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

Alternative splicing is the process by which the exons of a single gene may be differentially included or excluded in distinct mRNA transcripts to give rise to a plurality of distinct products. It is well known that much of the diversity in eukaryotic biology can be attributed to this alternative splicing of mRNA transcripts, and studies of high-throughput sequencing have shown 95-100% of human pre-mRNAs encompassing more than a single exon are processed to yield multiple mature mRNAs [1, 2]. The differential regulation and production of alternatively spliced mRNA transcripts can also be dictated by spatial or temporal cues, such as tissue specificity or developmental periods, respectively [3, 4], with the brain exhibiting higher numbers of alternative splicing events relative to other tissues[1, 5, 6].

While the specific functions of individual alternatively spliced isoforms are largely un-

explored, these may be inferred through the use of isoform co-expression networks under the idea that, although co-regulation of a pair of isoforms may not necessarily imply that they are related, large sets of isoforms that are co-expressed in a similar manner are likely to be enriched in a central function [7, 8].

1.2 Results

1.2.1 Specific co-expression modules are highly impacted by *de novo* ASD SNVs

We performed isoform co-expression network analysis using the *WGCNA* R package, which resulted in 50 modules of highly co-expressed isoforms.

In order to assess the impact of *de novo* ASD splice-site mutations and *de novo* LoF mutations on these co-expression modules, we first scored each module by their splicing mutation rates. We identify several relatively high scoring modules in the cases of either splice-site or total LoF mutations derived from the case cohort. In the case of LoF case mutation rate, M39 exhibits the highest score. M7, M22, M29, and M30 all exhibit high scores in the case of splice-site case mutation rates. We also consider mutations only affecting the known 99 ASD genes[CITATION] and find that, in the case of *de novo* LoF SNV rates, M24 exhibits the highest score, and in the case of *de novo* splice-site SNV rates, M17 shows the highest score. It should be noted that the transcripts affected by these splice-site SNVs in M17 are all derived from TCF7L2, which is a key player in canonical Wnt signalling in addition to being a high-confidence ASD gene [CITATION].

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 counts and TPM matrices, with samples 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. These matrices were filtered by applying a filter of TPM ≥ 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 100,834 retained isoforms.

1.3.2 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 (see figure []). Here, we proposed to use 11 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.

1.3.3 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 correlation with the scale-free topology, and the network was constructed blockwise using three blocks and the power estimate result from the scale-free topology correlation calculations.

1.3.4 *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.

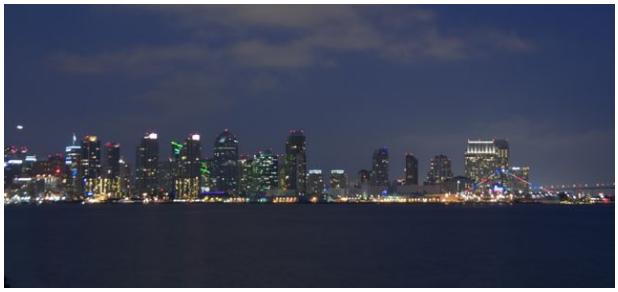

Figure 3.1: A picture of San Diego. Short figure caption must be < 4 lines in the list of figures and match the start of the main figure caption verbatim. Note that figures must be on their own line (no neighboring text) and captions must be single-spaced and appear *below* the figure. Captions can be 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.