especially true for long reads [33, 84]. Supporting previous findings in mRNA, our data suggest that
725 pH values <6.3 impacted the quality of post-mortem human brain RNA, which is especially critical
726 for robust amplification of longer (>5 Kb) CDS [85]. In general, PCR cycling was kept as low as
727 possible to avoid PCR bias towards shorter isoforms and other artifacts. However, we noted that lower
728 RINs, as recorded for Ind04, appeared to impact amplicons of longer CDS. To help overcome such
729 issues, future long-read amplicon sequencing could incorporate unique molecular identifiers to tag
730 molecules prior to PCR to ensure an accurate representation of the original RNA isoforms [86].

731 The risk genes profiled in this study were selected based on multiple levels of evidence for their
732 involvement in risk, not only from GWAS but from meta-analyses and further independent studies

[22, 25]. Whilst this approach was expected to produce a set of genes with high-confidence of their involvement in disorder risk, it is not exhaustive and it will be important to ensure risk gene lists are updated as more evidence from GWAS and other studies becomes available [25, 63]. Genes that are thought to confer resistance to the development of neuropsychiatric disease are also beginning to emerge and the addition investigation of their expression and isoform profiles may also provide valuable insight into disease risk and progression [87]. Additionally, combining whole or amplicon transcriptomic data, large-scale proteomic data and machine-learning predictive models like TRIFID can help to identify and prioritise functional proteomic isoforms [88].

Conclusion

In conclusion, we identified several hundred unreported RNA isoforms and novel exons, many of which could impact the function of known neuropsychiatric risk genes that also play crucial roles in normal neuronal development and activity. An understanding of the regulatory and functional impacts of these novel isoforms and incorporating long-read Nanopore data into existing repositories will help form an important knowledge base of alternative splicing in the human brain [89, 90]. Some novel isoforms or exons may also be future therapeutic targets through the modulation of splice isoforms using antisense oligonucleotides or CRISPR technology.

Conflict of interest

RDP, YP, YY, JG and MBC have received financial support from Oxford Nanopore Technologies (ONT) to present their findings at scientific conferences. ONT played no role in study design, execution, analysis or publication.

Data availability

All raw nanopore long-read data (fastq) generated for each of the genes reported in this manuscript will be uploaded to The European Genome-Phenome Archive (EGA) upon publication. Isoform GTFs and counts tables from IsoLamp analysis are available upon request. Scripts used for long-read data preparation and downstream data analysis are also available in supplementary material or upon request. IsoLamp is open source and freely available (https://github.com/ClarkLaboratory/IsoLamp). IsoVis is freely available (https://isomix.org/isovis/).

Acknowledgements

The authors would like to acknowledge that brain tissues were received from the Victorian Brain Bank (VBB), supported by The Florey, The Alfred and the Victorian Institute of Forensic Medicine and funded in part by Parkinson's Victoria, MND Victoria and FightMND. The authors would also like to thank Geoff Pavey at the VBB for his assistance with frozen tissue preparation. This research was supported by The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

767 # References

768 1. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing**
769 **complexity in the human transcriptome by high-throughput sequencing.** *Nature genetics*
770 2008, **40:**1413-1415.
771 2. Kelemen O, Convertini P, Zhang Z, Wen Y, Shen M, Falaleeva M, Stamm S: **Function of**
772 **alternative splicing.** *Gene* 2013, **514:**1-30.
773 3. Nilsen TW, Graveley BR: **Expansion of the eukaryotic proteome by alternative splicing.**
774 *Nature* 2010, **463:**457-463.
775 4. Leung SK, Jeffries AR, Castanho I, Jordan BT, Moore K, Davies JP, Dempster EL, Bray NJ,
776 O'Neill P, Tseng E: **Full-length transcript sequencing of human and mouse cerebral cortex**
777 **identifies widespread isoform diversity and alternative splicing.** *Cell reports* 2021, **37**.
778 5. Mazin P, Xiong J, Liu X, Yan Z, Zhang X, Li M, He L, Somel M, Yuan Y, Phoebe Chen YP:
779 **Widespread splicing changes in human brain development and aging.** *Molecular systems*
780 *biology* 2013, **9:**633.
781 6. Baralle FE, Giudice J: **Alternative splicing as a regulator of development and tissue identity.**
782 *Nature reviews Molecular cell biology* 2017, **18:**437-451.
783 7. De Paoli-Iseppi R, Gleeson J, Clark MB: **Isoform Age-Splice Isoform Profiling Using Long-**
784 **Read Technologies.** *Frontiers in Molecular Biosciences* 2021, **8**.
785 8. Castaldi PJ, Abood A, Farber CR, Sheynkman GM: **Bridging the splicing gap in human**
786 **genetics with long-read RNA sequencing: finding the protein isoform drivers of disease.**
787 *Human Molecular Genetics* 2022, **31:**R123-R136.
788 9. Stanley RF, Abdel-Wahab O: **Dysregulation and therapeutic targeting of RNA splicing in**
789 **cancer.** *Nature cancer* 2022, **3:**536-546.
790 10. Vitting-Seerup K, Sandelin A: **IsoformSwitchAnalyzeR: analysis of changes in genome-wide**
791 **patterns of alternative splicing and its functional consequences.** *Bioinformatics* 2019,
792 **35:**4469-4471.
793 11. Manuel JM, Guilloy N, Khatir I, Roucou X, Laurent B: **Re-evaluating the impact of alternative**
794 **RNA splicing on proteomic diversity.** *Frontiers in Genetics* 2023, **14:**1089053.
795 12. Melé M, Ferreira PG, Reverter F, DeLuca DS, Monlong J, Sammeth M, Young TR, Goldmann
796 JM, Pervouchine DD, Sullivan TJ: **The human transcriptome across tissues and individuals.**
797 *Science* 2015, **348:**660-665.
798 13. Carvill GL, Engel KL, Ramamurthy A, Cochran JN, Roovers J, Stamberger H, Lim N, Schneider
799 AL, Hollingsworth G, Holder DH: **Aberrant inclusion of a poison exon causes dravet**
800 **syndrome and related SCN1A-associated genetic epilepsies.** *The American Journal of*
801 *Human Genetics* 2018, **103:**1022-1029.
802 14. Lara-Pezzi E, Desco M, Gatto A, Gómez-Gaviro MV: **Neurogenesis: regulation by alternative**
803 **splicing and related posttranscriptional processes.** *The Neuroscientist* 2017, **23:**466-477.
804 15. Rehm J, Shield KD: **Global Burden of Disease and the Impact of Mental and Addictive**
805 **Disorders.** *Current Psychiatry Reports* 2019, **21:**10.
806 16. Sandell C, Kjellberg A, Taylor RR: **Participating in diagnostic experience: adults with**
807 **neuropsychiatric disorders.** *Scandinavian Journal of Occupational Therapy* 2013, **20:**136-
808 142.
809 17. Bray NJ, O'Donovan MC: **The genetics of neuropsychiatric disorders.** *Brain and neuroscience*
810 *advances* 2018, **2:**2398212818799271.
811 18. Medalia A, Saperstein AM, Hansen MC, Lee S: **Personalised treatment for cognitive**
812 **dysfunction in individuals with schizophrenia spectrum disorders.** *Neuropsychological*
813 *rehabilitation* 2018, **28:**602-613.
814 19. Mora C, Zonca V, Riva MA, Cattaneo A: **Blood biomarkers and treatment response in major**
815 **depression.** *Expert review of molecular diagnostics* 2018, **18:**513-529.
816 20. Anney RJL, Ripke S, Anttila V, Grove J, Holmans P, Huang H, Klei L, Lee PH, Medland SE, Neale
817 B, et al: **Meta-analysis of GWAS of over 16,000 individuals with autism spectrum disorder**

highlights a novel locus at 10q24.32 and a significant overlap with schizophrenia. *Molecular Autism* 2017, **8:**21.

21. Grove J, Ripke S, Als TD, Mattheisen M, Walters RK, Won H, Pallesen J, Agerbo E, Andreassen OA, Anney R: **Identification of common genetic risk variants for autism spectrum disorder.** *Nature genetics* 2019, **51:**431-444.
22. Pardiñas AF, Holmans P, Pocklington AJ, Escott-Price V, Ripke S, Carrera N, Legge SE, Bishop S, Cameron D, Hamshere ML: **Common schizophrenia alleles are enriched in mutation-intolerant genes and in regions under strong background selection.** *Nature genetics* 2018, **50:**381.
23. Ripke S, Neale BM, Corvin A, Walters JT, Farh K-H, Holmans PA, Lee P, Bulik-Sullivan B, Collier DA, Huang H: **Biological insights from 108 schizophrenia-associated genetic loci.** *Nature* 2014, **511:**421.
24. Stahl EA, Breen G, Forstner AJ, McQuillin A, Ripke S, Trubetskoy V, Mattheisen M, Wang Y, Coleman JR, Gaspar HA: **Genome-wide association study identifies 30 loci associated with bipolar disorder.** *Nature genetics* 2019, **51:**793-803.
25. Trubetskoy V, Pardiñas AF, Qi T, Panagiotaropoulou G, Awasthi S, Bigdeli TB, Bryois J, Chen C-Y, Dennison CA, Hall LS: **Mapping genomic loci implicates genes and synaptic biology in schizophrenia.** *Nature* 2022, **604:**502-508.
26. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, Montgomery GW, Goddard ME, Wray NR, Visscher PM: **Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.** *Nature genetics* 2016, **48:**481-487.
27. Sey NY, Hu B, Mah W, Fauni H, McAfee JC, Rajarajan P, Brennand KJ, Akbarian S, Won H: **A computational tool (H-MAGMA) for improved prediction of brain-disorder risk genes by incorporating brain chromatin interaction profiles.** *Nature Neuroscience* 2020, **23:**583-593.
28. Yang A, Chen J, Zhao X-M: **nMAGMA: a network-enhanced method for inferring risk genes from GWAS summary statistics and its application to schizophrenia.** *Briefings in bioinformatics* 2021, **22:**bbaa298.
29. Devlin B, Kelsoe JR, Sklar P, Daly MJ, O'Donovan MC, Craddock N, Sullivan PF, Smoller JW, Kendler KS: **Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs.** *Nature genetics* 2013, **45:**984-994.
30. Steijger T, Abril JF, Engström PG, Kokocinski F, Hubbard TJ, Guigó R, Harrow J, Bertone P: **Assessment of transcript reconstruction methods for RNA-seq.** *Nature methods* 2013, **10:**1177-1184.
31. Amarasinghe SL, Su S, Dong X, Zappia L, Ritchie ME, Gouil Q: **Opportunities and challenges in long-read sequencing data analysis.** *Genome biology* 2020, **21:**1-16.
32. Glinos DA, Garborcauskas G, Hoffman P, Ehsan N, Jiang L, Gokden A, Dai X, Aguet F, Brown KL, Garimella K: **Transcriptome variation in human tissues revealed by long-read sequencing.** *Nature* 2022, **608:**353-359.
33. Clark MB, Wrzesinski T, Garcia AB, Hall NAL, Kleinman JE, Hyde T, Weinberger DR, Harrison PJ, Haerty W, Tunbridge EM: **Long-read sequencing reveals the complex splicing profile of the psychiatric risk gene CACNA1C in human brain.** *Molecular Psychiatry* 2020, **25:**37-47.
34. Ma L, Semick SA, Chen Q, Li C, Tao R, Price AJ, Shin JH, Jia Y, Consortium B, Brandon NJ: **Schizophrenia risk variants influence multiple classes of transcripts of sorting nexin 19 (SNX19).** *Molecular psychiatry* 2020, **25:**831-843.
35. Buniello A, MacArthur JAL, Cerezo M, Harris LW, Hayhurst J, Malangone C, McMahon A, Morales J, Mountjoy E, Sollis E: **The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019.** *Nucleic acids research* 2019, **47:**D1005-D1012.
36. Nassar LR, Barber GP, Benet-Pagès A, Casper J, Clawson H, Diekhans M, Fischer C, Gonzalez JN, Hinrichs AS, Lee BT: **The UCSC genome browser database: 2023 update.** *Nucleic acids research* 2023, **51:**D1188-D1195.

37. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA: **Primer3Plus, an enhanced web interface to Primer3.** *Nucleic acids research* 2007, **35:**W71-W74.
38. Raney BJ, Dreszer TR, Barber GP, Clawson H, Fujita PA, Wang T, Nguyen N, Paten B, Zweig AS, Karolchik D: **Track data hubs enable visualization of user-defined genome-wide annotations on the UCSC Genome Browser.** *Bioinformatics* 2014, **30:**1003-1005.
39. Kuhn RM, Haussler D, Kent WJ: **The UCSC genome browser and associated tools.** *Briefings in bioinformatics* 2013, **14:**144-161.
40. Bushnell B: **BBMap: a fast, accurate, splice-aware aligner.** Lawrence Berkeley National Lab.(LBNL), Berkeley, CA (United States); 2014.
41. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018, **34:**3094-3100.
42. Chen Y, Sim A, Wan YK, Yeo K, Lee JJX, Ling MH, Love MI, Göke J: **Context-aware transcript quantification from long-read RNA-seq data with Bambu.** *Nature Methods* 2023**:**1-9.
43. Quinlan AR: **BEDTools: the Swiss-army tool for genome feature analysis.** *Current protocols in bioinformatics* 2014, **47:**11.12. 11-11.12. 34.
44. Pertea G, Pertea M: **GFF utilities: GffRead and GffCompare.** *F1000Research* 2020, **9**.
45. Patro R, Duggal G, Kingsford C: **Salmon: accurate, versatile and ultrafast quantification from RNA-seq data using lightweight-alignment.** 2015.
46. Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C: **Salmon provides fast and bias-aware quantification of transcript expression.** *Nature methods* 2017, **14:**417-419.
47. Kovaka S, Zimin AV, Pertea GM, Razaghi R, Salzberg SL, Pertea M: **Transcriptome assembly from long-read RNA-seq alignments with StringTie2.** *Genome biology* 2019, **20:**1-13.
48. Tang AD, Soulette CM, van Baren MJ, Hart K, Hrabeta-Robinson E, Wu CJ, Brooks AN: **Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns.** *Nature communications* 2020, **11:**1438.
49. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden TL: **Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction.** *BMC bioinformatics* 2012, **13:**1-11.
50. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A: **Highly accurate protein structure prediction with AlphaFold.** *Nature* 2021, **596:**583-589.
51. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE: **UCSF ChimeraX: Structure visualization for researchers, educators, and developers.** *Protein Science* 2021, **30:**70-82.
52. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M: **ColabFold: making protein folding accessible to all.** *Nature methods* 2022, **19:**679-682.
53. Tian L, Jabbari JS, Thijssen R, Gouil Q, Amarasinghe SL, Voogd O, Kariyawasam H, Du MR, Schuster J, Wang C: **Comprehensive characterization of single-cell full-length isoforms in human and mouse with long-read sequencing.** *Genome biology* 2021, **22:**1-24.
54. Tardaguila M, De La Fuente L, Marti C, Pereira C, Pardo-Palacios FJ, Del Risco H, Ferrell M, Mellado M, Macchietto M, Verheggen K: **SQANTI: extensive characterization of long-read transcript sequences for quality control in full-length transcriptome identification and quantification.** *Genome research* 2018, **28:**396-411.
55. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, Bairoch A: **ExPASy: the proteomics server for in-depth protein knowledge and analysis.** *Nucleic acids research* 2003, **31:**3784-3788.
56. Pihl R, Jensen RK, Poulsen EC, Jensen L, Hansen AG, Thøgersen IB, Dobó J, Gál P, Andersen GR, Enghild JJ: **ITIH4 acts as a protease inhibitor by a novel inhibitory mechanism.** *Science advances* 2021, **7:**eaba7381.

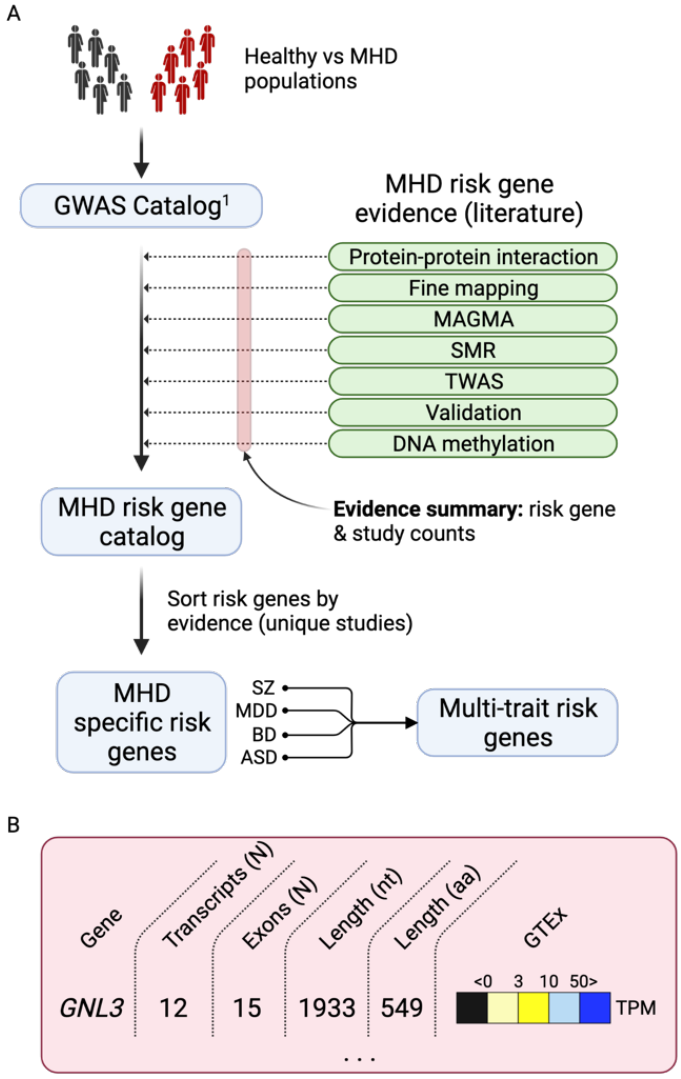
57. Schreiner D, Nguyen T-M, Russo G, Heber S, Patrignani A, Ahrné E, Scheiffele P: **Targeted combinatorial alternative splicing generates brain region-specific repertoires of neurexins.** *Neuron* 2014, **84:**386-398.
58. Joglekar A, Prjibelski A, Mahfouz A, Collier P, Lin S, Schlusche AK, Marrocco J, Williams SR, Haase B, Hayes A, et al: **A spatially resolved brain region- and cell type-specific isoform atlas of the postnatal mouse brain.** *Nature Communications* 2021, **12:**463.
59. Kamran M, Laighneach A, Bibi F, Donohoe G, Ahmed N, Rehman AU, Morris DW: **Independent Associated SNPs at SORCS3 and Its Protein Interactors for Multiple Brain-Related Disorders and Traits.** *Genes* 2023, **14:**482.
60. Dong F, Wu C, Jiang W, Zhai M, Li H, Zhai L, Zhang X: **Cryo-EM structure studies of the human VPS10 domain-containing receptor SorCS3.** *Biochemical and Biophysical Research Communications* 2022, **624:**89-94.
61. Breiderhoff T, Christiansen GB, Pallesen LT, Vaegter C, Nykjaer A, Holm MM, Glerup S, Willnow TE: **Sortilin-related receptor SORCS3 is a postsynaptic modulator of synaptic depression and fear extinction.** *PloS one* 2013, **8:**e75006.
62. Mullins N, Forstner AJ, O'Connell KS, Coombes B, Coleman JR, Qiao Z, Als TD, Bigdeli TB, Børte S, Bryois J: **Genome-wide association study of more than 40,000 bipolar disorder cases provides new insights into the underlying biology.** *Nature genetics* 2021, **53:**817-829.
63. Kim M, Vo DD, Jops CT, Wen C, Patowary A, Bhattacharya A, Yap CX, Zhou H, Gandal MJ: **Multivariate variance components analysis uncovers genetic architecture of brain isoform expression and novel psychiatric disease mechanisms.** *medRxiv* 2022**:**2022.2010.2018.22281204.
64. Yeo G, Holste D, Kreiman G, Burge CB: **Variation in alternative splicing across human tissues.** *Genome biology* 2004, **5:**1-15.
65. Sarantopoulou D, Brooks TG, Nayak S, Mrčela A, Lahens NF, Grant GR: **Comparative evaluation of full-length isoform quantification from RNA-Seq.** *BMC bioinformatics* 2021, **22:**1-24.
66. Hu Y, Fang L, Chen X, Zhong JF, Li M, Wang K: **LIQA: long-read isoform quantification and analysis.** *Genome biology* 2021, **22:**182.
67. Zhang C, Zhang B, Lin L-L, Zhao S: **Evaluation and comparison of computational tools for RNA-seq isoform quantification.** *BMC genomics* 2017, **18:**1-11.
68. Arendt-Tranholm A, Mwirigi JM, Price TJ: **RNA isoform expression landscape of the human dorsal root ganglion (DRG) generated from long read sequencing.** *bioRxiv* 2023**:**2023.2010.2028.564535.
69. Imbrici P, Conte Camerino D, Tricarico D: **Major channels involved in neuropsychiatric disorders and therapeutic perspectives.** *Frontiers in Genetics* 2013, **4**.
70. Smolin B, Karry R, Gal-Ben-Ari S, Ben-Shachar D: **Differential expression of genes encoding neuronal ion-channel subunits in major depression, bipolar disorder and schizophrenia: implications for pathophysiology.** *International Journal of Neuropsychopharmacology* 2012, **15:**869-882.
71. Guzman RE, Miranda-Laferte E, Franzen A, Fahlke C: **Neuronal ClC-3 splice variants differ in subcellular localizations, but mediate identical transport functions.** *Journal of Biological Chemistry* 2015, **290:**25851-25862.
72. Duncan AR, Polovitskaya MM, Gaitán-Peñas H, Bertelli S, VanNoy GE, Grant PE, O'Donnell-Luria A, Valivullah Z, Lovgren AK, England EM: **Unique variants in CLCN3, encoding an endosomal anion/proton exchanger, underlie a spectrum of neurodevelopmental disorders.** *The American Journal of Human Genetics* 2021, **108:**1450-1465.
73. Leppek K, Das R, Barna M: **Functional 5' UTR mRNA structures in eukaryotic translation regulation and how to find them.** *Nature reviews Molecular cell biology* 2018, **19:**158-174.

74. Roca-Umbert A, Garcia-Calleja J, Vogel-González M, Fierro-Villegas A, Ill-Raga G, Herrera-Fernández V, Bosnjak A, Muntané G, Gutiérrez E, Campelo F: **Human genetic adaptation related to cellular zinc homeostasis.** *Plos Genetics* 2023, **19:**e1010950.
75. Perez Y, Shorer Z, Liani-Leibson K, Chabosseau P, Kadir R, Volodarsky M, Halperin D, Barber-Zucker S, Shalev H, Schreiber R: **SLC30A9 mutation affecting intracellular zinc homeostasis causes a novel cerebro-renal syndrome.** *Brain* 2017, **140:**928-939.
76. Willekens J, Runnels LW: **Impact of Zinc Transport Mechanisms on Embryonic and Brain Development.** *Nutrients* 2022, **14:**2526.
77. Gonatopoulos-Pournatzis T, Blencowe BJ: **Microexons: at the nexus of nervous system development, behaviour and autism spectrum disorder.** *Current Opinion in Genetics & Development* 2020, **65:**22-33.
78. Irimia M, Weatheritt RJ, Ellis JD, Parikshak NN, Gonatopoulos-Pournatzis T, Babor M, Quesnel-Vallières M, Tapial J, Raj B, O'Hanlon D: **A highly conserved program of neuronal microexons is misregulated in autistic brains.** *Cell* 2014, **159:**1511-1523.
79. Li YI, Sanchez-Pulido L, Haerty W, Ponting CP: **RBFOX and PTBP1 proteins regulate the alternative splicing of micro-exons in human brain transcripts.** *Genome research* 2015, **25:**1-13.
80. Chau KK, Zhang P, Urresti J, Amar M, Pramod AB, Chen J, Thomas A, Corominas R, Lin GN, Iakoucheva LM: **Full-length isoform transcriptome of the developing human brain provides further insights into autism.** *Cell reports* 2021, **36**.
81. Wu Y, Cao H, Baranova A, Huang H, Li S, Cai L, Rao S, Dai M, Xie M, Dou Y: **Multi-trait analysis for genome-wide association study of five psychiatric disorders.** *Translational psychiatry* 2020, **10:**209.
82. Peters MJ, Joehanes R, Pilling LC, Schurmann C, Conneely KN, Powell J, Reinmaa E, Sutphin GL, Zhernakova A, Schramm K: **The transcriptional landscape of age in human peripheral blood.** *Nature communications* 2015, **6:**1-14.
83. Işıldak U, Somel M, Thornton JM, Dönertaş HM: **Temporal changes in the gene expression heterogeneity during brain development and aging.** *Scientific reports* 2020, **10:**4080.
84. Prawer YD, Gleeson J, De Paoli-Iseppi R, Clark MB: **Pervasive effects of RNA degradation on Nanopore direct RNA sequencing.** *NAR Genomics and Bioinformatics* 2023, **5:**lqad060.
85. Harrison PJ, Heath PR, Eastwood SL, Burnet PWJ, McDonald B, Pearson RCA: **The relative importance of premortem acidosis and postmortem interval for human brain gene expression studies: selective mRNA vulnerability and comparison with their encoded proteins.** *Neuroscience Letters* 1995, **200:**151-154.
86. Karst SM, Ziels RM, Kirkegaard RH, Sørensen EA, McDonald D, Zhu Q, Knight R, Albertsen M: **High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing.** *Nature methods* 2021, **18:**165-169.
87. Hess JL, Tylee DS, Mattheisen M, Børglum AD, Als TD, Grove J, Werge T, Mortensen PB: **A polygenic resilience score moderates the genetic risk for schizophrenia.** *Molecular psychiatry* 2021, **26:**800-815.
88. Pozo F, Martinez-Gomez L, Walsh TA, Rodriguez JM, Di Domenico T, Abascal F, Vazquez J, Tress ML: **Assessing the functional relevance of splice isoforms.** *NAR Genomics and Bioinformatics* 2021, **3:**lqab044.
89. Dawes R, Bournazos AM, Bryen SJ, Bommireddipalli S, Marchant RG, Joshi H, Cooper ST: **SpliceVault predicts the precise nature of variant-associated mis-splicing.** *Nature Genetics* 2023, **55:**324-332.
90. Frankish A, Carbonell-Sala S, Diekhans M, Jungreis I, Loveland JE, Mudge JM, Sisu C, Wright JC, Arnan C, Barnes I: **GENCODE: reference annotation for the human and mouse genomes in 2023.** *Nucleic acids research* 2023, **51:**D942-D949.
91. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, Hasz R, Walters G, Garcia F, Young N: **The genotype-tissue expression (GTEx) project.** *Nature genetics* 2013, **45:**580-585.

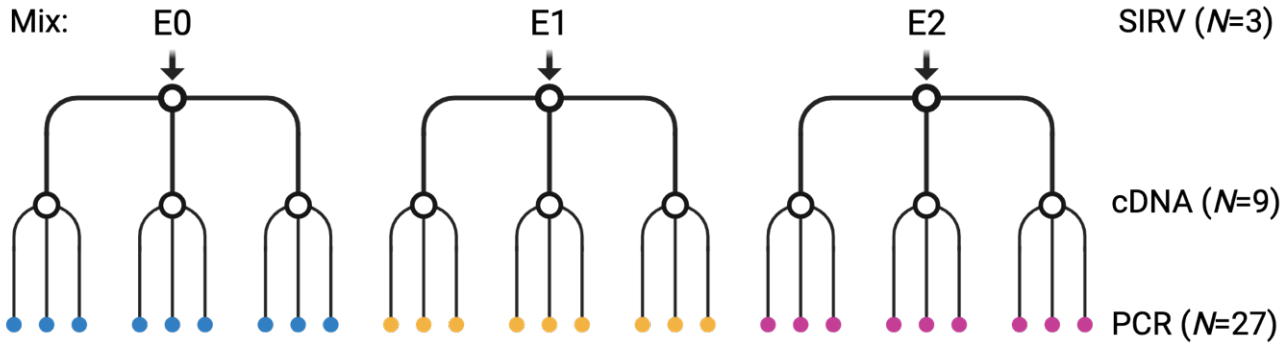
1020 ## Supplementary Figures

1021

1022 **Supplementary Figure 1. Mental health disorder (MHD) risk gene list curation pipeline. A.**
1023 Single nucleotide polymorphism (SNP) catalogues form the foundation of the final risk gene evidence
1024 lists used to select genes for amplicon sequencing. These catalogues are collated from up-to-date,
1025 large-scale genome wide association studies (GWAS) of MHDs. A single GWAS catalogue was
1026 downloaded for schizophrenia (SZ), major depressive disorder (MDD), bipolar disorder (BD) and
1027 autism spectrum disorder (ASD), and associations were filtered according to criteria in Supplementary
1028 Table 2. GWAS data generally had either a mapped or reported gene associated with each SNP.
1029 Further evidence (i.e. reported genes) from categorised literature sources (shown in green) was then
1030 added to the list. This list was then sorted (high to low) by the number of occurrences of a risk gene in
1031 unique studies across all evidence/validation categories. No weighting was applied to any category. A
1032 multi-trait list was also made containing evidence for risk genes across all four MHDs so shared risk
1033 genes could be identified. **B.** An example of additional risk gene information included in the list e.g.
1034 for *GNL3*, count of known transcripts, count of coding exons for the canonical isoform, length in
1035 nucleotides (nt), length of protein in amino acids (aa) and the categorised Genotype-Tissue
1036 Expression (GTEx) in transcript per million (TPM) for each brain tissue [91]. *Definitions*: multi-
1037 marker analysis of Genomic annotation (MAGMA), summary-based Mendelian randomisation
1038 (SMR), transcriptome wide association study (TWAS).

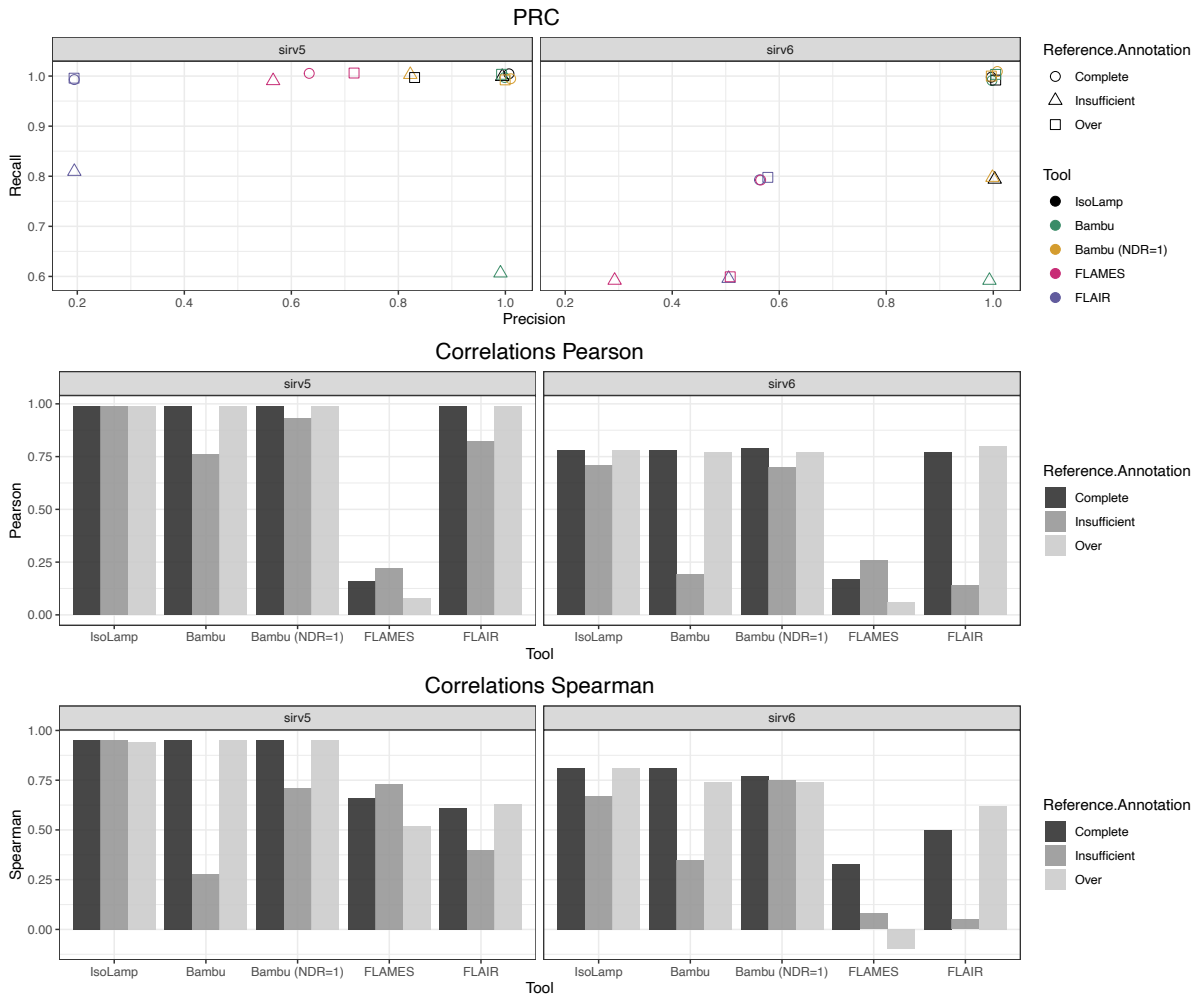


1039 **Supplementary Figure 2.** Experimental design of SIRV amplicon controls. E0, E1 and E2 represent
1040 each SIRV mix of known isoform concentrations. Each mix was converted into cDNA in triplicate
1041 and finally, the full length of synthetic genes SIRV5 and 6 were amplified using PCR in triplicate for
1042 each cDNA replicate.



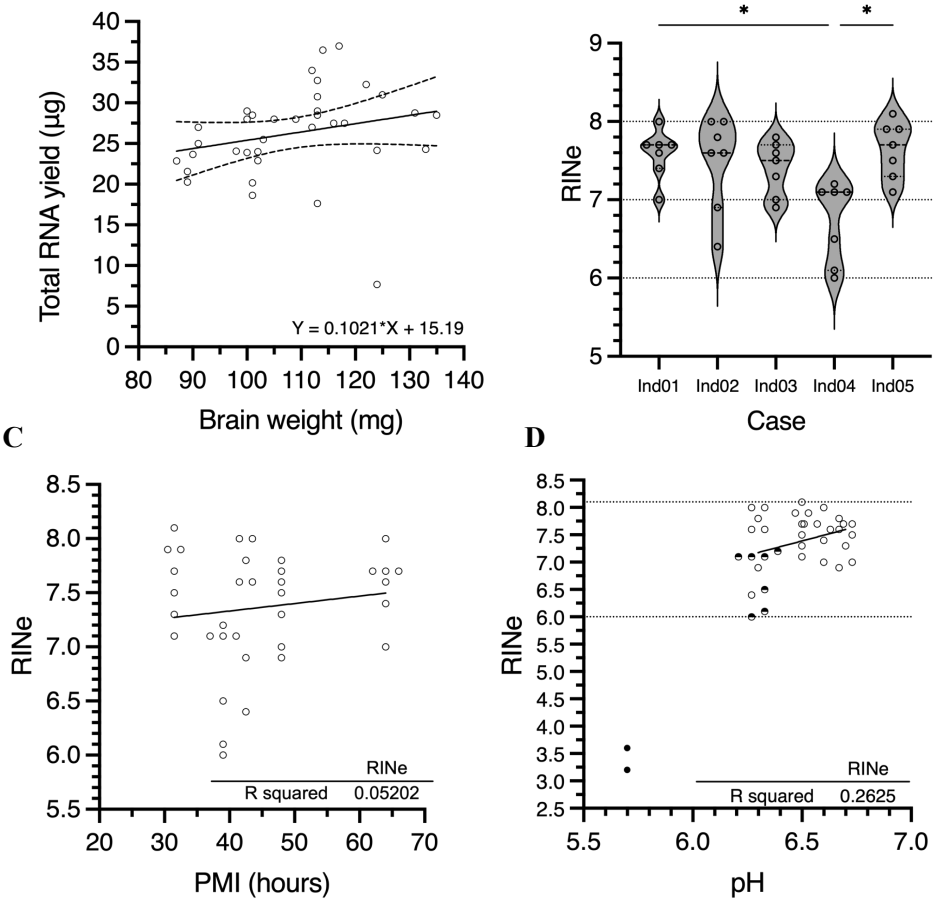
1043

Supplementary Figure 3. Benchmarking IsoLamp using spike-in SIRVs and the optimised IsoLamp expression-based filter. A. Precision recall of each tested pipeline with the complete, insufficient or over annotated SIRV reference, filtering all results using the IsoLamp expression-based filter. IsoLamp (black) returned high quality isoforms from amplicon data of both SIRV5 and 6. Pearson **(B)** and Spearman **(C)** correlations for each pipeline between known and observed expression values for SIRV 5 and 6 mixes.



1052 **Supplementary Figure 4. Post-mortem human brain RNA QC. A.** Yield of isolated total
1053 RNA generally increases with increased tissue weight (mg) input to homogenisation using the
1054 RNeasy Lipid Tissue Mini Kit (QIAGEN:74804). **B.** The RNA integrity number equivalent
1055 (RINe) for brain tissue was generally between 7 – 8. RINe from individual (Ind) 04 was
1056 significantly lower when compared to Individuals 01 and 05. **C.** No correlation was detected
1057 between RNA quality (RINe) and individual post-mortem interval (PMI). **D.** Decreasing
1058 individual brain pH appears to impact RINe. Half-circles indicate samples from individual 04
1059 (pH = 6.3). Full circles indicate degraded RNA isolated from individual 06 which were not
1060 included for further analysis. Data in B, C and D are staggered for clarity.

1061 **A** **B**

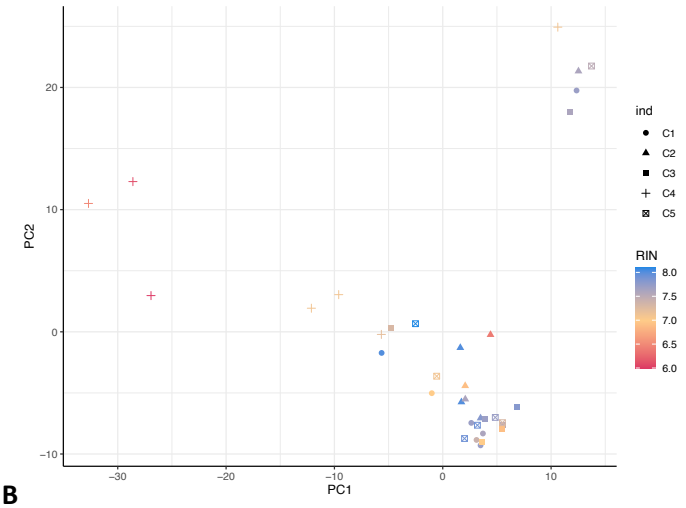


1062 **C** **D**

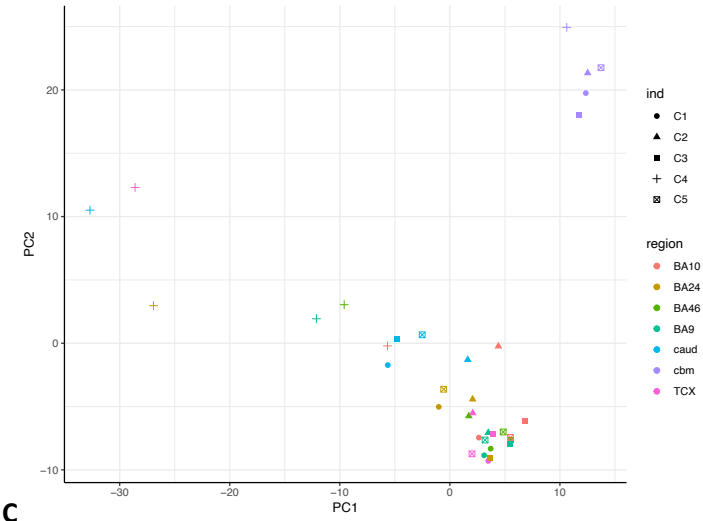
1063

Supplementary Figure 5. Principal component analyses (PCA) of brain samples. PC1 and PC2 coloured by RNA integrity (RIN) **(A)** and brain region **(B)**. **C.** PC4 and PC5 coloured by donor age (years). Key: individual (ind), Brodmann's area (BA), caudate (caud), cerebellum (cbm) and temporal cortex (TCX).

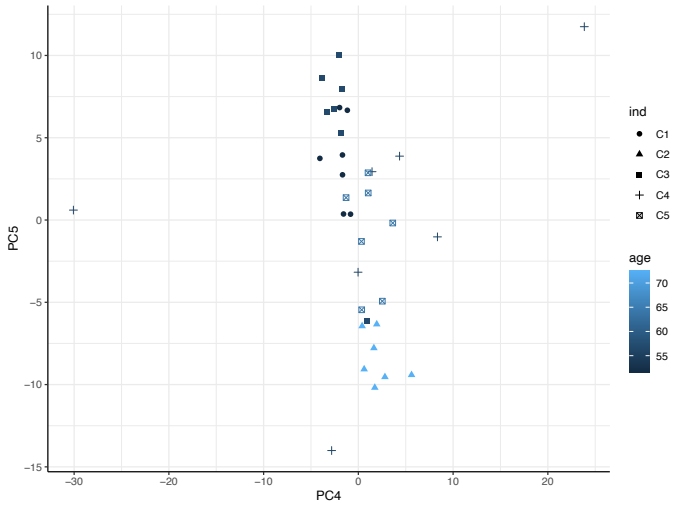
A



B

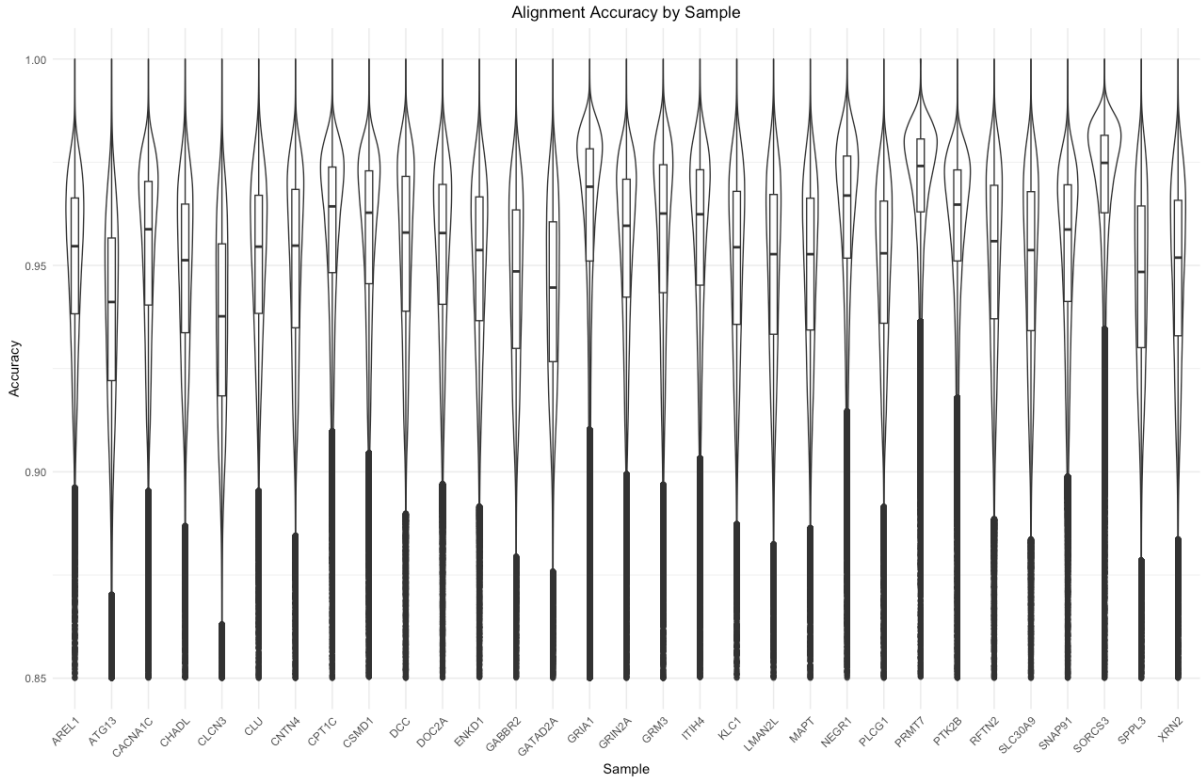


C

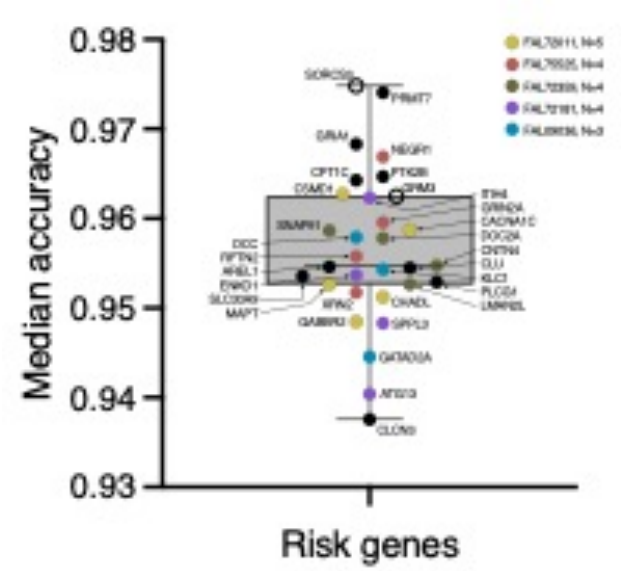


Supplementary Figure 6. Long-read amplicon mapping accuracy. A. All sequenced risk genes. Plotted range 0.85 – 1.00. **B.** A box and whiskers plot of the median accuracy for each risk gene. Open circles indicate the library was prepared with ligation sequencing kit 110 (ONT). Colours indicate flow cells that were used multiple times (N=3-5).

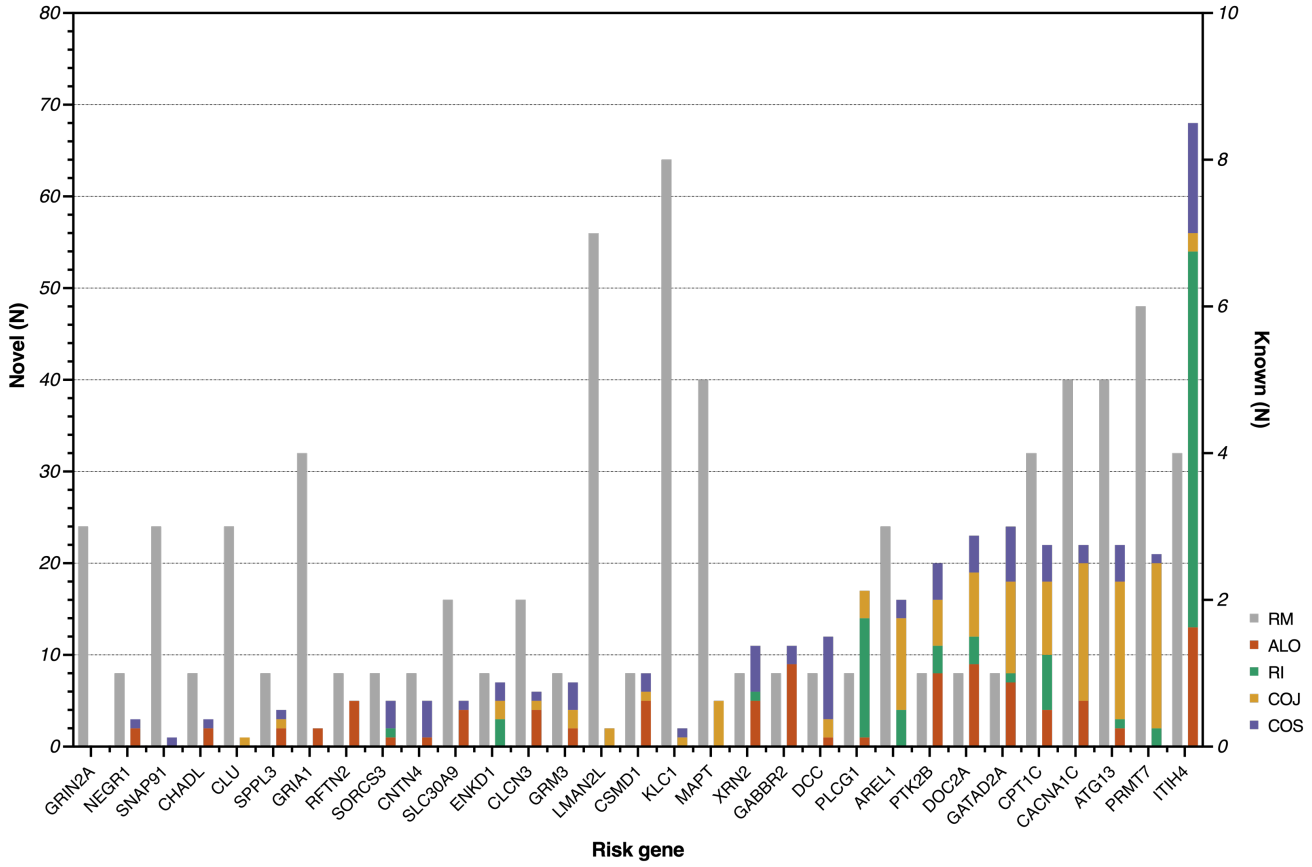
A



B



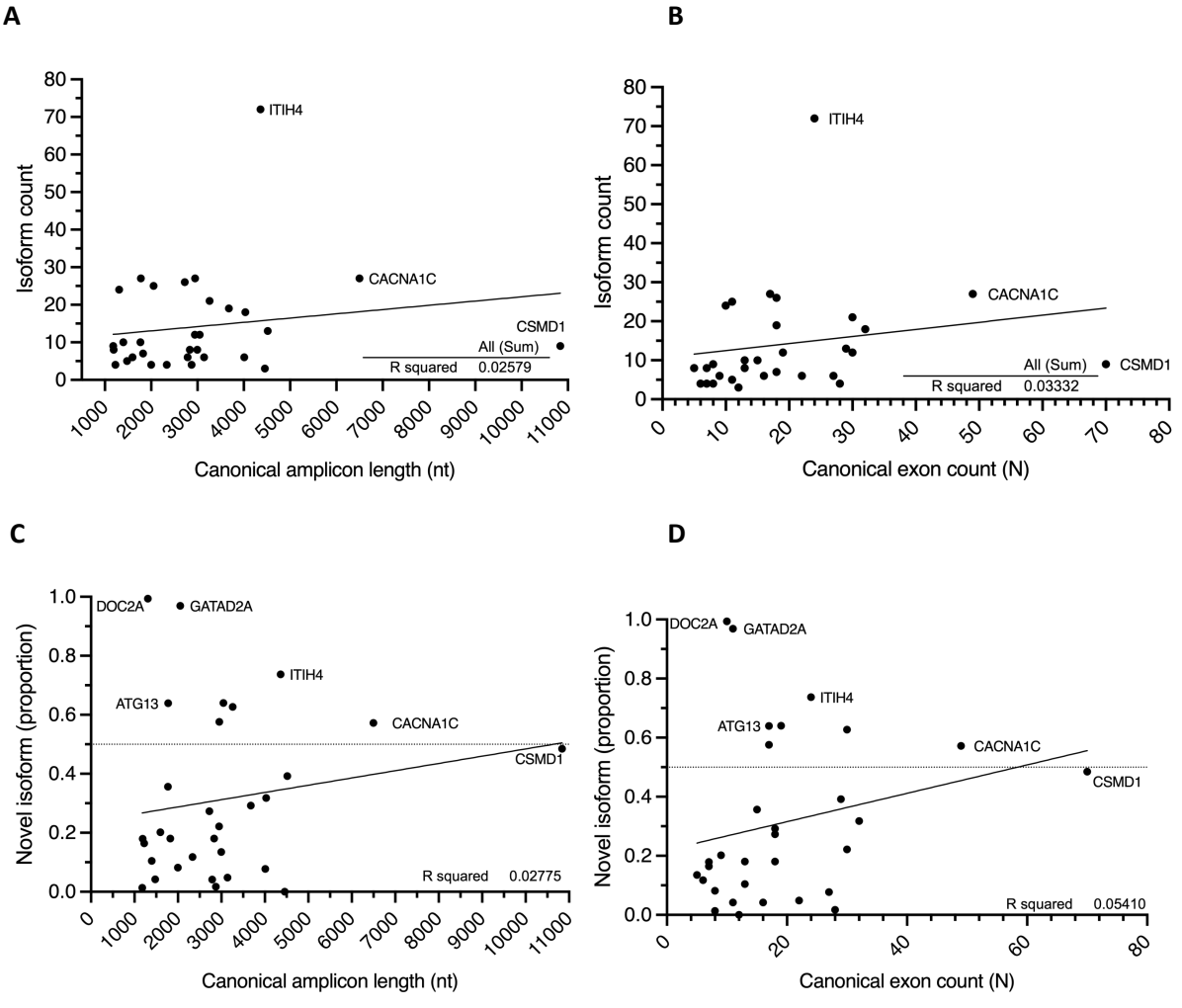
1080 **Supplementary Figure 7. Risk gene isoform counts.** The number of detected isoforms (known and
1081 novel) is shown for each risk gene sorted from lowest (*GRIN2A*) to highest (*ITIH4*). Each isoform was
1082 classified into a SQANTI subcategory: reference match (RM), containing at least one novel splice site
1083 (ALO), retained intron (RI), combination of known junctions (COJ) or splice sites (COS). Known
1084 (RM) isoform counts are plotted on the right Y-axis and novel isoform counts are plotted on the left
1085 Y-axis.



1086

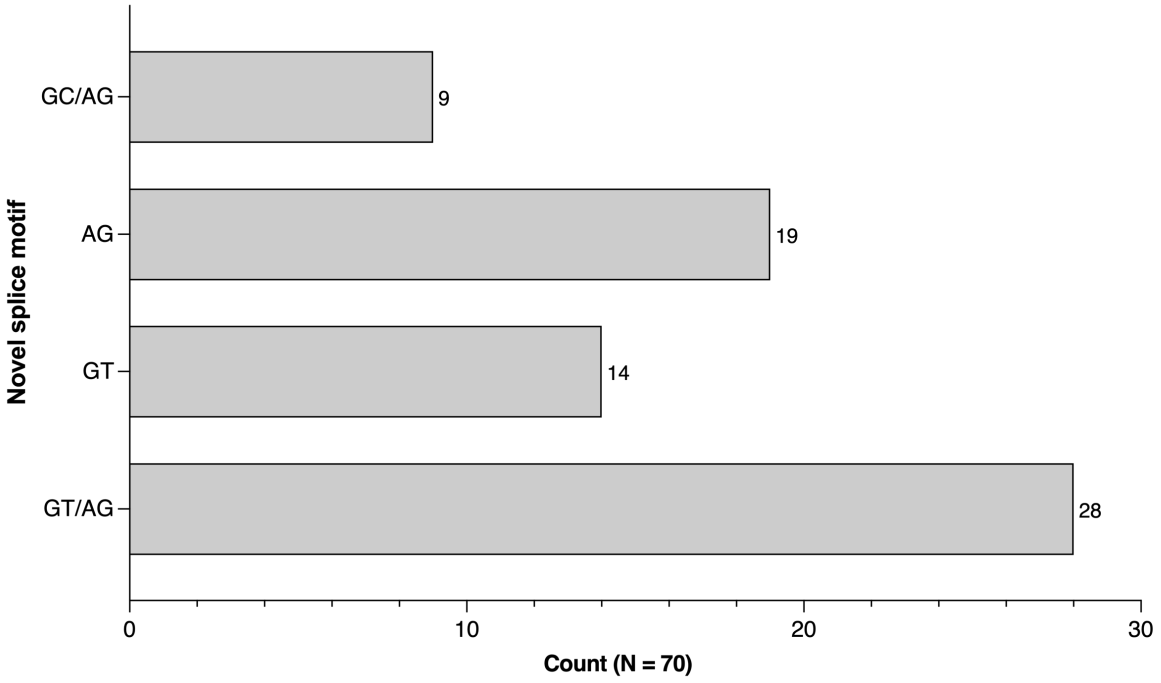
1087

Supplementary Figure 8. Linear regression of amplicon length or canonical exon count against isoform count and novel isoform TPM proportion does not deviate significantly from zero. Linear regression of known and novel isoform counts with expected canonical amplicon length (**A**) and number of canonical exons (**B**). Linear regression of novel isoform read proportion with expected canonical amplicon length (**C**) and number of canonical exons (**D**).



1098 **Supplementary Figure 9. A.** Count of novel isoform splice pairing. Isoforms classified as containing
1099 at least one novel splice site (ALO) were examined and the novel pair or donor/acceptor was counted.

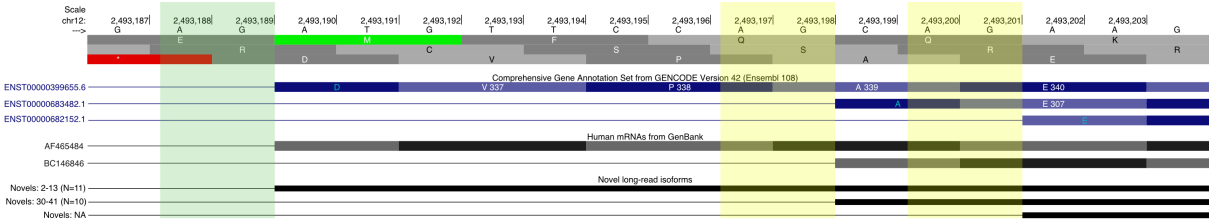
1100 **A**



1101

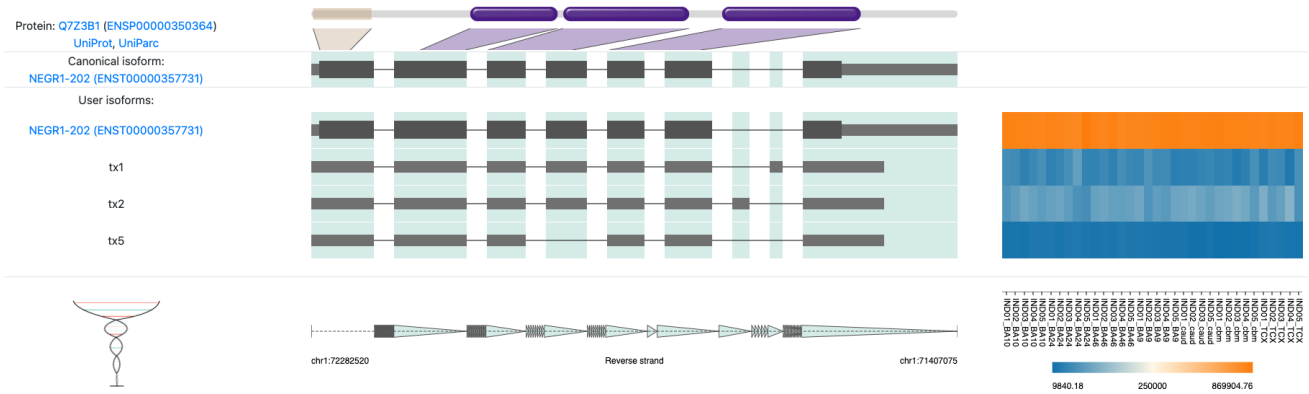
1102

Supplementary Figure 10. UCSC screenshot of *CACNA1C* splicing hotspot. Long-read sequencing identified 10 novel isoforms (black tracks) that support one of two annotated alternative splicing events (yellow boxes) within a 12 nt region (chr12:2,493,190-2,493,201) of exon 7 in *CACNA1C*. 11 novel isoforms also supported the use of the canonical (ENST00000399655.6) acceptor site (green box).

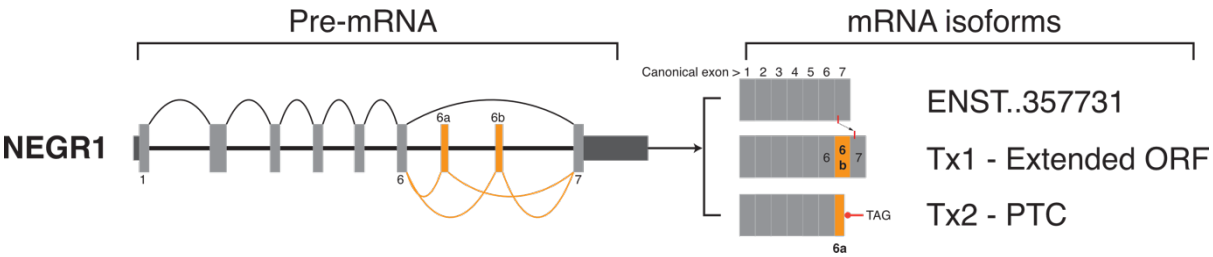


Supplementary Figure 11. *NEGR1* splice isoforms and protein prediction. A. The IsoVis schematic of *NEGR1* known and novel isoforms. Protein track indicates 5' signal peptide (brown) and three immunoglobulin (Ig)-like domains (red). **B.** NEGR1 pre-mRNA splice graph and predicted mRNA outcome. **C.** *NEGR1* novel exons were validated using Sanger sequencing of PCR amplicons and sequence reads were aligned and viewed using UCSC Genome Browser. Orange boxes indicate the novel exons and highlight high vertebrate conservation for exon 6b and predicted (NCBI RefSeq) termination site for exon 6a. **D.** AlphaFold protein prediction of the canonical *NEGR1* isoform (ENST00000357731). Three Ig-like domains are coloured according to the reported amino-acid positions (UniProt: Q7Z3B1); 1 (orange), 2 (blue) and 3 (pink). A GPI anchor residue (red) is shown at position 324 aa (C-terminal, red arrow). **E.** Overlaid AlphaFold protein predictions of novel Tx1 (purple) and Tx2 (orange). Ig-like domains are numbered 1-3, the GPI anchor (red) is present in Tx1 and the termination of Tx is indicated by a red arrow. Novel residues (pink) are indicated at the C-terminal end for both transcripts, Tx1: 14 aa and Tx2: 7 aa.

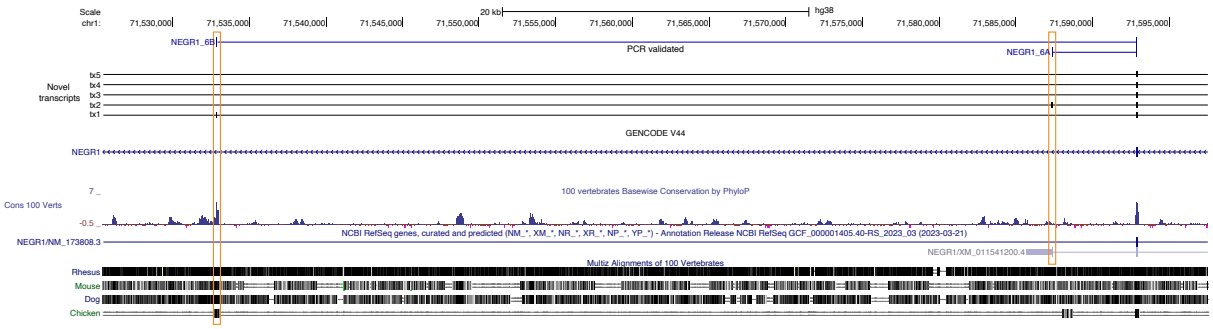
A



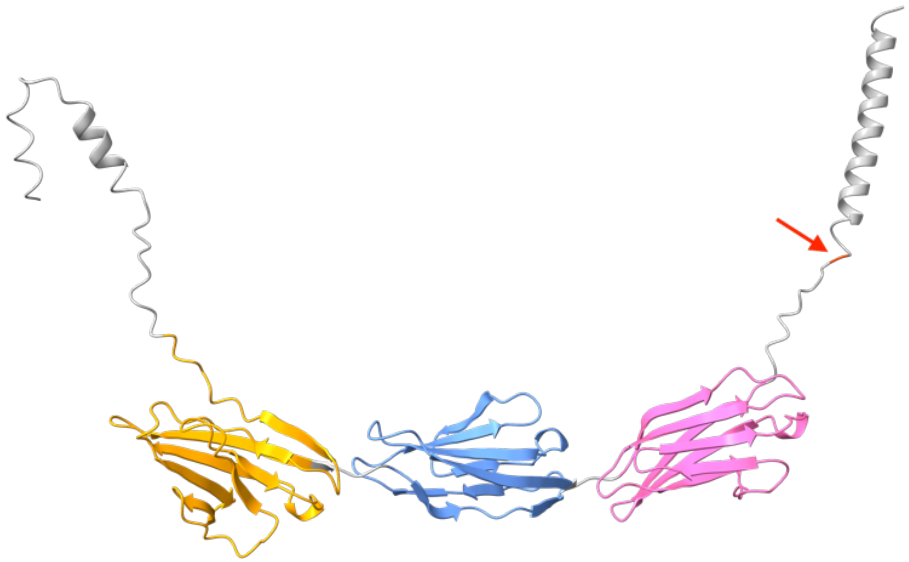
B



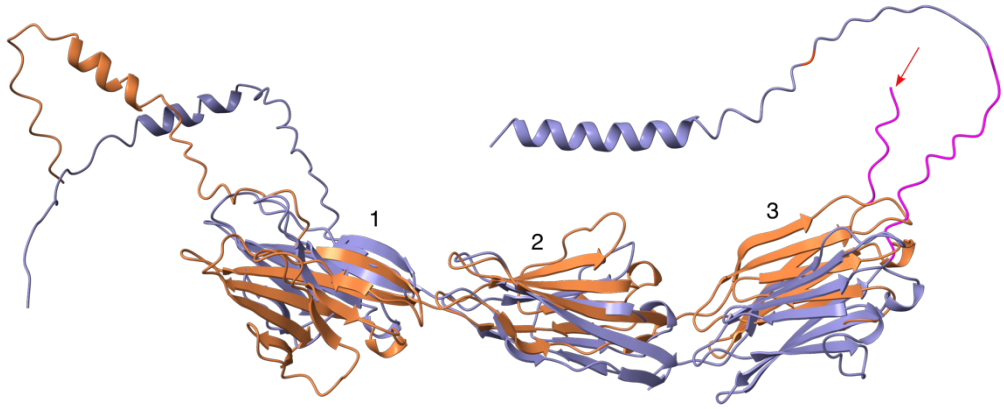
C



D

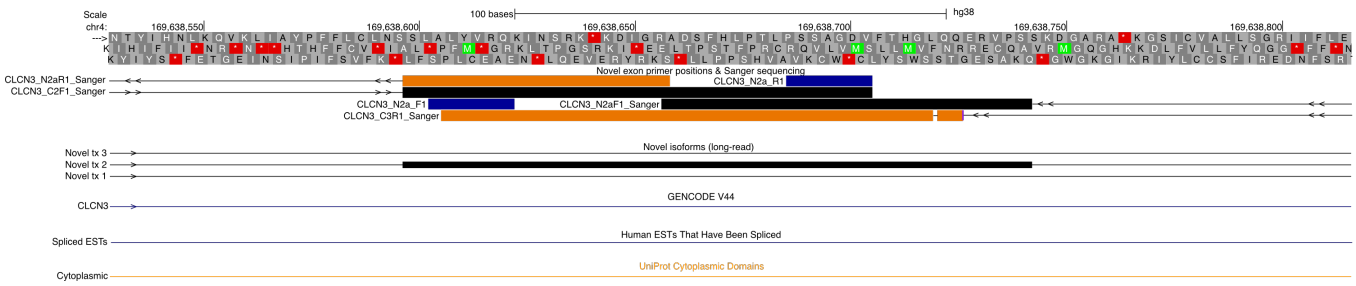


E

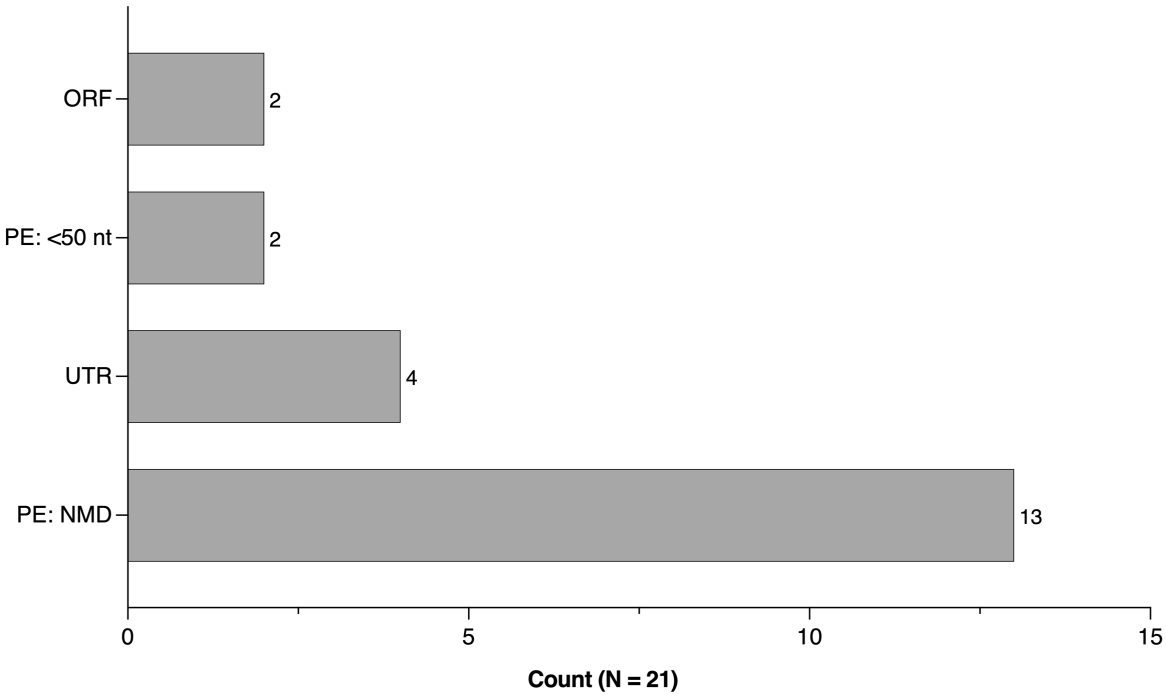


Supplementary Figure 12A. Novel exon validation in *CLCN3*. Novel exon 2a in the schizophrenia risk gene *CLCN3*, identified in long-read sequencing data, was validated by PCR using primers designed in the flanking cassette exons 1 and 3 (ENST00000513761). The novel exon sequence shown in novel transcript (Tx) 2 was validated with Sanger sequencing from the 5' canonical exon 2 (C2F1) to the reverse primer within the novel sequence (N2aR1) and from the novel sequence (N2aF1) to the 3' canonical exon 3 (C3R1). Direction of reads are indicated with arrows. Black and orange boxes indicate forward and reverse Sanger sequence respectively. Blue boxes indicate forward and reverse primers within the novel exon. A known cytoplasmic domain is shown by an orange track. Key: expressed sequence tag (EST). **B. Novel exon categories.** The impact of novel exon inclusion on the open reading frame of novel isoforms was predicted using Expasy [55] and then classed into groups, predicted to retain the open reading frame (ORF), inclusion of premature termination codon or 'poison exon' that was predicted to lead to nonsense-mediated decay (PE:NMD) or was <50 nt from the final exon junction (PE: <50 nt) or was with the 5' or 3' untranslated regions (UTR).

A

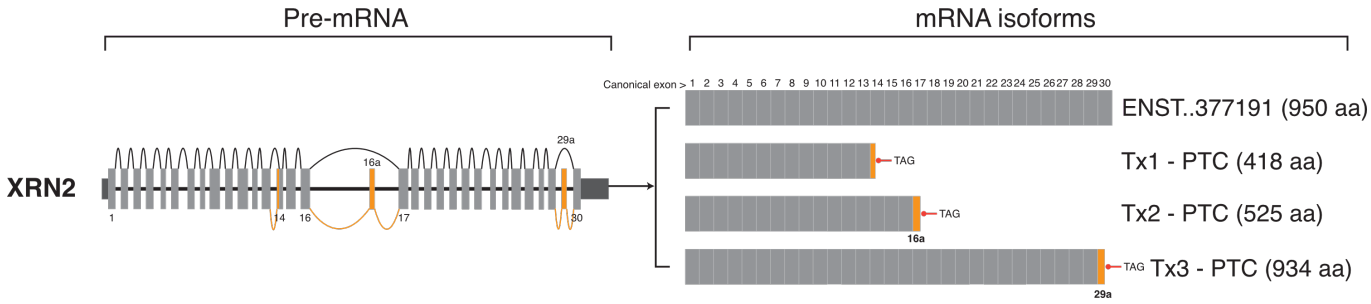


B



1156 **Supplementary Figure 13. Splice graph of *XRN2* novel isoforms containing novel exons.** Dark
1157 and light grey boxes indicate 5' and 3' UTR and coding exons respectively. Orange lines and boxes
1158 indicate novel splicing events and exons. Novel transcript 1 (Tx1) contains a novel splice acceptor
1159 (AG) within exon 14 (+24 nt) leading to a premature termination codon (PTC) and predicted nonsense
1160 mediated decay. Novel transcript 2 (Tx2) includes the validated novel exon 16a (54 nt) which was
1161 also predicted to encode a PTC. Novel isoform 3 (Tx3) contains a validated novel exon (29a) which
1162 encodes a PTC <50 nt from the final exon junction. Tx3 was predicted to lead to a truncated protein
1163 (934 aa). ".." indicates 0's removed for brevity.

1164



1165