UNIVERSITY OF CALIFORNIA SAN DIEGO

Isoform transcriptome of developing brain provides new insights into autism risk variants

A thesis submitted in partial satisfaction of the requirements
for the degree Master of Science

in

Biology

by

Kevin Khai Chau

Committee in charge:

Lilia M. Iakoucheva, Chair
Scott Rifkin, Co-Chair
Barry Grant

2019

Copyright
Kevin Khai Chau, 2019
All rights reserved.

The thesis of Kevin Khai Chau is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____
Co-Chair

_____
Chair

University of California San Diego

2019

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF SUPPLEMENTARY FIGURES

LIST OF SUPPLEMENTARY FILES

Supplementary Table 1:	chau_01_supplementary_table_1.xlsx, This table consists of relevant data for validation of BrainSpan TPM expression values against age-matched qPCR results.
Supplementary Table 2:	chau_02_supplementary_table_2.xlsx, Full gene-level differential expression results.
Supplementary Table 3:	chau_03_supplementary_table_3.xlsx, Full isoform-level differential expression results.
Supplementary Table 4:	chau_04_supplementary_table_4.xlsx
Supplementary Table 5:	chau_05_supplementary_table_5.xlsx, Gene ontology enrichment analysis results for gene-level differential expression analysis.
Supplementary Table 6:	chau_06_supplementary_table_6.xlsx, Gene ontology enrichment analysis results for isoform-level differential expression analysis.
Supplementary Table 7:	chau_07_supplementary_table_7.xlsx, Compiled and annotated variants table.
Supplementary Table 8:	chau_08_supplementary_table_8.xlsx, Gene co-expression module assignments.
Supplementary Table 9:	chau_09_supplementary_table_9.xlsx, Isoform co-expression module assignments.
Supplementary Table 10:	chau_10_supplementary_table_10.xlsx, Gene ontology enrichment analysis of gene co-expression modules.
Supplementary Table 11:	chau_11_supplementary_table_11.xlsx, Gene ontology enrichment analysis of isoform co-expression modules
Supplementary Table 12:	chau_12_supplementary_table_12.xlsx, Module impact rate analysis results

ACKNOWLEDGEMENTS

I would like to acknowledge Professor Lilia M. Iakoucheva for her support as the chair of my committee.

I would also like to thank my post-doctoral mentor Dr. Pan Zhang for his guidance throughout this project, as well as Dr. Patricia M. Losada, Dr. Akula Bala Pramod, Dr. Jorge Urresti, and Dr. Megha Amar for all of their support.

This thesis, in full, is currently being prepared for submission for publication of the material. Chau, Kevin K.; Zhang, Pan; Urresti, Jorge; Amar, Megha; Pramod, Akula Bala; Corominas, Roser; Lin, Guan Ning; Iakoucheva, Lilia M. The thesis author was the primary investigator and author of this material.

ABSTRACT OF THE THESIS

Isoform transcriptome of developing brain provides new insights into autism risk variants

by

Kevin Khai Chau

Master of Science in Biology

University of California San Diego, 2019

Professor Lilia M. Iakoucheva, Chair
Professor Scott Rifkin, Co-Chair

Alternative splicing plays important role in brain development, however its global contribution to human neurodevelopmental diseases (NDD) has not been fully investigated. Here, we examined the relationship between splicing isoform expression and *de novo* loss-of-function mutations implicated in autism. We constructed isoform transcriptome of the developing human brain, and observed better resolution and stronger signals at the isoform-level compared to the gene-level transcriptome. We identified differentially expressed isoforms and isoform co-expression modules enriched in autism loss-of-function mutations. These isoforms have higher prenatal expression, are enriched in microexons, and are co-expressed with a unique

set of partners. We experimentally test the impact of splice site mutations in five NDD risk genes, including *SCN2A*, *DYRK1A* and *BTRC*, and demonstrate exon skipping. Furthermore, our results suggest that the splice site mutation in *BTRC* reduces its translational efficiency, likely impacting Wnt signaling through impaired degradation of β-catenin. We propose that functional effect of mutations associated with human diseases should be investigated at isoform- rather than gene-level resolution.

1 Introduction

More than 95% of multi-exon human genes undergo alternative splicing (AS) and/or use alternative promoters to increase transcriptomic and proteomic diversity, with an estimated average of five to seven isoforms transcribed per gene [**?**, **?**, **?**]. Alternative splicing is highly specific, and expression of isoforms is often restricted to certain organs, tissues or cell types [**?**, **?**, **?**, **?**]. In addition, many isoforms are expressed only during specific developmental periods [**?**]. The developing human brain exhibits one of the highest frequencies of alternative splicing events [**?**, **?**, **?**, **?**]. Many of the processes occurring during neural development are controlled by differentially expressed alternatively spliced isoforms [**?**, **?**, **?**]. Several recent studies, including one by us, began to investigate isoform-level transcriptome dysregulation in psychiatric diseases [**?**, **?**, **?**]. However, spatiotemporal analyses of the full-length isoform transcriptome of the developing human brain remains relatively unexplored.

Integration of brain spatiotemporal transcriptome with the genetic data from exome and whole genome sequencing studies have provided important insights into neurodevelopmental diseases (NDDs) [**?**, **?**, **?**, **?**, **?**]. Most of the recent work in this area focused on understanding the effect of mutations at the gene-level resolution, whereas isoform-specific impact of loss-of-function (LoF) mutations in the context of brain development has not yet been investigated.

It is important to map LoFs to transcripts because protein isoforms, encoded by different transcripts, have drastically different protein interaction capabilities. As we have previously demonstrated, the majority of the isoforms encoded by the same gene share less than a half of

their interacting partners in the human interactome network [**?**]. This observation points to striking functional differences between splicing isoforms that are not accounted for by the majority of the existing gene-level studies. In addition, our recent work demonstrated that isoform-level networks provide better resolution and depth around disease candidate proteins compared to the gene-level networks [**?**].

To better understand how NDD risk mutations dysregulate neurodevelopment, we construct isoform transcriptome of the developing human brain using BrainSpan RNA-seq dataset [**?**] summarized to isoforms [**?**]. We perform isoform differential expression and co-expression analyses to identify differentially expressed isoforms (DEI) and co-expressed isoform modules (CIM) in the adjacent brain developmental periods starting from fetal to adult. When compared to gene-level analyses, isoform transcriptome provides more meaningful insights and paints a more complete picture of neurodevelopmental processes. Importantly, many DEIs and CIMs identified by our analyses are not detectable at the gene-level resolution. Mapping NDD risk mutations to DEI revealed that LoF-impacted transcripts have higher prenatal expression, more frequently carry microexons, and are preferentially involved in neuronal processes compared to non-impacted transcripts. Furthermore, isoform co-expression modules with splicing-related and synaptic functions are enriched in LoF-impacted transcripts implicating these functions in NDDs. Finally, we experimentally test the impact of several splice site LoFs and demonstrate that they cause exon skipping to produce novel isoforms with altered biological properties. Our study makes a strong case for investigation of disease mutations at isoform- rather than gene-level resolution.

2 Results

Construction and validation of isoform transcriptome of the developing human brain

To investigate global patterns of isoform expression across brain development, we built a temporal isoform transcriptome of the developing brain (Supplementary Figure 1). We used the BrainSpan dataset [**?**] (http://www.brainspan.org/) summarized to transcripts as previously described [**?**]. After rigorous quality control (Supplementary Figures 2-5, Materials and Methods), we obtained expression profiles for 100,754 unique isoforms corresponding to 26,307 brain-expressed human genes (3.8 isoforms/gene).

We experimentally validated the quality of isoform expression values for a subset of selected isoforms. We used quantitative PCR (qPCR) to estimate relative expression difference of 26 unique isoforms of 14 genes between two independent RNA samples that were age, sex and brain region-matched to the samples from the BrainSpan. The relative qPCR isoform expression values in the independent samples of frontal lobe of 22 weeks old fetus and cerebral cortex of 27 years old adult were compared to the values computationally assigned by us using BrainSpan. We observed positive correlation (R = 0.26) between experimental and computational values for these isoforms, despite using different samples for validation (Supplementary Table 1, Supplementary Figure 6, Materials and Methods).

Differential isoform expression reveals distinct signals relative to differential gene expression

We recently demonstrated that isoform-level changes capture larger disease effects than gene-level changes in the context of three major psychiatric disorders [**?**]. Here, we investigated the role of isoform expression in the context of the developing brain. We performed differential expression analysis among all pairs of adjacent developmental periods as well as between pooled prenatal (P02-P07) and pooled postnatal (P08-P13) (PrePost) samples using filtered gene-level and isoform-level data, yielding sets of differentially expressed genes (DEG) and differentially expressed isoforms (DEI) (Materials and Methods, Supplementary Tables 2 and 3). The P06/P07 (late mid-fetal/late fetal) and P07/P08 (late fetal/neonatal) developmental periods accumulated largest number of both DEGs and DEIs, supporting critical brain remodeling right before and after the birth (Fig. 1A). In P06/P07 8.3% of genes and 20.3% of isoforms are differentially expressed, whereas in P07/P08 13.2% of genes and 20.4% of isoforms are differentially expressed (Supplementary Table 4). Overall, 48.4% of genes and 64.9% of isoforms are differentially expressed between prenatal and postnatal (PrePost) periods. These results suggest a greater degree of transcriptomic remodeling during prenatal life, consistent with faster brain development in utero as compared to postnatal life.

In addition to the greater fraction of DEI among adjacent and PrePost periods, we also observed significantly increased effect sizes (absolute $\log_2$ fold changes) among DEI as compared to DEG, both overall and in nearly every developmental period (Fig. 1B). This suggests that levels of differential expression are more pronounced at the isoform-level relative to the gene-level. Thus, the isoform-level transcriptome is likely to provide additional information that is missed by the gene-level transcriptome.

To better understand whether isoform-level data is capturing additional information as compared to the gene-level data, we performed cell type and curated gene list enrichment analyses

of unique non-overlapping DEGs and DEIs (lightly shaded subsets from Fig. 1A) (Fig. 1C). We used published single-cell sequencing data (for Cell Type) along with NDD-related gene lists to detect enrichment in each period and in the PrePost dataset (Materials and Methods). Overall, DEGs are clearly capturing weaker enrichment signals than DEIs, potentially due to bigger dataset sizes. Among cell types, DEIs are significantly enriched in excitatory neuron markers, especially in the prenatal to early childhood developmental periods (Fisher-exact test, max Bonferroni-adjusted P $<$ 1E-09, OR = 2.39 – 3.29, min. 95% CI = 2.07, max 95% CI = 3.98 for P02/P03-P09/P10) (Fig. 1C, left panel). The DEIs from almost all periods are enriched in postsynaptically expressed genes, as well as FMRP and CHD8 targets, with most significant enrichment during P06/P07 (late mid-fetal/late fetal). Interestingly, the DEIs from only P04/P05 (early mid-fetal) are enriched in autism risk genes from a recent dataset [**?**] (Fisher-exact test, Bonferroni-adjusted P = 0.005, OR = 3.88, 95% CI = 2.11 – 3.68) (Fig. 1C, right panel), and this signal is not captured at the gene-level. Mid-to-late fetal developmental period was previously identified as critical to ASD pathogenesis [**?**, **?**].

Functional Gene Ontology (GO) enrichment analyses for unique DEGs and DEIs in P04/P05, P07/P08 and P08/P09 demonstrate more neurodevelopmentally-relevant processes with DEIs vs DEGs (Fig. 1D, Supplementary Tables 5 and 6). For example, "neuron projection development", "brain development", and "nervous system development" are enriched in DEI, but not in DEGs. This suggests that the isoform transcriptome may provide better insights into brain development.

Differentially expressed isoforms impacted by autism loss-of-function mutations have higher prenatal expression

To improve understanding of the impact of NDD mutations on brain development, we mapped rare de novo loss-of-function (LoF) variants identified in the largest autism spectrum

disorder (ASD) exome sequencing study [**?**] to the isoform transcriptome. A total of 12,111 ASD case variants and 3,588 control variants were processed through Ensembl's Variant Effect Predictor (VEP) and filtered for consequences likely to result in the loss-of-function of the impacted gene or isoform (Materials and Methods, Supplementary Table 7). In total, 1,132 ASD case and 262 control variants fit this criterion, impacting 4,050 isoforms from 1189 genes. At the isoform level, 3,128 isoforms were impacted by ASD case variants (ASD LoF), 848 isoforms by control variants (Control LoF), and 74 isoforms by both. We also defined a dataset of isoforms that were not impacted by ASD variants (Non-impacted by ASD LoF) as a control.

In every prenatal developmental period, as well as in the pooled prenatal sample, the expression of the ASD LoF isoforms was found to be significantly higher than Control LoF isoforms or Non-impacted by ASD LoF isoforms (Mann-Whitney test, BH-adjusted P-value $\leq$ 0.05) (Fig. 2A). This suggests that the potential loss of expression of these highly expressed isoforms in the normal prenatal human brain as a result of LoF mutations may contribute to ASD pathogenesis.

We then selected genes with differentially expressed isoforms, for which at least one isoform is ASD LoF, at least one other isoform is non-impacted by ASD LoF, and were not differentially expressed at the gene level; 26 genes out of 102 Satterstrom genes satisfied this criterion (Fig. 2B). Hierarchical clustering of the isoforms from these genes based on expression values identified a prenatally expressed cluster consisting largely of the ASD LoF impacted isoforms (Fig. 2C). These isoforms are also significantly enriched in microexons (i.e. short exons of 3-27bp in length) (Permutation test, n = 1000 permutations, P = 0.04) (Fig. 2D), in agreement with previous observations about their role in autism [**?**, **?**]. The impacted and non-impacted isoforms of some genes (*KMT2C*, *MBD5*, and *PTK7*) have opposite developmental trajectories, whereas for other genes (*GABRB3*) the impacted isoforms are highly expressed throughout brain development (Fig. 2E). It is likely that LoF mutation that impacts highly prenatally expressed isoform can severely disrupt early brain development and lead to NDD.

Overall, isoform transcriptome analyses could provide a more detailed picture of the functional impact of NDD risk mutations.

Isoform co-expression modules capture trajectories of brain development

To better understand how brain development is regulated at the transcriptional level, we carried out Weighted Gene Co-expression Network Analyses (*WGCNA*) [**?**] (Materials and Methods). Co-expression modules were defined as clusters of genes or isoforms with highly correlated expression profiles across all BrainSpan samples. We identified a total of 8 gene modules and 55 isoform modules, with one additional grey module in each network (Supplementary Tables 8 and 9).

Hierarchical clustering of the modules by their eigengenes demonstrates that each gene co-expression module closely clusters with a corresponding isoform co-expression module (Fig. 3A). Further characterization of these gene/isoform module pairs via GO annotations shows overlapping functions and pathways (Supplementary Tables 10 and 11). For example, gene module gM2 and isoform module iM2 are both enriched for GO terms related to synaptic transmission. This indicates that the isoform co-expression network recapitulates the gene co-expression network.

In order to relate each co-expression module with brain developmental periods, we calculated module-period associations using linear mixed effects models (Materials and Methods). We found modules that are significantly associated with several developmental periods (Fig. 3A, top panel); iM1 is significantly associated with prenatal periods P02 (FDR-adjusted P = 0.009), P03 (FDR-adjusted P = 0.003), and P04 (FDR-adjusted P = 0.008), and with iM10 (FDR-adjusted P = 6.59E-04) and iM39 (FDR-adjusted P = 0.026). Functional GO analyses of these modules demonstrates that iM1 is enriched in splicing functions, iM10 in cell cycle-related processes,

and iM39 is enriched in embryonic development; all functions are related to early fetal brain development.

There are also several modules (gM4, iM35, iM7, and iM38) strongly associated with the late fetal period P07, and these modules cluster together (gM4: FDR-adjusted P = 1.78E-09; iM35: FDR-adjusted P = 8.23E-04; iM7: FDR-adjusted P = 3.83E-04; iM38: FDR-adjusted P = 0.009). Collectively, these modules are enriched for angiogenesis and extracellular matrix organization GO functions (Supplementary Table 11).

Analysis of cell type enrichment shows modules that are significantly enriched in specific cell types (Fig. 3A, middle panel). For example, iM10 that is associated with very early P02 period, is also enriched in neuroprogenitors (NPCs), the cells that give rise to other neuronal cell populations and are often found very early in brain development. Likewise, iM2 is primarily associated with postnatal periods, and is strongly enriched in excitatory neurons, which represent mature neuronal population. Interestingly, the cluster of modules that is strongly associated with late fetal P07 period (gM4, iM35, iM7, and iM38), is enriched in microglia, or innate immune cells of the brain, that peak around late mid-fetal to late fetal development. Furthermore, isoform module eigengene trajectories are capturing the appropriate signals from each cell type, with NPC steadily decreasing and neuronal cell types increasing from prenatal to postnatal brain development (Fig. 3B).

Enrichment analysis co-expression modules for curated gene lists identified modules gM1 and iM1 as enriched in ASD risk genes, CHD8 targets, and functionally constrained and mutation intolerant (pLI $>$ 0.99) genes (Fig. 3A, bottom panel). The same modules are significantly associated with prenatal periods, and are enriched in RNA processing and splicing GO functions (Fig. 3D, upper panel). Another module that is enriched in ASD risk genes is iM19, and it is annotated with chromatin and histone-related GO functions. This is consistent with previous observations about chromatin modifier genes enrichment among ASD risk genes [**?**]. In summary, the analyses of isoform co-expression modules further broaden our knowledge of the developing

human brain at the transcriptome level.

LoF-impacted co-expression modules point to dysregulation of RNA splicing and synaptic organization

We next calculated impact rates of rare de novo ASD variants from cases and controls [**?**], and identified co-expression modules that are significantly more impacted by LoF case mutations relative to control mutations (Supplementary Table 12) (Materials and Methods). We observed three modules that are significantly impacted by case ASD variants, one gene module (gM1) and two isoform modules (iM1 and iM30) (Fig. 3C). Unsurprisingly, gM1 and iM1 cluster together and are enriched in similar GO functions that are related to RNA processing and splicing, including non-coding RNA splicing (Fig. 3D, upper panel). This agrees with the already-demonstrated crucial role of splicing dysregulation in ASD [**?**, **?**]. Functional enrichment of isoform co-expression module iM30 points to dysregulation of synapse organization and neuronal projection pathways (Fig. 3D, bottom panel), pathways which are strongly implicated in ASD [**?**, **?**]. Thus, isoform modules reflect processes previously implicated in ASD, and point to specific isoforms (rather than genes) that can contribute to this dysregulation.

To demonstrate how isoform co-expression modules could be useful for future studies, we built isoform co-expression protein-protein interaction (PPI) networks from the gM1 and iM1 modules (Fig. 3E). The network is focused on ASD risk genes that have at least one isoform impacted by LoF mutation, and the edges that have gene-level PPI support (due to scarcity of isoform-level PPIs) are filtered for the top 10% of connections by Pearson correlation coefficient (PCC) (Materials and Methods). Clearly, gM1 has fewer connections than iM1, and iM1 highlights some interesting isoform co-expressed PPIs that are not discernable from gene-level co-expression network. For example, 9 genes from this module (*ARID1B*, *CHD8*, *KMD5B—*, *KMT2A*, *MED13L*, *PCM1*, *PHF12*, *POGZ*, and *TCF4*) have at least one LoF impacted isoform and at least one

that is not impacted by mutation. These isoforms could also be co-expressed and interact with different partner isoforms. For example, one LoF-impacted isoform of the *KMT2A* gene (KMT2A-017) interacts with CREBBP-001 and CREBBP-003 whereas non-impacted KMT2A-014 has completely different partners (LEO1-001, SIN3A-002 and CDCZ3-001). This leads to different networks being disrupted as a result of *KMT2A* mutations, and these networks could not be discerned from gene-level information. Another interesting observation from co-expressed PPI networks is that LoF impacted isoforms tend to have higher PCC with the corresponding partners than non-impacted isoform (Mann-Whitney test, P = 1.53E-05), suggesting potentially greater impact on networks. Thus, we demonstrate the utility of isoform networks for investigating functional impact of mutations.

De novo splice site mutations of NDD risk genes cause exon skipping, partial intron retention, or have no effect on isoforms

One type of LoF mutations are mutations that affect splice sites directly. Here, we tested the effect of de novo splice site mutations identified in exome sequencing studies in four NDD risk genes (*DYRK1A*, *SCN2A*, *DLG2*, and *CELF2*) to better understand their functional impact. All highly prenatally expressed isoforms of these genes are found in iM1. We used exon trapping assay (Materials and Methods) to investigate the following de novo splice site mutations identified from exome sequencing studies of NDDs: *SCN2A* (chr2:166187838, A:G, acceptor site) [**?**]; *DYRK1A* (chr21: 38865466, G:A, donor site) [**?**]; DLG2 (chr11: 83194295, G:A, donor site) [**?**]; and *CELF2* (chr10: 11356223, T:C, donor site) [**?**]. Mutation in *SCN2A* causes out-of-frame exon skipping and potential inclusion of 30 new amino acids into the translated protein before ending with a premature stop codon, that most likely will result in nonsense-mediated decay (NMD) (Fig. 4A). In contrast, mutation in *DYRK1A* causes an in-frame exon skipping, potentially producing a different variant of the same protein and thus is expected to have milder functional effect

(Fig. 4B). In the case of *DLG2*, mutation affects a splice site adjacent to the exon five, which is alternatively spliced in the WT isoforms (Fig. 4C). We constructed a minigene that includes exon five together with the preceding exon four and observed that exon five is constitutively spliced out from our construct independently on the presence of mutation. However, the mutation caused partial (i.e. 65bp) intron inclusion downstream from exon four. At the translational level, this mutation would likely result in a truncated protein one residue after the end of exon 4 due to a premature stop codon. Finally, *CELF2* mutation affects an alternative splice site, which also maps to an exonic region of another alternatively spliced isoform. When cloned into the exon trapping vector, the transcript generated from the WT minigene included the isoform carrying longer exon with mutation (Fig. 4D). Thus, after introducing the mutation, no difference between WT and mutant constructs was observed. This is not surprising given the fact that the splice site mutation behaves like exonic missense mutation in the isoform predominantly expressed from our construct. These results suggest that mutations could impact different isoforms of the same gene by different mechanisms, i.e. splice site mutation in one isoform could represent a missense mutation in another isoform.

Further analysis of expression profiles of the brain-expressed isoforms transcribed by these genes (Fig. 4E) suggest that highly prenatally expressed isoforms (SCN2A-201, DYRK1A-001, DLG2-016 and CELF2-201) are most likely targets for the "pathogenic" effect of mutations. In summary, our experiments showcase different scenarios of the impact of splice site mutations and confirms the need to investigate their functional impact at the isoform- rather than the gene-level resolution.

The splice site mutation in *BTRC* reduces its translational efficiency

We next decided to investigate the pathways that specific isoform mutations can disrupt. For this, we selected the isoforms of a newly identified ASD risk gene, *BTRC* (also known as

β-*TrCP* or *FBXW1A*) [**?**], based on their availability from our previous study [**?**]. We used three full-length isoforms (BTRC-001, BTRC-002, and BTRC-003) for this study (Fig. 5A). The mutation in *BTRC* (chr10: 103221816, G:A, donor site), detected in the patient with ASD [**?**], causes in-frame exon four (78bp) skipping in the exon trapping assay (Fig. 5B). To further test the effect of this mutation on different *BTRC* transcripts, we generated additional constructs by inserting abridged introns surrounding exon 4 into the coding sequence (CDS) of two isoforms, BTRC-001 and BTRC-002 (Fig. 5C, Materials and Methods). The third isoform, BTRC-003, does not carry exon 4, and its structure and size are identical to the BTRC-001, when exon 4 is skipped. We also generated mutant constructs BTRC-001Mut and BTRC-002Mut carrying the mutation in the abridged intron (Fig. 5A). The RT-PCR following exon trapping assays on the full-length CDS, as well as WT and mutant constructs with abridged introns, confirmed the correct sizes of all constructs, and validated exon skipping event due to splice site mutation (Fig. 5C). Furthermore, Western blot confirmed the expected sizes of the protein products produced from the WT and mutant constructs (Fig. 5D). The splice site mutation significantly reduces the amount of the protein produced from mutant transcripts, suggesting their decreased translational efficiency (Fig. 5E). Higher amount of protein product produced from all constructs with abridged introns compared to CDSs is consistent with previous observations of increased translational efficiency of RNAs produced by splicing compared to their intron-less counterparts [**?**]. Further, BTRC-001 and BTRC-002 are highly expressed relative to the non-impacted BTRC-003 (and other *BTRC* isoforms), indicating that these two isoforms are likely key players in neurodevelopment (Fig. 5F).

Next, we investigated binding properties of all isoforms using co-immunoprecipitation (co-IP) (Fig. 5D). The *BTRC* gene encodes a protein of the F-box family and is a component of the SCF (Skp1-Cul1-F-box protein) E3 ubiquitin-protein ligase complex. One of the well-known substrates of this complex is β-catenin (*CTNNB1*). The SCF complex ubiquitinates and regulates degradation of β-catenin, an essential component of the *Wnt* signaling pathway [**?**]. *Wnt*

plays key roles in cell patterning, proliferation, polarity and differentiation during the embryonic development of the nervous system [**?**] and have been consistently implicated in ASD [**?**, **?**]. Both β-catenin and *Cul1* carry de novo mutations identified in patients with NDD [**?**].

The interaction of *BTRC* with its partners, Cul1, Skp1 and β-catenin, demonstrates reduced binding with mutant *BTRC* (Fig. 5D-E). In agreement with previous observations, we found that *BTRC* only binds to the phosphorylated form of β-catenin [**?**]. This suggests that the number of complexes is strongly dependent on the availability of *BTRC* protein, which is significantly reduced due to splice site mutation. Thus, our results indicate that the *BTRC* splice site mutation causes exon skipping in *BTRC* isoforms and reduces translational efficiency of the resulting protein product. This, in turn, decreases the amount of ligase complexes that are available for β-catenin ubiquitination. We hypothesize that this may lead to impaired degradation of β-catenin, its cellular accumulation and upregulation of *Wnt* signaling as a result of this ASD risk mutation. Further studies are needed to test this hypothesis.

3 Discussion

One of the greatest bottlenecks in the analysis of large-scale whole genome sequencing and whole exome sequencing brain development data, and its integration into knowledge regarding the pathogenesis of complex neurodevelopmental disorders such as ASD, is our lack of understanding of the transcriptional and translational mechanisms governing brain development. Given such a high amount of alternative splicing events and regulation thereof, knowledge and understanding of these underlying molecular mechanisms is paramount to the analysis of complicated neurodevelopmental disorders, which would pave the way for better therapeutics of those at risk [**?**, **?**].

Previous studies demonstrated that the integration of genetic data with isoform-level co-expression and/or protein interaction networks is able to capture molecular mechanisms of disease that could not be inferred from the gene-level analyses [**?**, **?**]. The importance of the isoform-level networks is further emphasized by our recent observation that protein products encoded by different splicing isoforms of the same gene share less than half of their interacting partners [**?**]. Thus, gene-level resolution is not sufficient to precisely map molecular networks dysregulated by genetic mutations in human diseases. Here, we analyzed the isoform-level transcriptome of the developing human brain to gain insights into neurodevelopmental disorders like ASD.

We first observed distinct forms of dysregulation across neurodevelopment when considering genomic features quantified either at the gene level or the isoform level (Fig. 1A). Isoform-level

quantification and differential expression results that did not overlap with gene-level analyses revealed distinct neurodevelopmental pathways, such as neuron projection development and more generally central nervous system development, that were not discernable at the gene resolution. This indicates that we were able to identify distinct groups of alternatively spliced isoforms that are likely to be essential to neurodevelopment through isoform-level resolution, and thus solely gene-level resolution analysis is not sufficient to fully capture the temporal dynamics of neurodevelopment (Fig. 1D).

Application of this concept of higher resolution isoform-level analysis relative to gene analysis allowed us to discern distinct alternatively spliced protein-coding isoforms of ASD risk genes that are a) significantly more highly expressed prenatally (Fig. 2A) and b) have distinct expression profiles relative to their sibling protein-coding isoforms (Fig. 2D). *PTK7* is one example of one such ASD risk gene with isoforms differentially impacted by rare de novo ASD variants. *PTK7* is a tyrosine-protein kinase and is involved in both the non-canonical and canonical *Wnt* signaling pathways and is has a role in embryogenesis and angiogenesis [**?**, **?**]. Two of the protein-coding isoforms transcribed by *PTK7* were found to "switch" expression profiles; one impacted isoform is relatively more highly expressed prenatally and decreases in expression across time, while another unimpacted isoform is lowly expressed prenatally and increases in expression and peaks in developmental period 10 (early childhood). This switch in expression between differentially impacted sibling isoforms, implies the relative importance of each isoform in the temporal context. Since *PTK7* is previously described as a player in the *Wnt* signaling pathway and embryogenesis, the impacted isoform that is highly expressed prenatally is likely the more relevant isoform in embryonic development and should be a focus of further study as compared to the other alternatively spliced protein-coding isoforms.

We find similar patterns in *KMT2C* and *MBD5*. Both of these high-risk ASD genes exhibit similar isoform patterns as PTK7, with higher relative expression in prenatal developmental periods and lower expression later. The dynamically expressed isoforms of these three genes

are expressed in radically different patterns as compared to the other 23 ASD genes identified as differentially impacted distinct DEI; therefore, it is likely that these genes (and these specific isoforms) are important specifically in prenatal neurodevelopment.

Through tandem co-expression network analysis utilizing both levels of quantification, we found the isoform-level networks essentially recapitulate the information provided by the gene-level co-expression modules given evidence by module eigengene hierarchical clustering and gene list enrichment analysis (Fig. 3A). Module-trait associations provide insights as to the biological pathways represented in each developmental period (Fig. 3A, middle). By enriching the co-expression modules for the biological processes and associating these modules with each developmental period, we can indirectly link those biological processes to each time point. Modules gM4, iM35, iM7, and iM38 most closely associated with P7 (late fetal to neonatal ages). Module gM4 was heavily characterized as a circulatory system development and blood vessel development. Naturally, its isoform module partner, iM35, was also enriched for terms related to angiogenesis. The next most closely clustered module, iM7, was heavily enriched for synaptic signaling related processes, and iM38, the next clustered module, was enriched for extracellular matrix organization. Furthermore, these modules are enriched in markers for endothelial cells and glial cells (Fig. 3C, bottom), indicating that these modules are very likely to be representative of the previously studied development of the blood-brain barrier and neuronal and glial/endothelial interactions during late-fetal and neonatal developmental periods [**?**, **?**, **?**, **?**, **?**, **?**].

Further, impact analysis revealed modules predicted to be significant in the context of autism spectrum disorder: gM1, iM1, and iM30. We found that these significantly impacted modules were enriched for features related to splicing and synapse structure, which is to be expected given the brain tissue context and impact by rare de novo ASD variants (Fig. 4C, Fig. 4D). Even further, by overlaying gene-level PPIs with the gM1 and iM1 co-expression networks, we identify several ASD risk genes transcribing differentially impacted isoforms. These isoforms exhibited greater co-expression with the gene-level protein interacting partners

relative to unaffected sibling isoforms, highlighting the relative importance of the specifically impacted alternatively spliced isoforms. Thus, by combining gene-level and isoform-level expression results we find specific signal for this disorder context, and more so than just using a gene-level analysis.

The investigation of the impact of the *de novo* splice site mutations in four genes important in neurodevelopment demonstrated exon skipping and disruption of normal splicing pattern. However, a more detailed analysis at the isoform-level suggested that not all isoforms could be affected by mutations. For example, at least one known isoform of *BTRC* does not carry an exon with the *de novo* mutation, and therefore is not expected to be impacted by this mutation (Fig. 5A, Fig. 5C). We also demonstrate that *BTRC* mutation decreases translational efficiency of the impacted isoforms, since lower amount of the resulting protein is observed (Fig. 5D). This, in turn, leads to reduced interaction between BTRC and its protein partners, potentially disrupting *Wnt* signaling (Fig. 5E). Since β-catenin is a substrate of the BTRC-Cul1-Skp1 ubiquitin ligase complex, the shortage of this complex may lead to impaired ubiquitination and degradation of β-catenin and its neuronal accumulation. Interestingly, transgenic mice overexpressing β-catenin have enlarged forebrains, arrest of neuronal migration and dramatic disorganization of the layering of the cerebral cortex [**?**]. It would be interesting to investigate whether the patient carrying the *de novo BTRC* splice site mutation has similar brain abnormalities.

Typically, mutations affecting essential splice sites are automatically classified as loss of function mutations when considering gene-level analyses. Here, we demonstrate that this is not always the case, and that splice site mutation affecting one isoform of the gene may serve as a missense mutation in another isoform that carries a longer exon spanning the splice site, like in the case of *CELF2* (Fig. 4D). Thus, depending on where, when, at what level and which isoform of the gene is expressed, the functional impact of the same mutation may differ dramatically. In addition, the mutation could also be "silent" if the isoform is highly expressed but does not carry an exon affected by a specific mutation. This suggests that the impact of mutations should be

investigated at the isoform-level rather than the gene-level resolution, and expression levels of splicing isoforms in disease-relevant tissues should be taken into consideration to better guide hypotheses regarding potential mechanisms of the disease and its future treatments.

Typically, mutations affecting essential splice sites are automatically classified as loss of function mutations when considering gene-level analyses. Here, we demonstrate that this is not always the case, and that splice site mutation affecting one isoform of the gene may serve as a missense mutation in another isoform that carries a longer exon spanning the splice site, like in the case of *CELF2* (Fig. 4D). Thus, depending on where, when, at what level and which isoform of the gene is expressed, the functional impact of the same mutation may differ dramatically. In addition, the mutation could also be "silent" if the isoform is highly expressed but does not carry an exon affected by a specific mutation. This suggests that the impact of mutations should be investigated at the isoform-level rather than the gene-level resolution, and expression levels of splicing isoforms in disease-relevant tissues should be taken into consideration to better guide hypotheses regarding potential mechanisms of the disease and its future treatments.

4 Materials and Methods

Throughout this analysis, R version 3.6.0 was used. Downstream bioinformatics analysis is outlined in Supplementary Figure 1A.

Preprocessing of RNA-Seq data

RNA-Seq quantification data at both gene and isoform levels was obtained from the BrainSpan Atlas of the Developing Human Brain [**?**, **?**, **?**] (Supplementary Figure 1B). This data was sequenced from post-mortem brain tissue from 57 donors aged between 8 weeks post-conception to 40 years, across a number of different brain regions, for a total of 606 initial samples. This data was processed as previously described [**?**]; to summarize, data was aligned with STAR (v2.4.2a) [**?**] and quantified RSEM (v1.2.29) [**?**] using the GRCH37.p13 human reference genome, resulting in counts and TPM matrices for both datasets. These matrices were filtered for features with TPM ¿= 0.1 in at least 25% of samples in both datasets.

Sample connectivity analysis was performed to detect sample outliers as previously described [**?**]. In brief, biweight midcorrelation was calculated among sample expression vectors in both the expression-filtered gene-level and expression-filtered isoform-level datasets. These values were converted into connectivity Z-scores. 55 samples were identified as having sample connectivity Z-scores ¡= -2 and were removed from downstream analysis, resulting in 551 final samples.

Surrogate variable analysis was performed to remove latent batch effects in the data, taking into consideration age, brain region, sex, ethnicity, and study site [**?**, **?**]. The number of surrogate variables was chosen to minimize apparent batch effects while avoiding overfitting based on evidence from principal components analysis and relative log expression (Supplemental Figures 2 – 5). 16 surrogate variables were found to be sufficient for downstream analysis of both gene and isoform data.

Validation of isoform expression with qPCR

Computationally assigned BrainSpan values were compared against relative isoform expression of 14 genes from independent brain samples quantified with qPCR. (Supplementary Table 1, Supplementary Figure 6). RNA from a frontal lobe tissue sample of a 22 weeks old female (fetal brain), and RNA from cerebral cortex tissue sample of a 27 years old female (adult brain) (AMSBIO, UK), corresponding to P06 (late mid-fetal) and P12 (young adult) in the BrainSpan data, was used. The BrainSpan isoform expression data was then compared to the qPCR experimental expression results as described below.

Multi-isoform genes carrying at least two isoforms that are expressed during P6 and P9 periods were selected. To select the genes, the following criteria were used: (1) computationally assigned expression differences between two isoforms had to be at least 2-fold; (2) to ensure that isoforms are expressed within the detection limits, the expression of one isoform had to be ¿= 40 TPM, while the expression of the other isoform had to be ¿= 10 TPM. 14 genes were randomly selected from the ones that passed these criteria to test by qPCR from independent samples. Primers were designed using exon-exon junctions specific for each of the selected isoforms. 3 μg of RNA using SuperScript II Kit (Invitrogen) were reverse transcribed to cDNA, following manufacturer's instructions. Then, the cDNA was diluted ten times to use as a template for the qPCR reaction. SYBR Green II Master Mix (Invitrogen) was used for the qPCR reaction,

performed in a CFX Connect 96X Thermal Cycler, using standard parameters for SYBR Green. Relative expression between each isoform in the two samples was calculated by normalizing each expression value against two housekeeping genes (RPL28 and MRSP36) as control using QIAGEN control primers, and $\Delta\Delta$t method was applied using the CFX Manager Software. Comparison of the directionality of these relative expressions against the BrainSpan expressions resulted in positive correlation (Supplementary Figure 6).

Differential expression analysis

Differential expression analysis of gene and isoform counts data was performed using the *limma* (v3.40.6) R package [**?**]. The *limma* package performs differential expression analysis by fitting a linear model to each feature (gene or isoform) expression vector. Parametric empirical Bayes is used to incorporate expression variabilities among all tests for estimation of common feature variance and increasing statistical power. Relevant covariates and surrogate variables were included in the linear model as fixed effects. The *duplicateCorrelation* function of the package was used to fit the donor identifier as a random effect for the blocking variable in the model, estimating a linear mixed effect model, to account for the nested expression measurements from individual regions measurements per donor [**?**]. Significantly differentially expressed features were defined as features with an absolute fold change ≥ 1.5 and FDR-adjusted p-value ≤ 0.05.

Cell type and literature curated lists enrichment analyses

Fisher-exact tests were performed on gene lists and isoform lists (converted to gene identifiers) against curated gene lists: Mutationally Constraint Genes (Mut. Const. Genes) [**?**], FMRP Target genes [**?**], high risk ASD genes (Satterstrom ASD) [**?**], CHD8 Target genes [**?**], synaptic genes (Synaptome DB) [**?**], genes intolerant to mutations (pLI099) [**?**], Syndromic and

rank 1 and 2 ASd risk genes (SFARISyndromicRisk12) (https://gene.sfari.org/). Cell types were extracted from two recent single cell sequencing studies [**?**, **?**].

Gene set enrichment analysis

Gene set enrichment analysis was performed using the *gprofiler2* v0.1.5 R package [**?**]. Ensembl gene identifiers (or transcript identifiers converted to gene identifiers) were enriched for Gene Ontology: Biological Processes and Gene Ontology: Molecular Functions terms. Enrichment p-values were adjusted for multiple hypothesis testing with Benjamini-Hochberg FDR, and overly general terms (i.e., terms with more than 1000 members) were removed.

Rare *de novo* ASD loss of function variants

Rare *de novo* variant data was downloaded from [**?**], and was processed using Ensembl's Variant Effect Predictor v96 tool using human genome version GRCh37 to annotate variants for predicted functional consequences [**?**]. Loss of function consequences were defined as variants impacting essential splice donor/acceptor sites, frameshift insertions and deletions, predicted start losses, and predicted stop gains.

Weighted co-expression network construction

Co-expression networks were constructed using the WGCNA (v1.68) R package [**?**]. All relevant covariates and surrogate variables were first regressed out of both gene and isoform expression sets using linear mixed effects models. Each transformed expression matrix was then tested for scale-free topology to estimate a soft thresholding power (2 for the gene co-expression network and 3 for the isoform co-expression network), and each signed network was constructed

blockwise, using a single block for the gene data and three blocks for the isoform data using deepSplit=2 and minModuleSize=20 for module detection in both networks.

Co-expression module characterization

Module eigengene-developmental period association analysis was performed using linear mixed effects models considering the previously stated fixed effects and random donor effects to account for multiple, unpaired brain region samples per donor. Enrichment of each module was performed using Fisher-exact tests against curated gene lists; isoform module members were converted to parent gene identifiers for this purpose. Gene set functional enrichment analysis for Gene Ontology terms was performed using Ensembl gene/isoform identifier queries with inputs ordered by module membership (kME).

Co-expression module variant impact analysis

Co-expression modules were measured for their rate of module member impact by distinct rare *de novo* loss of function variants. Given that modules with isoforms or genes that cover more positions in the genome are more likely to be impacted by any given genomic variant, we first calculated the genomic coverage of the module members. We then measured the number of ASD case variants and control variants that impact each module, normalized by this module genomic coverage in kilobases. These values were finally scaled by the total number of variants analyzed, and further scaled by a factor of 1,000,000. Differences in impact rates between cases and controls for each module were tested with permutation tests. 1000 iterations of module member resampling were performed, selecting based on similar feature GC content and feature length (+/- 10% for each attribute) (Supplementary Table 12).

Protein-protein interaction network overlay with co-expression modules

Gene-level PPI network data was manually curated, filtering for physical interactions and co-complex associations, from HuRI [**?**], Bioplex [**?**], HPRD [**?**], Inweb [**?**], HINT [**?**], BioGRID [**?**], GeneMANIA [**?**], STRING [**?**], and CORUM [**?**].

Gene co-expression module 1 and isoform co-expression module 1 were filtered for edges supported by the gene-level PPI; edges among isoforms were retained if the transcribing gene connection was supported by a PPI edge. Networks were then filtered for the top 10% of edges by Pearson correlation coefficient and singletons were removed.

Minigenes cloning

Five genes (*SNC2A*, *DYRK1A*, *CELF2*, *DLG2*, and *BTRC*) were selected that were impacted by variants expressed by patients with autism spectrum disorder or schizophrenia for the experiments. The exons of these genes that are likely impacted by splice site mutations were cloned, together with the 1kb of their flanking intronic sequence. The constructs were cloned into pDESTSplice exon trapping expression vector [**?**]. The site-directed mutagenesis by two-step stich PCR was performed to introduce the mutation affecting the splice site.

The minigenes were generated by PCR-amplifying the desired sequences from genomic DNA (Clontech). Primers were designed for each minigene, and attB sites were added at the 5' end of the primers. The sequences of the primers were as follows: (1) *SCN2A*; Fw: GGAAGCTATGTTTAGCCAGGATACATTTGG, Rv: CCAGATGATGTCCCCTCCCTACATAGTCC;
(2) *DYRK1A*: Fw: GTTGGGAAAATTTCCCCCTATTTAAGC, Rv: CCCAGAGGCTTAATAAAGTATGGACC;
(3) *CELF2*: Fw: GGAGTTGGAATGACAGACGTTCACATGC, Rv: CCGCTGTGGGCTGAGGATCAGTTTCC;
(4) *DLG2*: Fw: GAGGTTCAGAGACATTCAATTCCC, Rv: CTTGATGCTGTCCAGATAATGC;

(5) *BTRC*: Fw: GGGCCTCAGAATGACACAGTACG, Rv: GAACTTGCGTTTCTTGTTTTTGCC. After PCR amplification, amplicons were loaded in a 1% low EEO agarose gel (G-BioSciences) and purified using the QIAquick Gel Extraction Kit (QIAGEN) following manufacturer's instructions. Purified amplicons were subcloned into pDON223.1 expression vector using the BP-Gateway System (Invitrogen). At least six different clones for each minigene were sequenced to verify correct sequences of the minigenes. The clone with the desired sequence and highest DNA concentration was used for subcloning into the pDESTSplice expression vector (Addgene) using the LR-Gateway System (Invitrogen).

Co-Immunoprecipitation and Western Blot

HeLa cells were harvested and rinsed once with ice-cold 1xPBS, pH 7.2, and lysed in immunoprecipitation lysis buffer (20 mM Tris, pH 7.4, 140 mM NaCl, 10% glycerol, and 1% Triton X-100) supplemented with 1xEDTA-free complete protease inhibitor mixture (Roche) and phosphatase inhibitor cocktails-I,II (Sigma Aldrich). The cells were centrifuged at 16,000xg at 4°C for 30min, and the supernatants were collected. Protein concentration was quantified by modified Lowry assay (DC protein assay; Bio- Rad). The cell lysates were resolved by SDS-PAGE and transferred onto PVDF Immobilon-P membranes (Millipore). After blocking with 5% nonfat dry milk in TBS containing 0.1% Tween 20 for 1hr at room temperature, membranes were probed overnight with the appropriate primary antibodies. They were then incubated for 1h with the species-specific peroxidase-conjugated secondary antibody. Membranes were developed using the Pierce-ECL Western Blotting Substrate Kit (Thermo Scientific).

For immunoprecipitation experiments, samples were lysed and quantified as described above. Then, 3 mg of total protein was diluted with immunoprecipitation buffer to achieve a concentration of 3 mg/ml. A total of 30μl of anti-V5-magnetic beads-coupled antibody (MBL) was added to each sample and incubated for 4h at 4°C in tube rotator. Beads were then washed

twice with immunoprecipitation buffer and three more times with ice cold 1xPBS. The proteins were then eluted with 40μl of 2xLaemli buffer. After a short spin, supernatants were carefully removed, and SDS-PAGE was performed. The following primary antibodies were used: anti-V5 (1:1000; Invitrogen), anti-β-catenin (1:1000; Abcam), anti-p-βcatenin (1:1000; Cell Signaling), anti-Cul1 (1:1000; Abcam), anti-SKP1 (1:1000; Cell Signaling), and anti-βactin (1:10000; Thermo Scientific).

A Figures

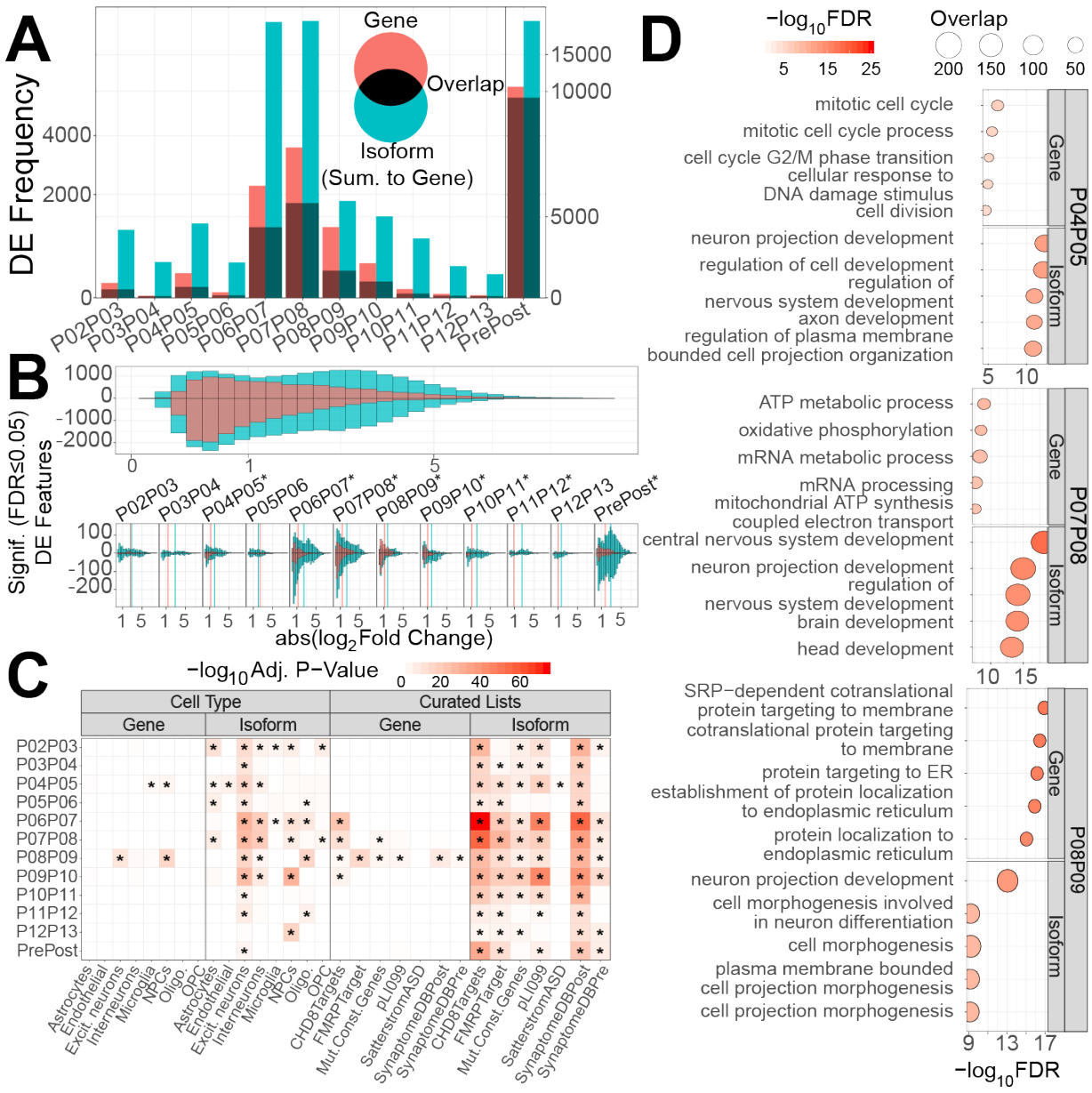


Figure 1: Differential expression analysis of gene and isoform quantification data. A) Frequency of significant differential expression results at gene-level and isoform-level (summarized to distinct gene parents). B) Effect size (absolute log2 fold change) distribution using total DE results (top) or DE per developmental period contrast (bottom). Average absolute effect sizes for gene data and isoform data are marked by colored vertical lines and differences were tested with two-sample T-tests (*FDR ¡ 0.05). C) Fisher-exact test of enrichment of genes differentially expressed specifically using isoform-level quantifications versus genes specifically differentially expressed with gene-level quantifications. D) Functional enrichment analysis of features specifically DE in either dataset from selected comparisons shows higher signals for nervous system related processes specifically in isoform-level analysis and not with gene-level analysis.

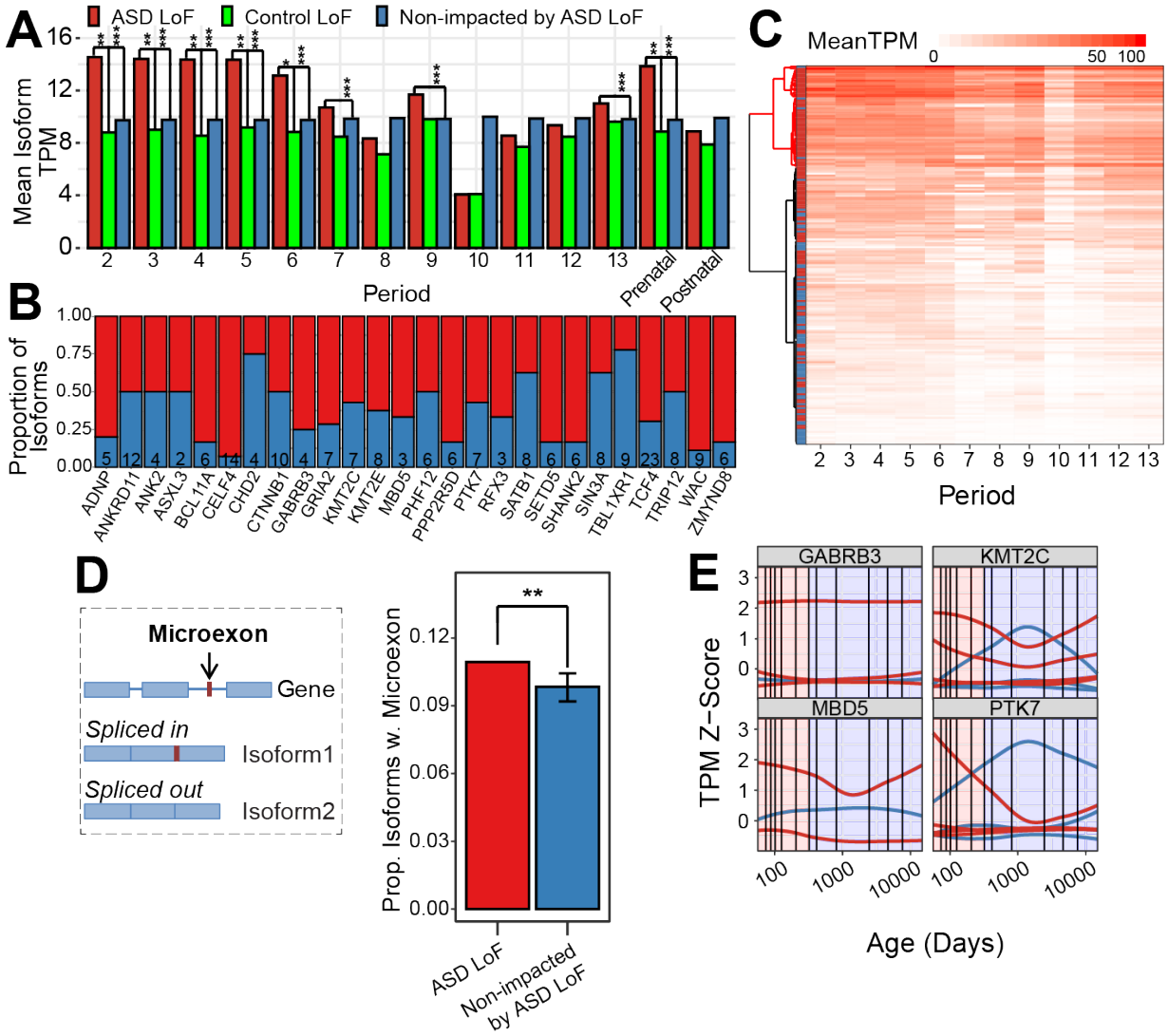
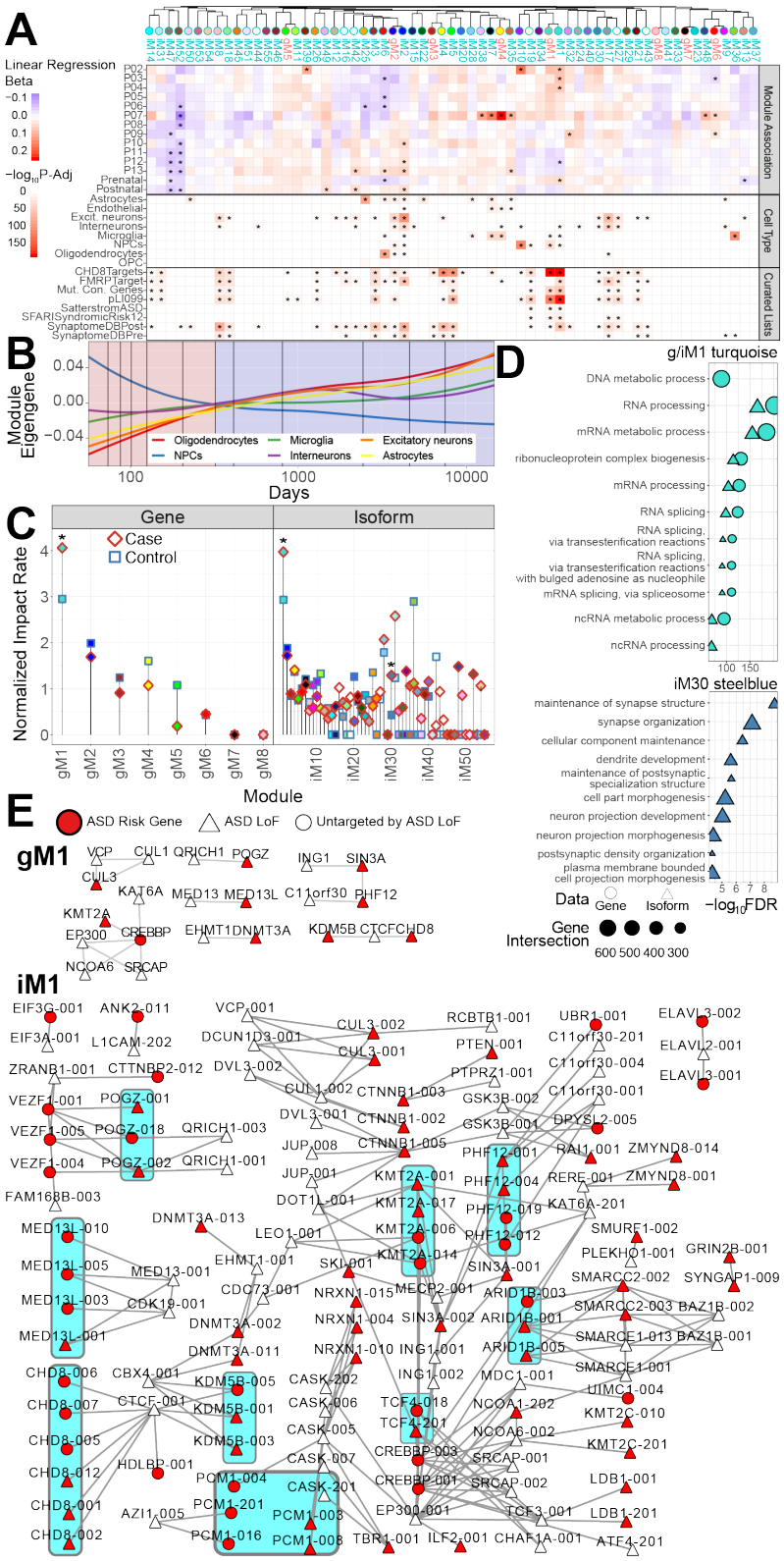


Figure 2: Rare de novo ASD loss of function variants. A) Mean isoform TPM expressions of isoforms impacted by rare de novo ASD LoF variants (ASD LoF) were compared to those of control targets (Control LoF) and non-impacted isoforms (Non-impacted by ASD LoF). B) Proportion of protein-coding isoforms of high-risk ASD genes, specifically differentially expressed at isoform level, either impacted or not impacted by rare de novo ASD LoF variants. C) Ward hierarchical clustering of isoforms from B) and heatmap of average expression per developmental period per isoform. D) Schematic of definition of alternatively regulated microexon (left), along with proportions of isoforms carrying microexons (right), stratified by impact status. E) Selected expression profiles show high variability of sibling isoform expressions transcribed from high-risk ASD genes, and these isoforms tend to be more highly expressed prenatally. $*P \leq 0.1, **P \leq 0.05, ***P \leq 0.01$

Figure 3: Gene and isoform co-expression modules reflect distinct signals in neurodevelopment. A) Modules clustered by module eigengene (top); module eigengene-developmental period associations measured by linear mixed effect model beta coefficients (middle); Fisher-exact enrichment tests against cell-type and literature-curated gene lists (bottom). B) Module eigengene expression profiles of modules most significantly associated with each cell type: Astrocytes, iM25; Oligodendrocytes, iM6; Microglia, iM36; NPCs, iM10; Excitatory neurons, iM2; Interneurons, iM17. C) Normalized module impact rate, per module and per status. Modules were selected based on significantly higher impact rate by ASD cases as compared to controls (permutation test, 1000 permutations, FDR ≤ 0.05). D) Functional enrichment for significantly impacted modules shows overrepresentation of processes related to DNA metabolic processes, RNA splicing, synapse structure and synapse organization. E) Gene-level PPIs overlaid onto gM1 and iM1, after filtering for direct connections between ASD risk features and ASD LoF features and retaining edges in the top 10% of Pearson correlation coefficients; both ASD LoF and Non-impacted by ASD LoF isoforms are shown for ASD risk genes, whereas only ASD LoF isoforms are shown for non-risk genes. Module iM1 shows differential connectivity strengths among sibling isoforms, and impacted isoforms show higher PCCs through gene-level PPIs relative to unaffected sibling isoforms. ASD risk genes with differentially impacted isoforms are highlighted in turquoise.


A
Linear Regression
Beta
−log10P-Adj
Module Association
Cell Type
Curated Lists
B
Module Eigengene
Days
Oligodendrocytes
Microglia
Excitatory neurons
NPCs
Interneurons
Astrocytes
C
Gene
Isoform
Case
Control
Normalized Impact Rate
Module
D
g/iM1 turquoise
DNA metabolic process
RNA processing
mRNA metabolic process
ribonucleoprotein complex biogenesis
mRNA processing
RNA splicing
ncRNA metabolic process
ncRNA processing
iM30 steelblue
maintenance of synapse structure
synapse organization
cellular component maintenance
dendrite development
cell part morphogenesis
neuron projection development
neuron projection morphogenesis
postsynaptic density organization
cell projection morphogenesis
Data
Gene
Isoform
−log10FDR
Gene Intersection
E
ASD Risk Gene
ASD LoF
Untargeted by ASD LoF
gM1
iM1

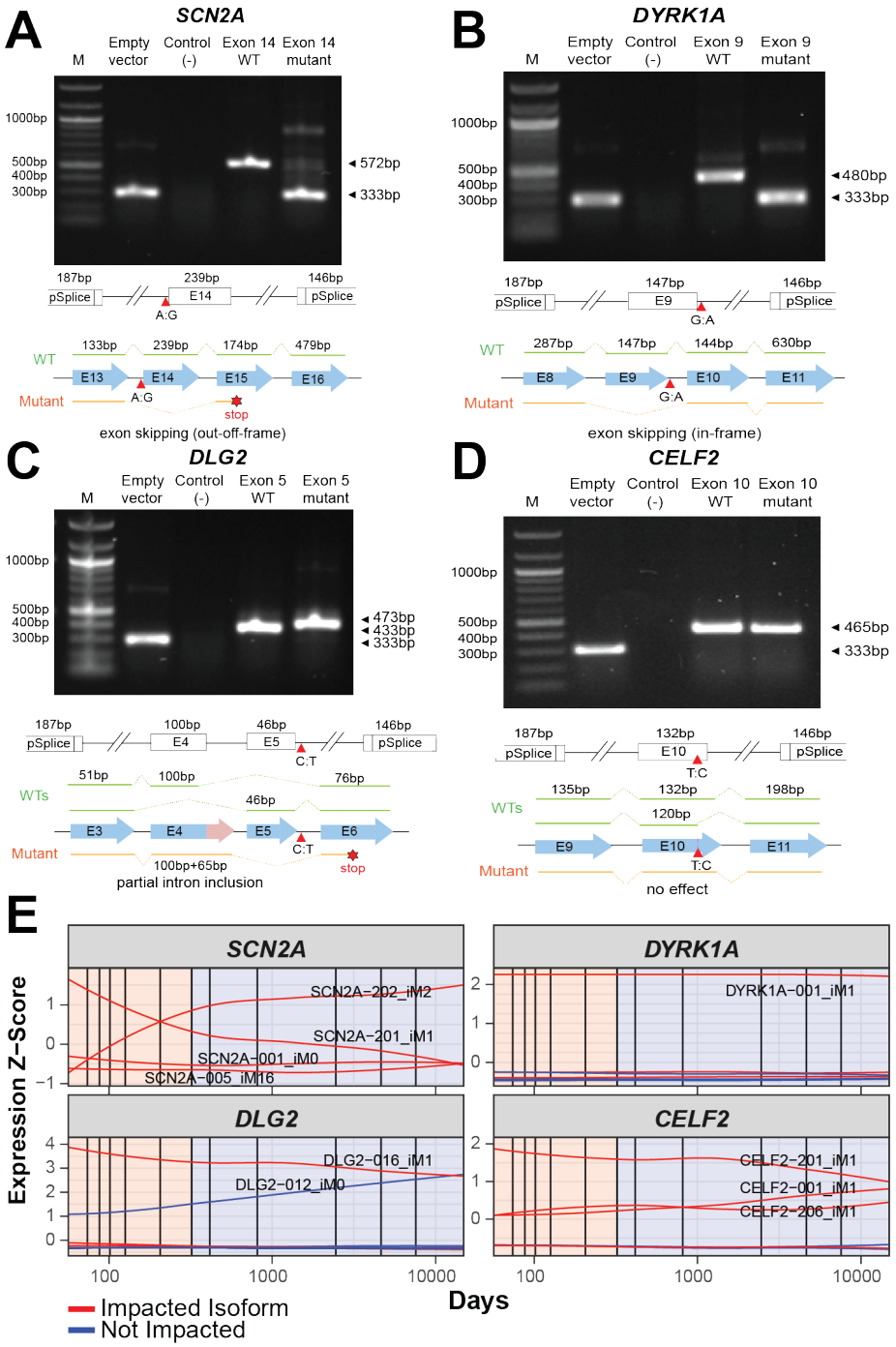
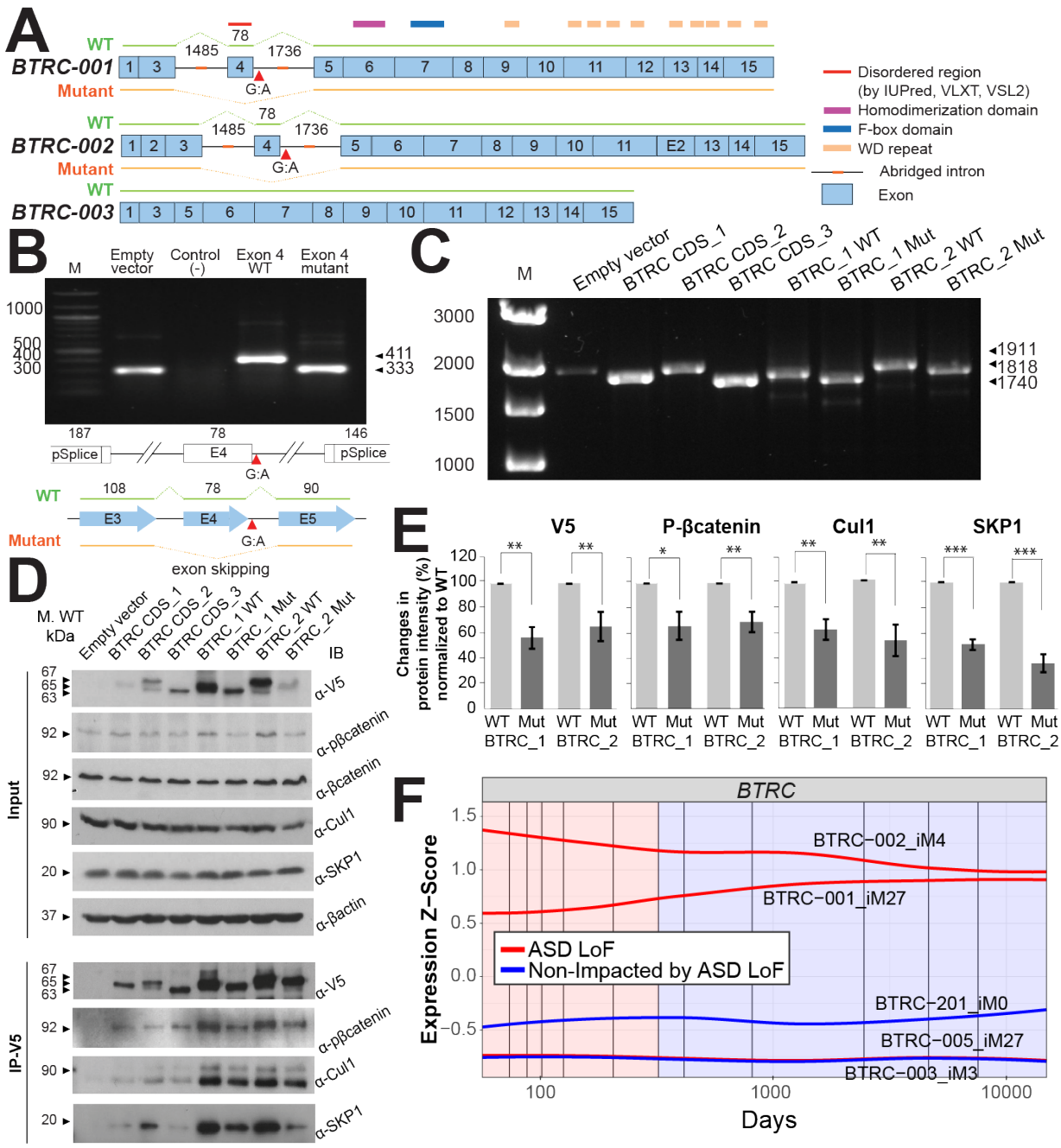
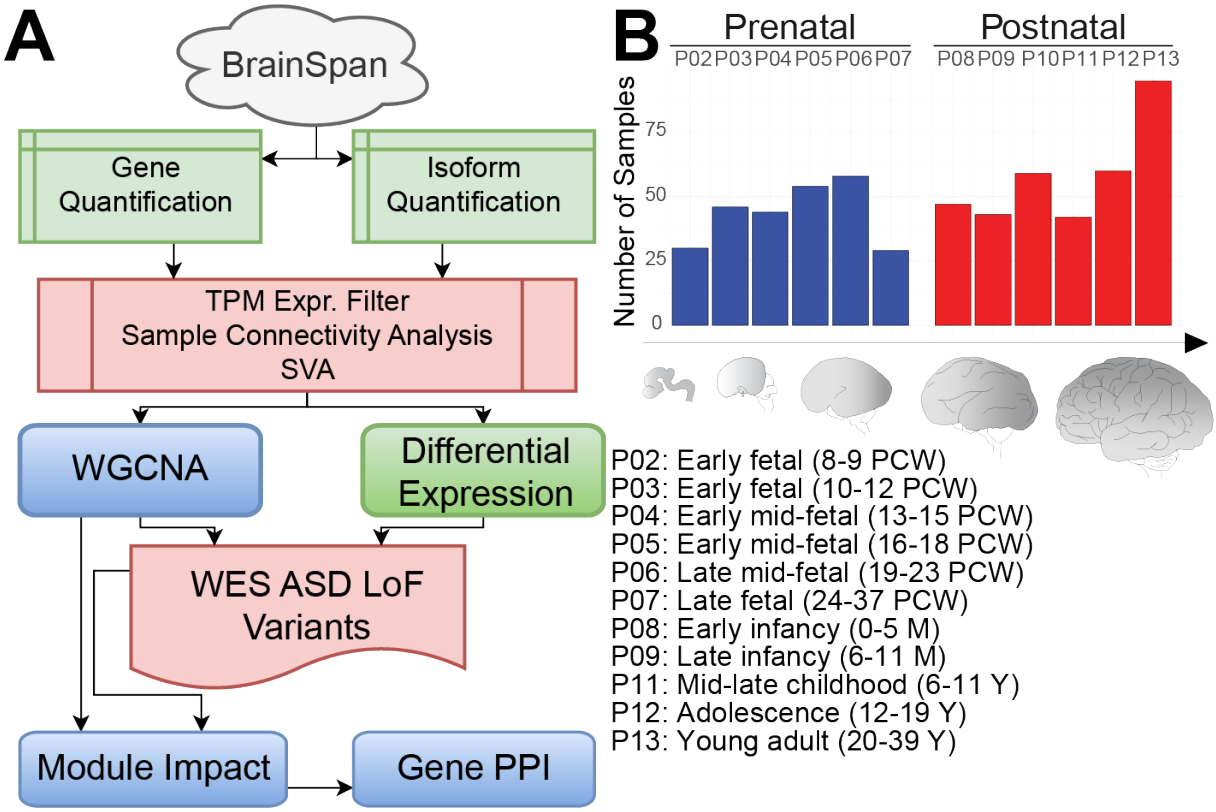


Figure 4: Functional effect of the de novo splice site mutations from the patients with neurodevelopmental diseases. Minigene assays demonstrate the effect of splice site mutations in four genes. (A) *SCN2A*; (B) *DYRK1A*; (C) *DLG2*; and (D) *CELF2*. Schematic representation of the cloned minigenes, the expected splicing patterns, and the impact of the mutations are shown below the gel image. Numbers denote base pairs; M: molecular marker; E: exon. E) Expression profiles of the brain-expressed isoforms transcribed by these four genes, annotated with module memberships; highly overlapping expression profiles are unlabeled for readability.

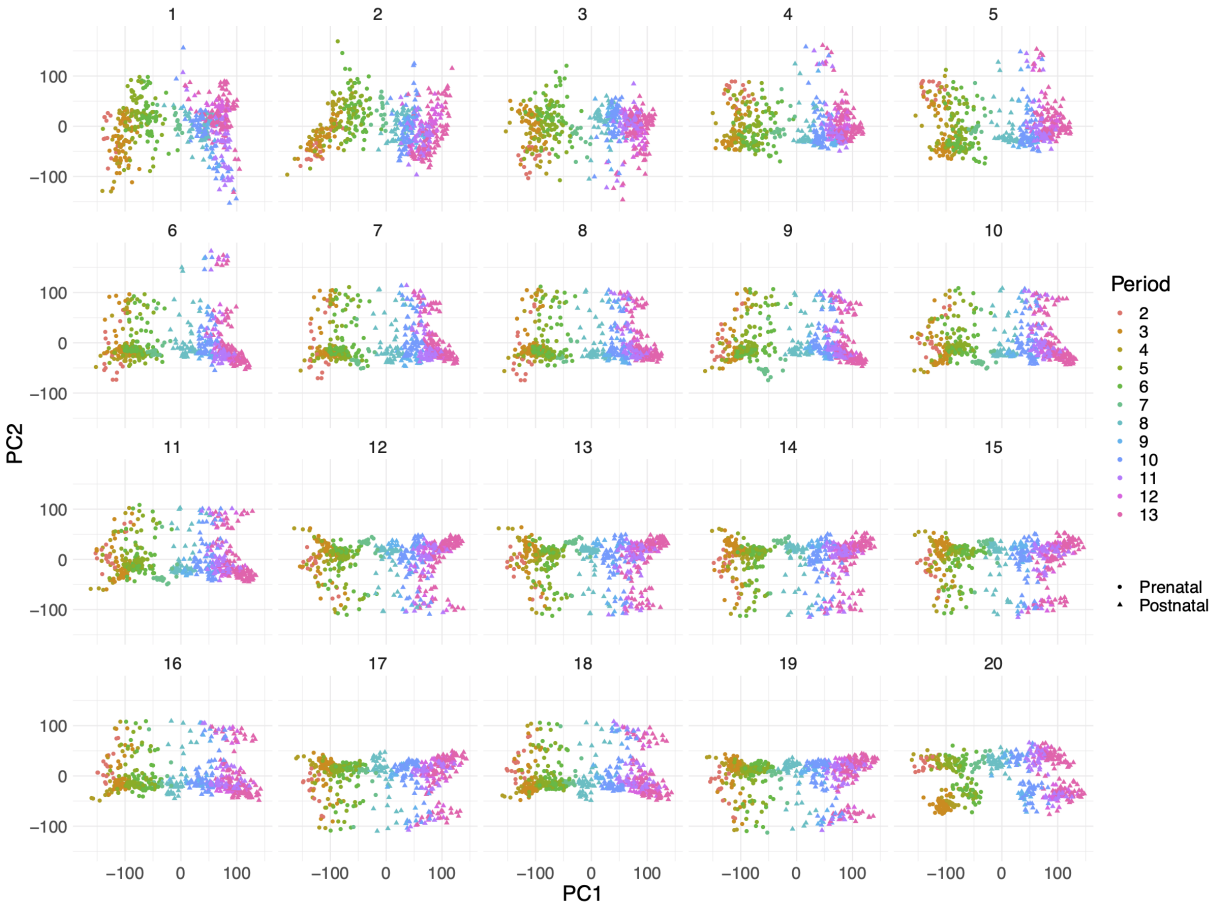
Figure 5: The de novo autism splice site mutation causes exon skipping in *BTRC* isoforms and reduces their translational efficiency. A) The exon structure of three splicing isoforms of the *BTRC* gene showing positions of the cloned abridged introns and the splice site mutation; numbers denote base pairs; B) Minigene assays demonstrate exon 4 skipping as a result of the splice site mutation. The assays show the RT-PCR results performed using total RNA from HeLa cells transfected with *BTRC* minigene constructs; numbers denote base pairs; C) Splicing assays with the full-length constructs carrying abridged introns confirm exon skipping observed in the minigene assays; D) Immunoblotting (IB) from the whole cell lysates of HeLa cells transfected with different *BTRC* minigene constructs and an empty vector, as indicated. Membranes were probed to observe *BTRC* overexpression, and to investigate expression of p-β-catenin, *Cul1* and *SKP1*. β-actin was used as loading control. Immunoprecipitation was performed with the antibody recognizing V5-tag and proteins were detected by IB with the p-β-catenin, *Cul1*, *SKP1* and V5 antibodies. The splice site mutation causes reduced translational efficiency of both BTRC_1Mut and BTRC_1Mut mutant isoforms as compared to their wild type counterparts; E) Quantification of protein pull-downs with V5-IP using ImageJ software. The band intensity values were normalized to WT expression levels. Error bars represent 95% confidence intervals (CI) based on 3 independent experiments. On average, 40% reduction of *BTRC* protein expression is observed as a result of a mutation. Consequently, the reduction of the corresponding *BTRC* binding partners (p-β-catenin, *Cul1*, and *SKP1*) is also observed. F) Expression profiles of brain-expressed *BTRC* isoforms show higher expression of ASD-impacted *BTRC*-001 and *BTRC*-002. Numbers denote base pairs (A, B, C) or kDa (D). $*P \leq 0.05, **P \leq 0.01, ***P \leq 0.001$.


A
WT
BTRC-001
Mutant
WT
BTRC-002
Mutant
WT
BTRC-003
78
1485
1736
G:A
Disordered region
(by IUPred, VLXT, VSL2)
Homodimerization domain
F-box domain
WD repeat
Abridged intron
Exon
B
M
Empty vector
Control (-)
Exon 4 WT
Exon 4 mutant
1000
500
400
300
411
333
187
78
146
pSplice
E4
pSplice
G:A
WT
108
78
90
E3
E4
E5
Mutant
G:A
exon skipping
C
M
Empty vector
BTRC CDS_1
BTRC CDS_2
BTRC CDS_3
BTRC_1 WT
BTRC_1 Mut
BTRC_2 WT
BTRC_2 Mut
3000
2000
1500
1000
1911
1818
1740
D
M. WT
kDa
IB
Input
IP-V5
α-V5
α-pβcatenin
α-βcatenin
α-Cul1
α-SKP1
α-βactin
67
65
63
92
90
20
37
E
V5
P-βcatenin
Cul1
SKP1
Changes in protein intensity (%) normalized to WT
WT Mut WT Mut
BTRC_1 BTRC_2
F
BTRC
Expression Z-Score
Days
BTRC−002_iM4
BTRC−001_iM27
BTRC−201_iM0
BTRC−005_iM27
BTRC−003_iM3
ASD LoF
Non-Impacted by ASD LoF

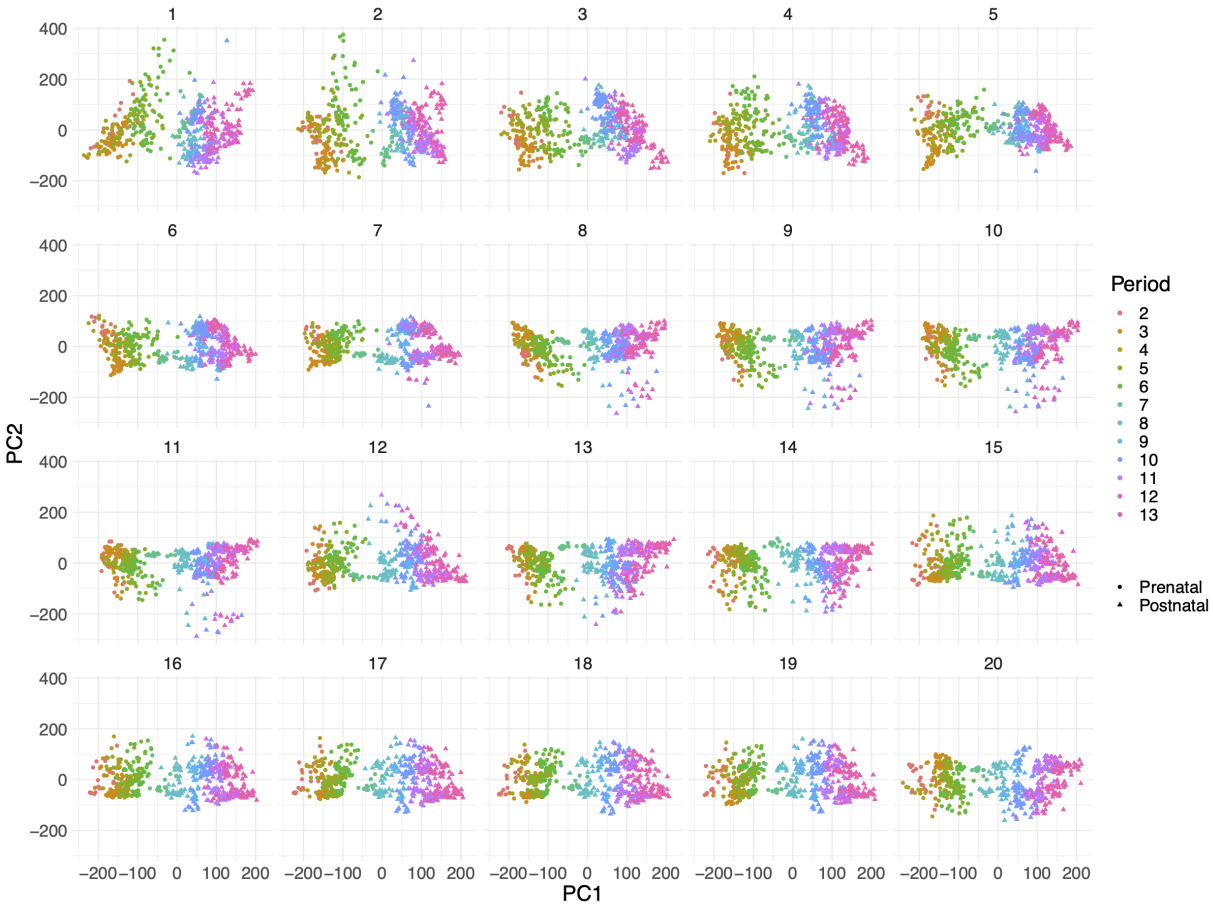
B Supplementary Figures



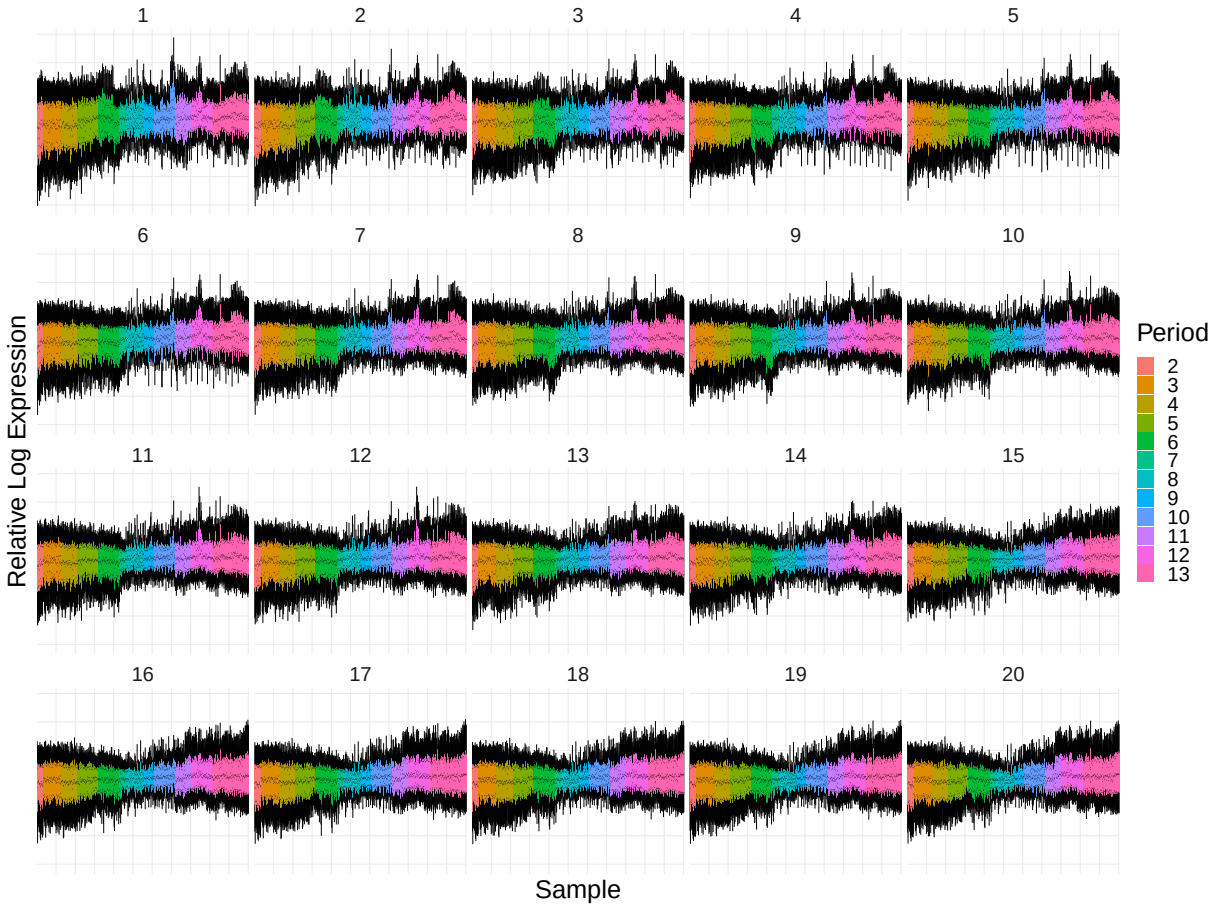
Supplementary Figure 1: RNA-Seq data was obtained from BrainSpan. A) Schematic of bioinformatics analysis of BrainSpan data: Beginning with gene and isoform quantifications downloaded from BrainSpan, features were filtered based on TPM. Outlier samples were detected and removed. Surrogate Variable Analysis was performed to account for latent batch effects. Temporal differential expression was performed on both datasets. WGCNA co-expression networks were created and analyzed on both datasets. Whole exome sequencing data was filtered for LoF variants and mapped to features. B) Initial samples were divided into distinct developmental periods.



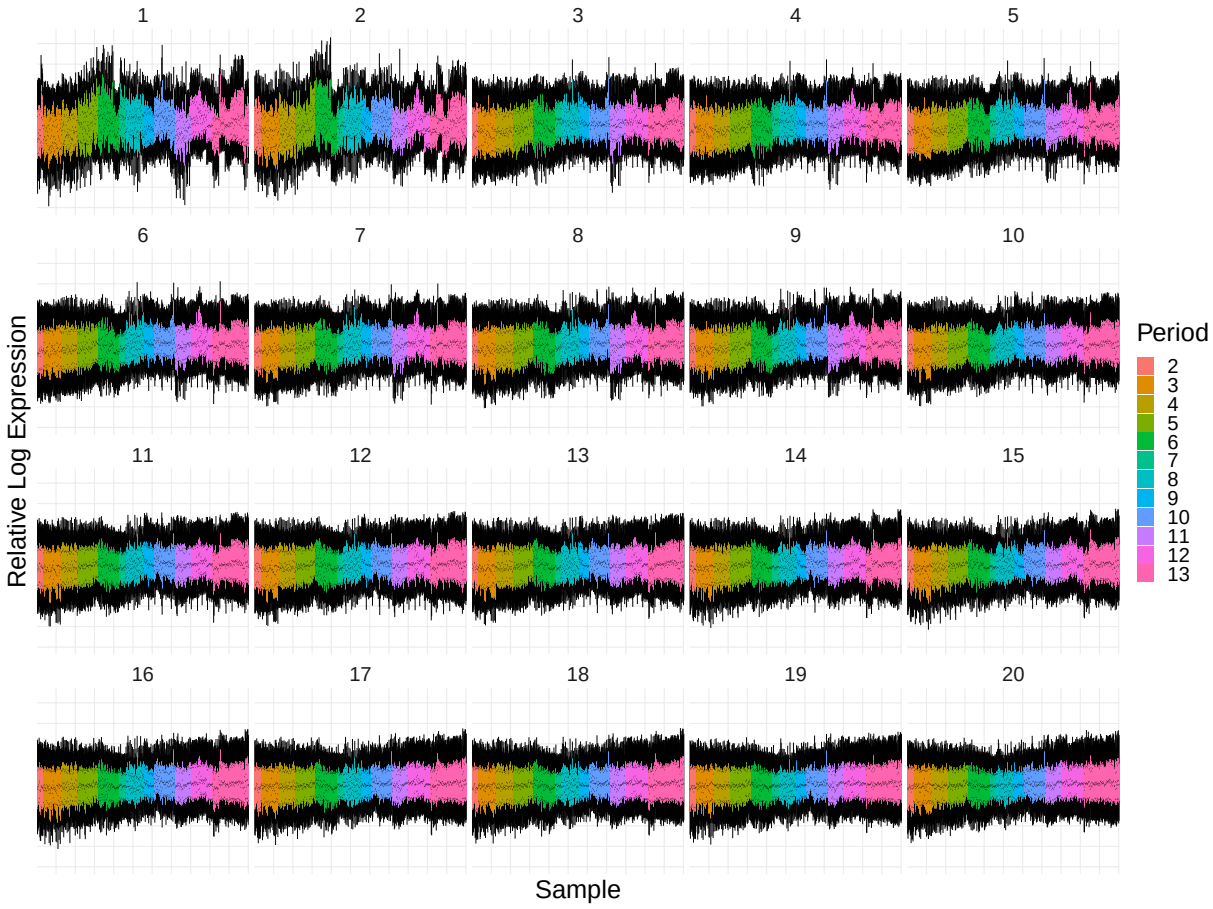
Supplementary Figure 2: Principal components analysis of transformed gene quantifications. Gene expression data was transformed through regression of relevant covariates (age, brain region, gender, ethnicity, study site, surrogate variables) for each count of surrogate variables analyzed to determine the appropriate number of surrogate variables.



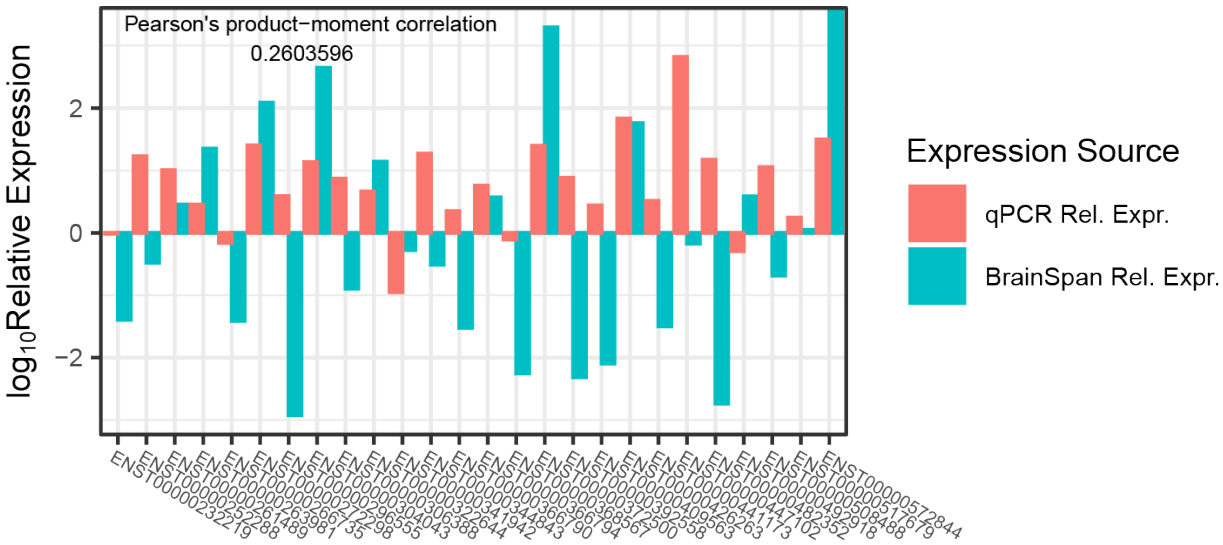
Supplementary Figure 3: Principal components analysis of transformed isoform quantifications. Isoform expression data was transformed through regression of relevant covariates (age, brain region, gender, ethnicity, study site, surrogate variables) for each count of surrogate variables analyzed to determine the appropriate number of surrogate variables.



Supplementary Figure 4: Relative log expression analysis of transformed gene quantifications. Gene relative log expressions were calculated, per sample, to find moment of relative log expression stability in surrogate variable selection.



Supplementary Figure 5: Relative log expression analysis of transformed isoform quantifications. Isoform relative log expressions were calculated, per sample, to find moment of relative log expression stability in surrogate variable selection.



Supplementary Figure 6: Comparison of relative expression from qPCR and BrainSpan. Age- and gender-matched samples were compared for the isoforms of 14 genes; positive Pearson correlation was found for the sign of relative expressions.

This thesis, in full, is currently being prepared for submission for publication of the material. Chau, Kevin K.; Zhang, Pan; Urresti, Jorge; Amar, Megha; Pramod, Akula Bala; Corominas, Roser; Lin, Guan Ning; Iakoucheva, Lilia M. The thesis author was the primary investigator and author of this material.

Supplementary Files

Supplementary Table 1: This table consists of relevant data for validation of BrainSpan TPM expression values against age-matched qPCR results.

Supplementary Table 2: Full gene-level differential expression results.

Supplementary Table 3: Full isoform-level differential expression results.

Supplementary Table 4: Composition of differential expression results.

Supplementary Table 5: Gene ontology enrichment analysis results for gene-level differential expression analysis.

Supplementary Table 6: Gene ontology enrichment analysis results for isoform-level differential expression analysis.

Supplementary Table 7: Compiled and annotated variants table.

Supplementary Table 8: Gene co-expression module assignments.

Supplementary Table 9: Isoform co-expression module assignments.

Supplementary Table 10: Gene ontology enrichment analysis of gene co-expression modules.

Supplementary Table 11: Gene ontology enrichment analysis of isoform co-expression modules

Supplementary Table 12: Module impact rate analysis results