UNIVERSITY OF CALIFORNIA SAN DIEGO

Impact of *de novo* Splicing Mutations on Human Brain Spliceoform Dynamics in Autism Spectrum Disorder

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

in

Biology

by

Kevin Khai Chau

Committee in charge:

Professor Lilia Iakoucheva, Chair
Professor Scott Rifkin, Co-Chair
Professor Barry Grant

2019

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

DEDICATION

To two, the loneliest number since the number one.

EPIGRAPH

A careful quotation

conveys brilliance.

—Smarty Pants

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

Thanks to whoever deserves credit for Blacks Beach, Porters Pub, and every coffee shop in San Diego.

Thanks also to hottubs.

VITA

2002	B. S. in Mathematics *cum laude*, University of Southern North Dakota, Hoople
2002-2007	Graduate Teaching Assistant, University of California, San Diego
2007	Ph. D. in Mathematics, University of California, San Diego

PUBLICATIONS

Your Name, "A Simple Proof Of The Riemann Hypothesis", *Annals of Math*, 314, 2007.

Your Name, Euclid, "There Are Lots Of Prime Numbers", *Journal of Primes*, 1, 300 B.C.

ABSTRACT OF THE DISSERTATION

Impact of *de novo* Splicing Mutations on Human Brain Spliceoform Dynamics in Autism Spectrum Disorder

by

Kevin Khai Chau

Master of Science in Biology

University of California San Diego, 2019

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

This dissertation will be abstract.

Chapter 1

Impact of *de novo* Splicing Mutations on Human Brain Spliceoform Dynamics in Autism Spectrum Disorder

1.1 Background

More than 95% of multi-exon human genes undergo alternative splicing (AS) and/or use alternative promoters to increase transcriptomic and proteomic diversity [1, 2] . As a result, multiple isoforms can be generated from each gene, with an estimated average of five to seven isoforms per gene (Pan et al., 2008; Steijger et al., 2013). Alternative splicing is highly specific, and expression of isoforms is often restricted to certain organs, tissues or even cell types within the same tissue (Barbosa-Morais et al., 2012; Shalek et al., 2013; Trapnell et al., 2010). In addition, many isoforms are expressed only during specific developmental periods (Kalsotra and Cooper, 2011) . The alternatively spliced isoforms encoded by the same gene can also be expressed at different levels in the same tissue or during the same developmental period [2] .

The brain is one of the tissues with the highest number of AS events (Calarco et al.,

2011; Mele et al., 2015; Raj and Blencowe, 2015; Yeo et al., 2004). Several recent studies demonstrated that processes occurring during neural development, such as cell-fate determination, neuronal migration, axon guidance and synaptogenesis, are controlled by differentially expressed alternatively spliced isoforms (Grabowski, 2011; Kim et al., 2013; Li et al., 2007). Quantitative differences in brain isoform expression have been previously analyzed primarily in the model organisms (Trapnell et al., 2010), and limited to specific brain regions, sets of genes, neuronal cell types or developmental periods (Chen et al., 2008; Shekhar et al., 2016; Zhang et al., 2014). However, the global spatiotemporal diversity of splicing isoform expression within the developing human brain remained relatively unexplored. Furthermore, how disease-associated mutations impact different isoforms is still an open question.

Since many exons are differentially used, it is likely that the disease mutations may selectively impact only isoforms with mutation-carrying exons. In addition, if some isoforms are not expressed at a particular developmental period or in a particular tissue, then the disease mutations affecting such isoforms may not manifest their functional impact at that period or in that tissue. Correlating period- and tissue-specific isoform expression with disease mutations is an important and necessary task for improving our understanding of human diseases.

Investigating the impact of the disease mutations at the isoform-level is also needed, because different protein isoforms have drastically different protein interaction capabilities. We showed that the majority of the isoforms encoded by the same gene share less than a half of their interacting partners in the human interactome network (Yang et al., 2016). This observation points to striking functional differences between splicing isoforms that are not accounted for by the majority of the existing gene-level studies. Our recent work demonstrated that isoform-level networks provide better resolution and depth around disease candidate proteins compared to the gene-level networks (Corominas et al., 2014).

Integration of brain spatiotemporal transcriptome with the genetic data from exome sequencing studies has provided important insights into neurodevelopmental diseases (NDDs)

(Gulsuner et al., 2013; Lin et al., 2015; Parikshak et al., 2013; Willsey et al., 2013). However, most of the recent work in this area was focused on investigation of protein-coding mutations at the gene-level resolution (Chang et al., 2014; Gilman et al., 2012; Gilman et al., 2011; Iossifov et al., 2015; Samocha et al., 2014; Sanders et al., 2015; Uddin et al., 2014). Several studies have explored the role of mutations that dysregulate alternative splicing in relation to other human diseases, and targeted antisense oligonucleotides that correct splicing defects are now being tested in the clinic for myotonic dystrophy, spinal muscular atrophy and Duchenne muscular dystrophy (Bader and Hogue, 2002; Kole et al., 2012). The investigation of splice site-disrupting mutations in relation to NDDs is still lagging behind.

1.2 Results

1.2.1 Initial RNA-Seq data processing

We obtained previously quantified gene-level and transcript-level RNA-Seq data from the BrainSpan Atlas of the Developing Human Brain, which includes data from multiple regions of post-mortem brain tissue from clinically unremarkable donors aged between 8 weeks post-conception to 40 years of age [CITATION]. We then filtered this dataset for those features that are expressed in the brain and performed sample connectivity calculations to identify sample outliers (**Materials and Methods**). As a result, we obtained spatiotemporal expression profiles for 100,754 unique brain-expressed isoforms corresponding to 26,307 genes ($\sim$ 3.83 isoforms/gene).

In order to model latent batch effects, we performed surrogate variable analysis, taking into account sample age, region, ethnicity, study site, and sex as covariates [CITATION]. We chose to use 16 surrogate variables in our downstream analyses based on the collective evidence from principal components analysis, relative log expression, and the distribution of null nominal p-values from our differential expression analyses (**Materials and Methods**, **Supplementary Figure 2**).

1.2.2 Modeling the timeline of expression across neurodevelopment

We performed differential expression analysis between all adjacent developmental periods as well as between all prenatal and postnatal samples using the *limma* R package ([CITATION], **Materials and Methods**). We find a total of 16.88% of brain-expressed isoforms as differentially expressed between any two adjacent developmental periods, with a majority of those DE isoforms contributed by P06P07 (6.86%) and P07P08 (6.83%), indicating high levels of expression regulation from late-fetal to neonatal ages (FDR-adjusted P-value ≤ 0.05, abs(FoldChange) ≥ 1.5). Among just prenatal contrasts, we find 9.83% of brain-expressed isoforms to be DE (P02P03, 1.44%; P03P04, 0.76%; P04P05, 1.61%; P05P06, 0.75%; P06P07, 6.86%); among postnatal contrasts, 5.06% (P08P09, 2.13%; P09P10, 1.76%; P10P11, 1.24%; P11P12, 0.67%; P12P13, 0.51%).

1.2.3 Functional impact of *de novo* splice site mutations in temporally regulated isoforms

We explored the functional impact of *de novo* mutations affecting splice sites of four genes that are implicated in autism spectrum disorder or schizophrenia that transcript differentially expressed transcripts (SCN2A, DYRK1A, DLG2, and CELF2) using exon trapping assays (**Experimental Procedures**). The de novo splice site mutation in SCN2A (chr2:166187838, A:G, acceptor site) from a schizophrenia patient (Fromer et al., 2014) caused out-of-frame exon skipping and potentially inclusion of 30 new amino acids into the translated protein and ending with a premature stop codon (Figure 4a). In contrast, the de novo splice site mutation in DYRK1A (chr21: 38865466, G:A, donor site), identified through exome sequencing of autism patients (O'Roak et al., 2012b), caused an in-frame exon skipping, potentially producing a different variant of the same protein (Figure 4b). In case of DLG2 (chr11: 83194295, G:A, donor site), the de novo mutation from a schizophrenia patient (Fromer et al., 2014) affects a splice site adjacent

to the exon five, which is alternatively spliced in the WT isoforms (Figure 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 the protein that is truncated one residue after the end of exon 4 due to a premature stop codon. Finally, CELF2 mutation (chr10: 11356223, T:C, donor site) identified in a patient with schizophrenia (Xu et al., 2011) affects an alternatively used 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 (Figure 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. In summary, the above observations showcase different scenarios of the impact of splice site mutations on genes and confirms the need to investigate functional impact of mutations at the isoform- rather than the gene-level resolution.

We mapped these mutations to our temporally differentially expressed isoforms

1.2.4 Weighted co-expression network analysis

We applied the *WGCNA* R package to the transformed counts matrix to construct three different isoform co-expression networks: one with all samples, one with just prenatal samples, and one with just postnatal samples. Interestingly, we identify far more modules in the postnatal-specific co-expression network (79 modules) than the full network (62 modules) and the prenatal network (35 modules).

1.3 Materials and Methods

All analyses were performed using R version $\geq$ 3.5.1. False discovery rate (FDR) adjustment was used to correct for multiple hypothesis testing with a significance threshold of 0.05.

1.3.1 Pre-processing of RNA-Seq data

We downloaded RNA-Seq quantification data from the BrainSpan Atlas of the Developing Human Brian [CITATION]. This resource consists of both gene-level and isoform-level RNA-Seq derived from post-mortem brain tissue from 57 donors aged between 8 weeks post-conception through 40 years, across a number of different brain regions, for a total of 606 initial samples. This data was aligned using STAR and quantified with RSEM with the Gencode V19 human reference genome, resulting in counts and TPM matrices for both datasets [REFERENCE]. These matrices were filtered by applying a filter of TPM $\geq$ 0.1 in at least 25% of samples in both data sets; we further restricted the data to only include genes with at least one retained isoform per the isoform-level filter and vice-versa, resulting in a total of 100,734 retained isoforms.

1.3.2 Validation of isoform expression with qPCR

We used qPCR to estimate the relative isoform expression of 14 genes from independent brain samples, and compared it to the computationally assigned values. We used 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. The BrainSpan isoform expression data was then compared to the qPCR experimental expression results as described below.

We first selected multi-isoform genes carrying at least two isoforms that are expressed during P6 and P9 periods. To select the genes, we used the following criteria: (1) computationally

assigned expression difference between two isoforms has to be at least 2-fold; (2) to ensure that isoforms are expressed within the detection limits, the expression of one isoform has to be $\geq$40 TPM, while the expression of the other isoform has to be $\geq$10 TPM. We randomly selected 14 genes 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. We reverse transcribed 3 μg of RNA using SuperScript II Kit (Invitrogen) to cDNA, following manufacturer's instructions. Then, the cDNA was diluted ten times to use it 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.

Sample connectivity analysis was performed in order to detect outliers as previously described [CITATION]. In summary, 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 Z-scores. 55 samples were identified as having sample connectivity Z-scores $\leq$ -2 and were removed from downstream analysis.

1.3.3 Normalization and differential expression analysis

To normalize the isoform counts data for between-sample comparability, we first performed surrogate variable analysis to detect latent batch effects [CITATION], relying on evidence from a combination of principal components analysis, relative log expression and p-value distribution visualizations to determine the number of surrogate variables that minimizes latent batch effects while avoiding the problem of overfitting (Supplemental Figure 1). Here, we proposed to use 16 surrogate variables for downstream analysis

Differential expression analysis of normalized isoform counts data was performed using

the *limma* R package. At its core, *limma* performs differential expression analysis by fitting a linear model to each isoform expression vector. However, since simply fitting a linear model generally produces low-powered results, *limma* leverages the highly-parallel nature of genomic data to borrow and incorporate strength from every isoform linear model. Further, it is highly flexible and able to model for many contrasts at once as one whole integrated experiment. These linear models are then processed using parametric empirical Bayes, which, given the parallel nature of these isoform models, incorporates global and local expression variabilities, thereby increasing the overall degrees of freedom for the estimation of isoform-wise variance. Further, *limma* is also able to account for nested experiment designs, similar to fitting a linear mixed effects model, through the *duplicateCorrelation* function. The BrainSpan data is designed with multiple region measurements per individual, such that there are multiple expression measurements per individual "block." The *duplicateCorrelation* function of *limma* is used to calculate the consensus correlation, with the constraint that all isoforms share the same intrablock consensus correlation, which is then incorporated into the linear model to account for this nested data structure.

1.3.4 Minigenes cloning

We selected five switch genes (SNC2A, DYRK1A, CELF2, DLG2 and BTRC) for the experiments. We cloned the exons of these genes that are likely impacted by splice site mutations, together with the 1kb of their flanking intronic sequence. The constructs were cloned into pDEST-Splice exon trapping expression vector (Kishore et al., 2008). 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: GGAAGC-TATGTTTAGCCAGGATACATTTGG, Rv: CCAGATGATGTCCCCTCCCTACATAGTCC;(2) DYRK1A: Fw: GTTGGGAAAATTTCCCCCTATTTAAGC, Rv: CCCAGAGGCTTAATAAAG-

TATGGACC; (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).

1.3.5 Weighted co-expression network analysis

Co-expression networks were constructed using the *WGCNA* R package. This network construction package operates under the scale-free topology criterion, such that the given network has a degree distribution which follows a power law, and calculates pairwise correlations among the isoform expression data. We first transformed the counts data by adjusting the counts values using information from the surrogate variable analysis and incorporating any relevant covariates [CITE LINEAR MODELS]. This transformed counts matrix was then tested for scale-free topology, and the network was constructed blockwise using three blocks and the power estimate result from the scale-free topology correlation calculations. Three networks were created: one using all available samples, one using just prenatal samples, and one using just postnatal samples

1.3.6 *de novo* splice site mutations

Variant calling was performedd on *de novo* mutations from [HOW??] from both case and control autism cohorts from the Simons Simplex Consortium (SSC) and REACH [EXPAND

ACRONYM]. These datasets include [N] families with autistic children.

1.4 Discussion

Chapter 2

Single-Cell RNA-Seq of Mouse Models

Single-cell RNA-Seq analysis was performed as part of pilot study in heterogeneous 16p11.2$^{+/-}$ and Cul3$^{+/-}$ deletion mouse models. For both experiments, scRNA-Seq data was extracted and quantified using the *Cell Ranger* software package by 10X Genomics. Downstream analysis was performed using the *Seurat* R package [SATIJA LAB REFERENCE HERE].

2.1 16p11.2

2.1.1 Background

¡Why is 16p11.2 important? -¿ Copy number variant implicated in macro/microcephaly in children with autism spectrum disorder¿

2.2 Cul3 Knockout Mice

2.2.1 Background

¡Why is Cul3 important -¿ Negative regulator of RhoA, micro/macrocephaly in children with autism spectrum disorder¿

Chapter 3

Just a Test

This is only a test.

3.1 A section

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Nulla odio sem, bibendum ut, aliquam ac, facilisis id, tellus. Nam posuere pede sit amet ipsum. Etiam dolor. In sodales eros quis pede. Quisque sed nulla et ligula vulputate lacinia. In venenatis, ligula id semper feugiat, ligula odio adipiscing libero, eget mollis nunc erat id orci. Nullam ante dolor, rutrum eget, vestibulum euismod, pulvinar at, nibh. In sapien. Quisque ut arcu. Suspendisse potenti. Cras consequat cursus nulla.

3.1.1 A Figure Example

This subsection shows a sample figure.

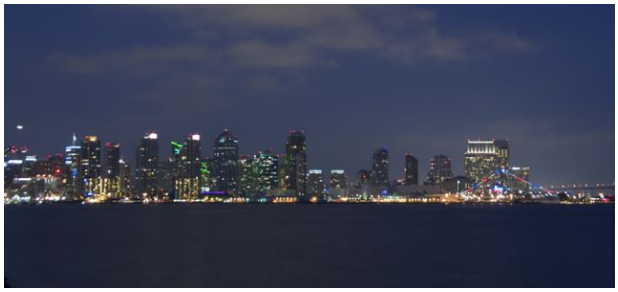

Figure 3.1: A picture of San Diego. Short figure caption must be < 4 lines in the list of figures and match the start of the main figure caption verbatim. Note that figures must be on their own line (no neighboring text) and captions must be single-spaced and appear *below* the figure. Captions can be as long as you want, but if they are longer than 4 lines in the list of figures, you must provide a short figure caption.

3.1.2 A Table Example

While in Section 3.1.1 Figure 3.1 we had a majestic figure, here we provide a crazy table example.

Table 3.1: A table of when I get hungry. Short table caption must be < 4 lines in the list of tables and match the start of the main table caption verbatim. Note that tables must be on their own line (no neighboring text) and captions must be single-spaced and appear *above* the table. Captions can be as long as you want, but if they are longer than 4 lines in the list of figures, you must provide a short figure caption.

Time of day	Hunger Level	Preferred Food
8am	high	IHOP (French Toast)
noon	medium	Croutons (Tomato Basil Soup & Granny Smith Chicken Salad)
5pm	high	Bombay Coast (Saag Paneer) or Hi Thai (Pad See Ew)
8pm	medium	Yogurt World (froyo!)

Appendix A

Final notes

Remove me in case of abdominal pain.

Bibliography

[1] Qun Pan, Ofer Shai, Leo J. Lee, Brendan J. Frey, and Benjamin J. Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40(12):1413–1415, 2008.

[2] Eric T. Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F. Kingsmore, Gary P. Schroth, and Christopher B. Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, 2008.

[3] Timothy W. Nilsen and Brenton R. Graveley. Expansion of the eukaryotic proteome by alternative splicing. *Nature*, 463(7280):457–463, 2010.

[4] Robert S. Porter, Farris Jaamour, and Shigeki Iwase. Neuron-specific alternative splicing of transcriptional machineries: Implications for neurodevelopmental disorders, mar 2018.

[5] Gene Yeo, Dirk Holste, Gabriel Kreiman, and Christopher B Burge. Variation in alternative splicing across human tissues. *Genome Biology*, 5(10):R74, 2004.

[6] Q. Xu. Genome-wide detection of tissue-specific alternative splicing in the human transcriptome. *Nucleic Acids Research*, 30(17):3754–3766, 2002.

[7] Scott L. Carter, Christian M. Brechbühler, Michael Griffin, and Andrew T. Bond. Gene co-expression network topology provides a framework for molecular characterization of cellular state. *Bioinformatics*, 20(14):2242–2250, 2004.

[8] Joshua M. Stuart, Eran Segal, Daphne Koller, and Stuart K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 2003.