Predicting Pathogenicity using Aberrant Splicing Predictions of Single Nucleotide Variants in Autism Spectrum Disorder

Bioinformatics Laboratory, BIMM 185

Kevin Chau

University of California, San Diego
16 June 2017

1 Introduction

The major question I wanted to answer was: "Can the likelihood that a single nucleotide variant causes cassette exon skipping be used as a predictor for that nucleotide's potential for pathogenticity?" I hypothesized that there is a correlation between an SNV's propensity to cause alternative splicing and that variant's pathogenicity probability. The evidence to either support or reject my hypothesis was the result of processing scores of variants known to be involved in autism spectrum disorder and control variants. This processing included scoring each variant for the predicted change in splicing, as well as scoring each variant for a pathogenicity score. Gene-gene coexpression networks were also constructed to calculate risk scores for each gene affected by an SNV. Much of the data processing was done with Python scripts; data management was performed using a local MySQL database.

2 Data and Processing

A list of variants classified by expressed phenotype was downloaded from DenovoDB, a database of nucleotide variants. Only variants tagged with the autism phenotype and control phenotype were downloaded. Overall, the taken from DenovoDB included the phenotype, the nucleotide variant, chromosome, affected gene, and starting position of the variant. This table was uploaded to a locally hosted MySQL database for simplified querying. Since only SNVs were analyzed, the dataset was filtered for only mutations that substitute one nucleotide for another. The SPIDEX splicing scoring tool was downloaded in order to score the nucleotide variants for their likelihood to cause cassette exon skipping.

2.1 Splicing Scores with SPIDEX

The downloaded SPIDEX tool was used to score each SNV for the likelihood that it causes cassette exon skipping. The scoring metric used was delta PSI (dpsi), the change in percent inclusion rate; negative values correspond to decreased inclusion (exon skipping) and positive values correspond to increased inclusion (exon retention). The criteria for scoring was that the SNV must have occurred within 300 nucleotides from a splice junction; that is, within 300 nucleotides from an exon-to-intron or intron-to-exon transition position.

Since not all variants fell within that distance, only scorable variants were retained for further analysis.

The SPIDEX tool is downloaded as a tab-indexed tar file. The contents of the file was queried using the *tabix* utility provided by the htslib C library (formerly packaged with samtools). Example query and output follows:

```
tabix spidex_public_noncommercial_v1.0/
    spidex_public_noncommercial_v1_0.tab.gz chr1:0 | head -10 |
    cut -f1,2,3,4,5,7
chr1    861181  G       A       0.5983  SAMD11
chr1    861181  G       C       0.3903  SAMD11
chr1    861181  G       T       0.6564  SAMD11
chr1    861182  T       A       0.4921  SAMD11
chr1    861182  T       C       0.4723  SAMD11
chr1    861182  T       G       0.7517  SAMD11
chr1    861183  G       A       -0.0870 SAMD11
chr1    861183  G       C       -0.0874 SAMD11
chr1    861183  G       T       0.1979  SAMD11
chr1    861184  G       A       -0.1115 SAMD11
```

with fields as Chromosome, Position, Reference Allele, Mutant Allele, dPSI, and Gene.

2.2 Pathogenicity Scores with UMD Predictor

Pathogenicity of the given variants were scored with the web tool UMD Predictor. This online tool takes in a list of formatted variants and positions and returns a score for pathogenicity of the variant for each transcript affected. These scores ranged from 0 to 100, with 0 being non-pathogenic and 100 being confirmed pathogenicity. These scores were converted to percentages in order to more closely conform with the splicing scores.

2.3 Gene Networks and Risk Analysis

Gene coexpression networks were constructed in order to take into account gene risk factors, with the reasoning that genes with many coexpression partners are more likely to play an important role in their respective biological pathways; disruption in their splicing would thus lead to pathogenicity. These coexpression networks were created using RNA-seq expression data from the BrainSpan database. A single network was developed for each of eight de-

velopmental periods, since expression values are likely to vary depending on the given stage of life. These developmental periods are as follows: 8 weeks postconception (PCW) to 12PCW, 13PCW to 18PCW, 19PCW to 23PCW, 24PCW to 37PCW, 0 months after birth (M) to 11M, 1 year (Y) to 11Y, 12Y to 19Y, and $21Y^{+}$. Pairwise comparisons were performed using the Pearson correlation metric and all gene-gene coexpression pairs with Pearson correlation coeffections of ≥ 0.7 were retained. The number of coexpression partners for each gene was calculated and divided by the maximum of the set, yielding the risk score for that gene. These scores were appended to all variants that affect that gene; the relevant variant information along with the calculated scores were all uploaded to a local MySQL database.

3 Data Analysis

The final dataset to analyze was compiled into a MySQL database with the following layout:

mysql> **DESCRIBE** scored_denovo_db;

Field	Type	**Null**	**Key**	**Default**
PrimaryPhenotype	**varchar**(125)	**NO**	MUL	**NULL**
Gene	**varchar**(125)	**NO**	MUL	**NULL**
Transcript	**varchar**(125)	**NO**	MUL	**NULL**
Chromosome	**varchar**(125)	**NO**		**NULL**
Position	bigint(15)	**NO**		**NULL**
Variant	**varchar**(500)	**NO**		**NULL**
SpliceScore	double	**NO**		**NULL**
PathogenScore	double	**NO**		**NULL**
P1Risk	double	**NO**		**NULL**
P2Risk	double	**NO**		**NULL**
P3Risk	double	**NO**		**NULL**
P4Risk	double	**NO**		**NULL**
P5Risk	double	**NO**		**NULL**
P6Risk	double	**NO**		**NULL**
P7Risk	double	**NO**		**NULL**
P8Risk	double	**NO**		**NULL**

Actual analysis is yet to be performed and will be the subject of this week's work.

My GitHub with relevant code can be found here: https://github.com/kkchau/bimm185/