Potential Transcription Factor Hierarchical System in Autism Spectrum Disorder Cortex Brain Samples

Benjamin Pham

Background:

Autism Spectrum Disorder (ASD) is a complex neurodevelopmental disorder that covers a wide spectrum of behavioral deficits[1]. The severity of the disease varies between afflicted persons in which they may have learning disabilities, communication deficits, or even a combination of these two[2]. The expansion of ADS diagnoses and advances in screening methods rapidly caused ASD diagnoses to be more prevalent[3]. These disorders are expressed uniquely in each afflicted person, which makes treatment an extremely difficult case-by-case basis. These neurodevelopmental disorders are thought to originate from a combination of both environmental and genetic factors that alters neuronal circuitry in the developing brain[4].

Transcription factors (TFs) act as regulatory switches that cause abnormal gene expression in ASD risk genes during early brain development. Previously, a dynamic transcription-factor hierarchical network model was hypothesized in Tsigelny et al. 2013 where "master-transcription factors" regulate dysfunctional co-expressed groups of ASD genes, known as coherent gene groups (CGGs)[5]. These CGGs would most likely contain ASD risk genes and are thought to be expressed during specific brain development periods resulting in a rise to ASD symptoms. Co-expressed genes are regulated by the same regulatory system, are involved in the same biological pathways, and are also functionally related[5,6]. In addition, multiple genes may have the same TF binding motifs and can be co-regulated by common TFs. These TFs can also be regulated by other TFs and even by themselves. Targeting these TFs that regulate CGGs at these key developmental time periods may prevent the manifestation of ASD. Most importantly, the methodology of this study can also be applied to other diseases that are mechanistically similar.

Materials and Methods:

This study identified powerful CGGs in a publicly available RNA-seq count dataset (GSE64018)[7] through Weighted Co-Expression Gene Network Analysis (WCGNA)[8]. This dataset contained 24 cortex brain samples in total evenly distributed as 12 of ASD subjects and 12 control subjects. Subject ages ranged from 15 years old to 60 years old. WCGNA is a form of hierarchical clustering that is usually used to relate phenotypical traits to clusters. In this study, the only "trait" for these samples is the sample type classification as ASD or control. Genes with a positive correlation would have high expression associated with ASD samples while genes with a negative correlation would have high expression associated with the normal samples. CGGs that are highly correlated to either classification in either direction would most likely have differential expression between sample type. Co-expressed gene groups that exist among different aged samples show that their co-expression may be persistent over time. These groups were constructed around extremely interconnected genes defined as "hub genes". The hub genes are significant because the expression of these genes control the CGGs. They are defined as genes with both a high module membership larger than 0.8 and high correlation with sample type greater than 0.6 as recommended in the documentation. The blockwiseModules process was used to automatically form these CGGs. The default settings of deepSplit = 2, signed network, and mergeCutHeight = 0.25 were used. A 9000 maxBlocksize was used to fully utilize the computational environment[8]. In addition, a co-expression module contains a minimum of 30 genes. The top 3 clusters with positive correlation to the ASD sample type were retrieved. Differential expression via DE-Seq2 was also conducted to determine the most probable expression of these CGGs[9]. Identifiable genes in each CGG were also retained. Previously known ASD risk genes recorded in the SFARI Autism gene database are checked for in each CGG[10]. The CGGs were then separately submitted to DAVID for pathway and functional analysis to determine the overall function of each CGG[11]. Significant biological processes were defined as those with an adjusted p-value less than 0.05. The promoter sequence of the hub genes was then retrieved from UCSC genome browser[12]. In this study, promoter sequences of genes are defined as a 2000 base pairs (bps) sequence away from the transcription start site (TSS). MEME (Multiple EM for Motif Elicitation)

was then used to identify TF binding motifs on the hub genes to elucidate common TFs that could bind via TomTom, a MEME-Suite motif search tool that compares query motifs to known motifs from various databases[13,14]. The promoter sequence retrieval and subsequent motif analysis process was repeated for these TFs to build a potential Transcription Factor Hierarchy.

Results:

DE-SEQ2 analysis

Across all 24 samples, there were initially 63,152 genes. Quality control on the gene counts reduced the number of analyzed genes to 53,521 genes. Genes that had an absolute value log2 fold change (log2FC) bigger than 1 and an adjusted p-value less than 0.05 were considered significantly differentially expressed. There were only 885 Differentially Expressed transcripts in which 681 of these were identifiable genes. There was a noticeable skew of differentially expressed genes that were highly expressed in the ASD samples than in the normal samples.

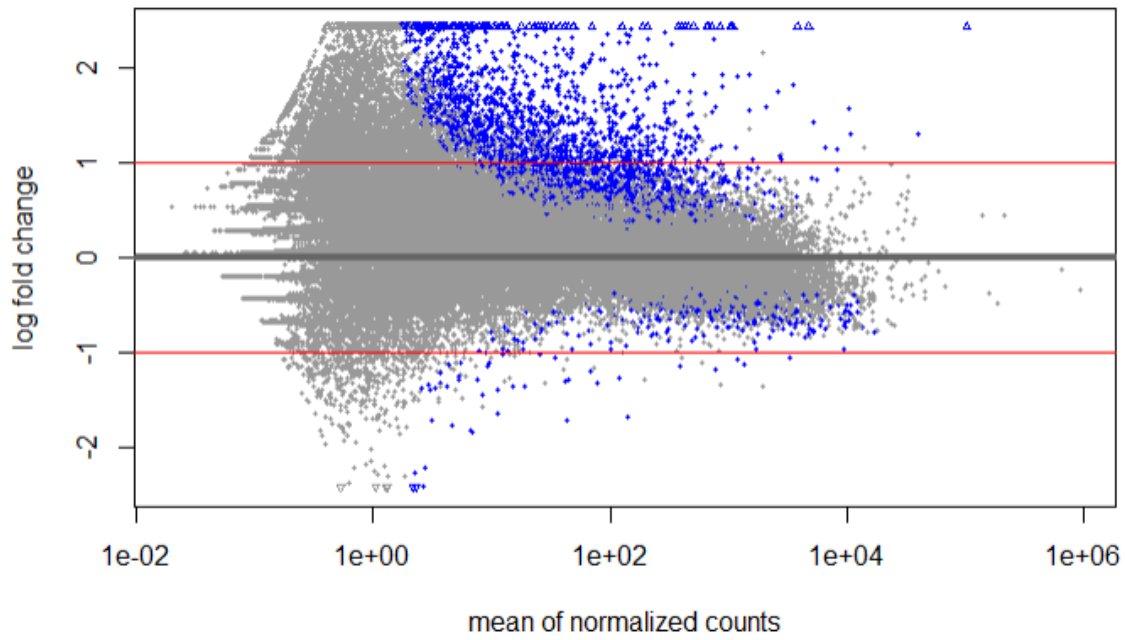

Figure 1. RNA-Seq Count Distribution.

The distribution of all 53,521 genes across the 24 samples are shown. The blue points represent genes with an adjusted p-value less than 0.05. The red lines represent boundaries where anything above abs(log2FC) are considered differentially expressed.

WCGNA Co-Expression Analysis

The 53,521 quality-controlled genes were then clustered via WCGNA. A total of 72 modules were created from these genes across the 24 samples and were labeled by a randomly assigned color. Only the top 3 clusters positively correlated to the ASD classification were able to be analyzed. These correlation plots of genes correlation to the classification vs connectivity with neighbors in the specific module are shown below for each cluster.

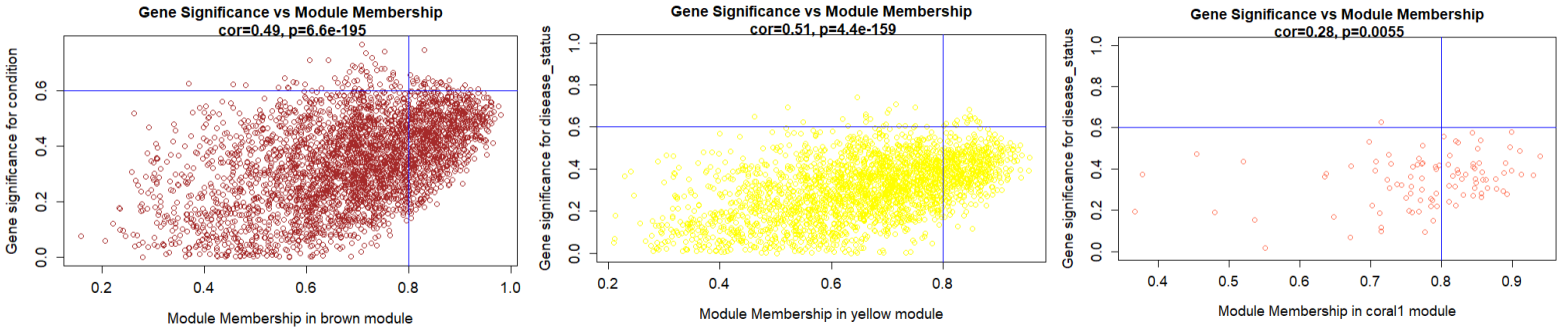

Figure 2. Gene Significance vs Module Membership Plots.

Genes in each cluster are shown correlated to either sample type (ASD and Control) against their Module Membership score. Module membership score shows how interconnected a gene is within the module. The genes within the blue lines are those that are considered "hub" genes with correlation > 0.8 and module membership > 0.6.

The brown cluster contains 2,349 annotated genes out of 3,233 transcripts, the yellow cluster contains 1,414 annotated genes out of 2,400 transcripts, and the coral cluster contains only 80 annotated genes out of 97 transcripts. Out of 1003 ASD risk genes in the SFARI database, ITPR1 was found to be a hub gene in the brown module while RAB43 was found to be a hub gene in the yellow module. The coral module does not contain any of the SFARI validated genes. Interestingly, the membership of differentially expressed genes from the DE-Seq2 analysis in the clusters are sparse. The brown module only contains 138 differentially expressed genes, the yellow module only contains 141 differentially expressed genes, and the coral module only contains 1 differentially expressed gene. Because of the sparse number of genes in the coral module and lack of ASD risk gene as a hub gene, it is discarded. The rest of the analysis considers only the brown module and the yellow module.

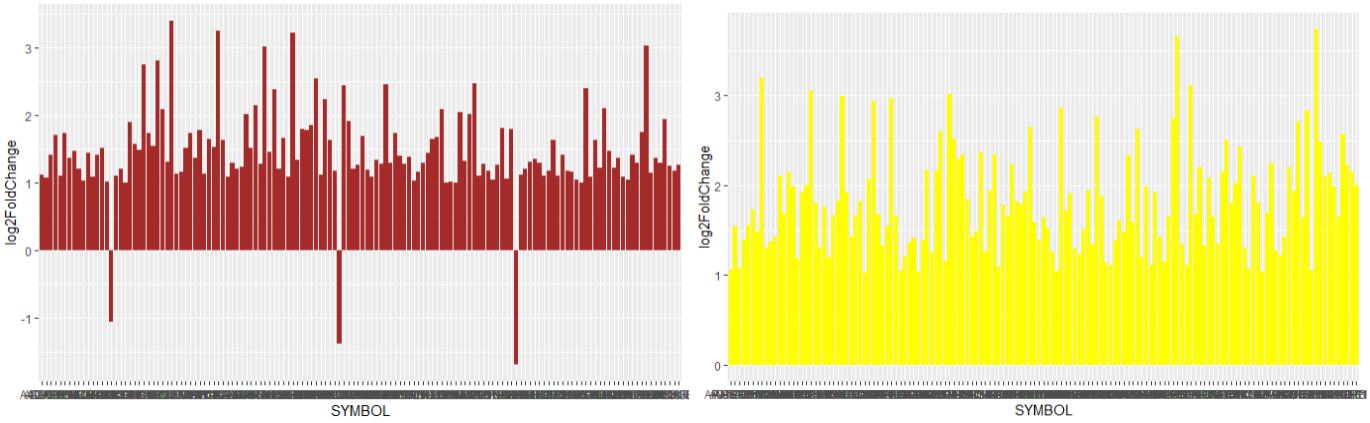

Figure 3. DE gene distribution in Modules of Interest.

The majority of the genes in the brown module have a large positive log2fold change. All genes in the yellow module have a large and positive log2foldchange.

Most genes inside of the brown module have an extremely high log2fold change. All genes inside of the yellow module have a high log2fold change. This indicates that both of these CGGs are more highly expressed in ASD brain tissue samples than in the control brain tissue samples.

Pathway and Functional Analysis

All genes from each module were submitted to DAVID for pathway and functional analysis. In the brown module, DAVID was able to detect 2,209 genes while in the yellow module, DAVID was able to detect 1,296 genes. The only pathway that was enriched in the brown module based on the proposed exclusion criteria is Phosphatidylinositol signaling. The yellow module did not have any pathways that met the statistical significance exclusion criteria.

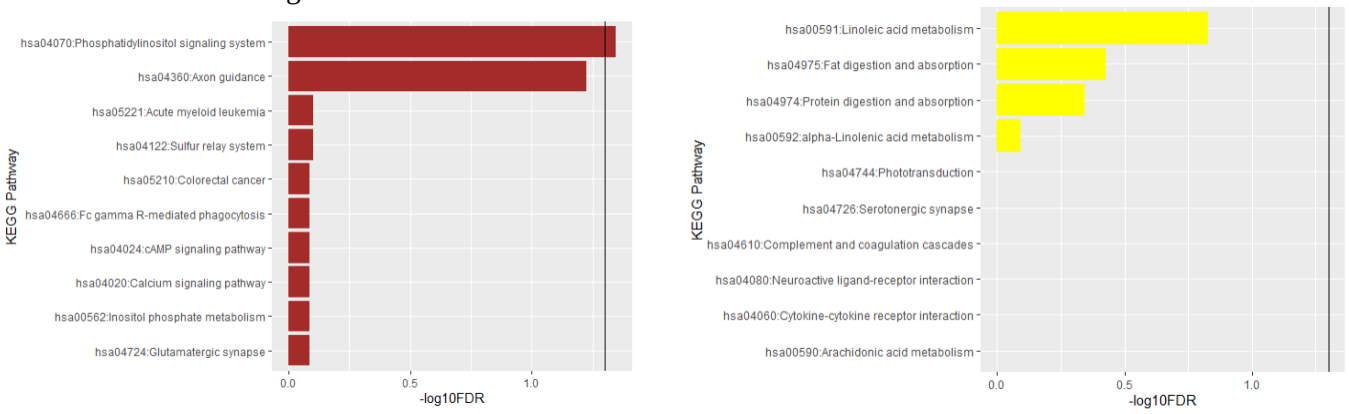

Figure 4. DAVID Pathway Analysis.

A larger -log10FDR signifies smaller FDR adjusted p-value. In the brown module, only the pathway Phosphatidylinositol Signaling System was statistically significant. The pathway Axon Guidance was nearly statistically significant. The other pathways had too large FDR. In the yellow module, no pathways were considered statistically significant.

The 2,000 bp promoter sequence for ITPR1 and RAB43 for the brown and yellow modules respectively was successfully retrieved from the refflat file provided by the UCSC genome browser and cross-validated for accuracy. These promoter sequences were then processed through the MEME-Suite to discover common binding motifs.

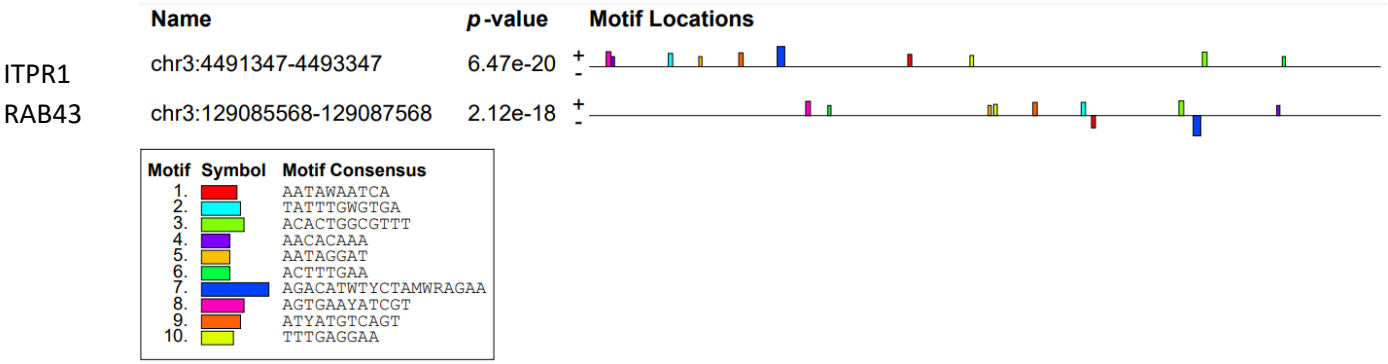

Figure 5. Top 10 Motif locations Identified Across the Hub Genes ITPR1 (top) and RAB43 (bottom).

The 10 motifs seem to be evenly distributed across both promoter sequences. The longer motifs (light green (3) and dark blue(7)) were deeply examined.

The existence of same-colored motifs across the two promoter sequences indicate that these genes could be bound by the same Transcription Factors. The larger motifs, which are colored light green and dark blue, are examined more closely with TomTom to determine if these motifs are already known. The light green motif matches the transcription factor binding motifs of YY1 and YY2. The dark blue motif matches three transcription factor binding motifs: BCL6, BHLHE23, and BHLHE22_DBD complex.

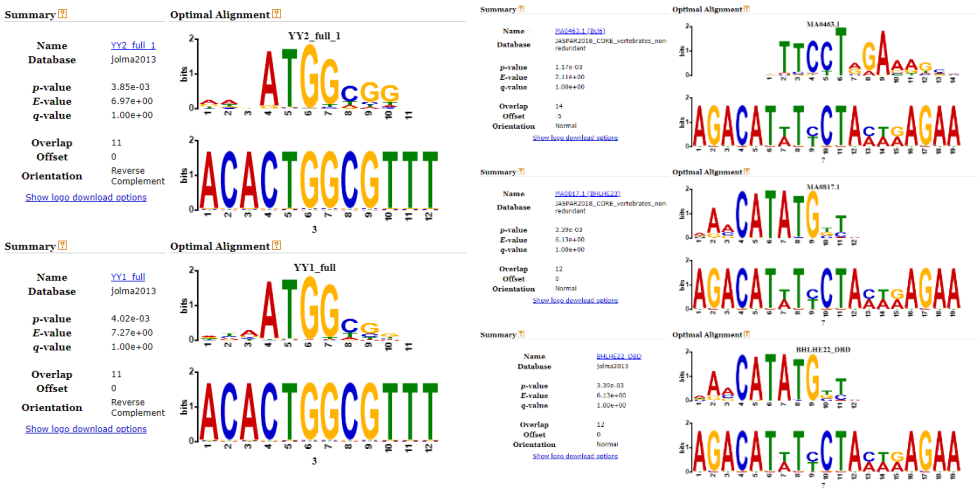

Figure 6. Known Transcription Factor Binding Motifs that are Similar to the Light Green Motif (left) and the Dark Blue Motif (right).

In total, two transcription factors could bind to the Light Green site while there are three transcription factors that could bind to the Dark Blue site.

Preliminary TF Hierarchy Construction

Promoter sequences of the 5 transcription factors shown here are retrieved successfully and run through the MEME-Suite.

YY1
YY2
BCL6
BHLHE22
BHLHE23

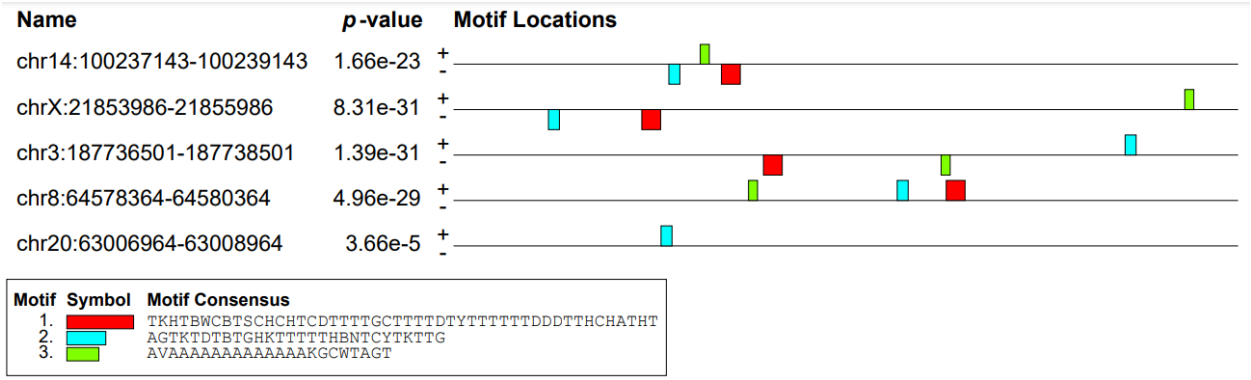

Figure 7. Top 3 Motif Locations on Transcription Factor Promoter Sequences.

The transcription factors are labeled on the left. 4 out of the 5 promoter sequences contain all three motifs. Only BHLHE23 contains one motif type: the cyan motif.

Three motifs across all 5 promoter sequences were queried in which four out of the five transcription factor promoter sequences had the same binding motifs, colored in red, green, and cyan. Only the cyan motif was present in all 5 transcription factor promoter sequences.

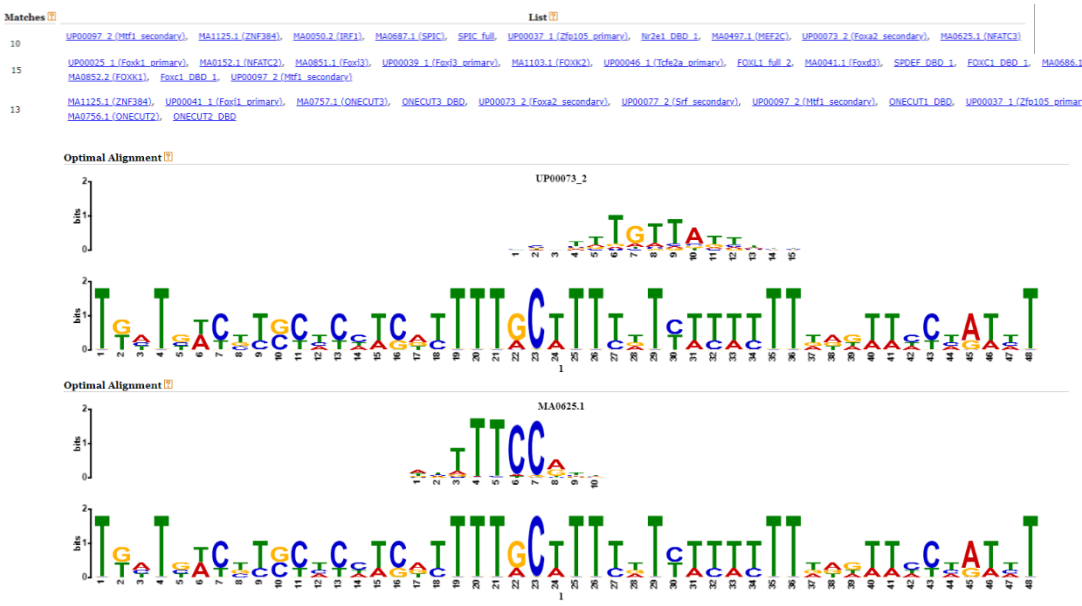

Figure 8. Matched TF Binding Motifs in All Three Query Motifs (top) and Matched Motifs Overlap on Long Query Motif (bottom).

The top table shows the known numerous TF binding motifs that bind to the three elucidated motifs found in the majority of the first-level TF promoters. 10 TFs can align to the first motif, 15 TFs can align to the second motif, and 13 TFs can align to the third motif. The bottom figure is an example of how the known motifs align separately to the long query motif that was discovered in the promoter sequence, an indication of a potential TF complex formation.

Each motif matched multiple transcription factors since the detected motifs are quite long, and these transcription factors are shown to bind to different parts of the query motif. From these analyses, a TF hierarchical regulatory system can be seen.

Discussion:

The work presented in this study show that there might be an underlying TF Hierarchical Regulatory System underlying co-expression modules that may be responsible for ASDs. The formed CCGs are the aggregation of other similar smaller modules. In both the brown and yellow CGGs, there are many genes associated with different biological functions. As a result, there were not a lot of results in the pathway analysis that were statistically significant. Using a more stringent criteria on module formation can differentiate these clusters to separate by functionality.

The discovery of the SFARI hub genes in each module shows the genes in the clusters could also be newly identified ASD risk genes since they are co-expressed with the SFARI hub gene as well. In addition, there are 117 SFARI genes in the brown module while there are 35 SFARI genes in the yellow module. This shows that the formed CGG modules are most likely functionally associated with ASD despite a lack of significant pathways from the DAVID analysis.

The predicted TFs that bind to the modules seem to also be associated with brain development or ASD. For instance, YY1[15,16], YY2[15], and BCL6[17], are known regulators relating to ASD. BHLHE22 and BHLHE23 are known to be involved in the development of cortical brain tissue as well as regulating the development of neuronal circuits[18,19]. The expression direction of the second level TFs that bind to the promoter sequences of these first level TFs are unknown, though they can be inferred based on literature review. For instance, previous studies implicate BCL6 as a co-repressor[20]. Since the CGGs that it regulates are over-expressed, BCL6 is most likely repressed in the regulatory system in the ASD brain. Deletions of the YY1 gene was found in those present in ASD, which could also lead to the prediction that YY1 ,and in extension YY2 since it is in the same TF family, can act as repressors[16]. To determine if

this regulatory system actually exists, separate in-vitro studies need to be conducted. CHIP-seq can be used to determine if the TFs bind to the promoters, and their effect on gene expression can be observed[21].

Conclusion:

This study allows for the formulation of a potential TF hierarchical regulatory system. CGGs highly associated with other validated ASD risk genes can be formed with some ASD risk genes that are found to be highly interconnected in the module. In addition, common TFbs motifs are found to in the promoters of the hub genes and additionally in the promoters of the TFs that govern them which indicates the existence of this system.

Bibliography

1. Hodges, H., Fealko, C. & Soares, N. Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation. *Transl. Pediatr.* **9**, S55–S65 (2020).

2. Shen, M. D. & Piven, J. Brain and behavior development in autism from birth through infancy. *Dialogues Clin. Neurosci.* **19**, 325–333 (2017).

3. Rice, C. E. *et al.* Evaluating Changes in the Prevalence of the Autism Spectrum Disorders (ASDs). *Public Health Rev.* **34**, 1–22 (2012).

4. Chaste, P. & Leboyer, M. Autism risk factors: genes, environment, and gene-environment interactions. *Dialogues Clin. Neurosci.* **14**, 281–292 (2012).

5. Tsigelny, I. F., Kouznetsova, V. L., Baitaluk, M. & Changeux, J.-P. A hierarchical coherent-gene-group model for brain development: A hierarchical gene groups model for brain development. *Genes Brain Behav.* **12**, 147–165 (2013).

6. Fessele, S., Maier, H., Zischek, C., Nelson, P. J. & Werner, T. Regulatory context is a crucial part of gene function. *Trends Genet. TIG* **18**, 60–63 (2002).

7. Irimia, M. *et al.* A highly conserved program of neuronal microexons is misregulated in autistic brains. *Cell* **159**, 1511–1523 (2014).

8. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).

9. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).

10. Banerjee-Basu, S. & Packer, A. SFARI Gene: an evolving database for the autism research community. *Dis. Model. Mech.* **3**, 133–135 (2010).

11. Huang, D. W. *et al.* The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol.* **8**, R183 (2007).

12. Karolchik, D. *et al.* The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**, 51–54 (2003).

13. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202-208 (2009).

14. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biol.* **8**, R24 (2007).

15. He, Y. & Casaccia-Bonnefil, P. The Yin and Yang of YY1 in the nervous system. *J. Neurochem.* **106**, 1493–1502 (2008).

16. Gabriele, M. *et al.* YY1 Haploinsufficiency Causes an Intellectual Disability Syndrome Featuring Transcriptional and Chromatin Dysfunction. *Am. J. Hum. Genet.* **100**, 907–925 (2017).

17. Diaz-Beltran, L., Esteban, F. J. & Wall, D. P. A common molecular signature in ASD gene expression: following Root 66 to autism. *Transl. Psychiatry* **6**, e705 (2016).

18. Dennis, D. J., Han, S. & Schuurmans, C. bHLH transcription factors in neural development, disease, and reprogramming. *Brain Res.* **1705**, 48–65 (2019).

19. Aviel-Shekler, K. *et al.* Gestational diabetes induces behavioral and brain gene transcription dysregulation in adult offspring. *Transl. Psychiatry* **10**, 412 (2020).

20. Shukla, A. *et al.* Variants in the transcriptional corepressor BCORL1 are associated with an X-linked disorder of intellectual disability, dysmorphic features, and behavioral abnormalities. *Am. J. Med. Genet. A.* **179**, 870–874 (2019).

21. Raha, D., Hong, M. & Snyder, M. ChIP-Seq: a method for global identification of regulatory elements in the genome. *Curr. Protoc. Mol. Biol.* **Chapter 21**, Unit 21.19.1-14 (2010).