Matthew Dyer

Introduction

This analysis comes from a data set that analyzed the transcriptome of post-mortem cortex tissue of those with autism spectrum disorder (ASD). 58 samples (29 with ASD, 29 controls) were sequenced with Illumina Ref8 v3 microarrays [1].

ASD is a heritable neurodevelopmental condition that has complex genetic heterogeneity [1]. The authors of the primary paper [1] (which paper I specifically avoided reading until I finished my analysis) found evidence that some genes do have a causal effect on autism but also found evidence that refuted this. My only intention in analyzing this data set was to reveal any interesting genes or pathways in those with ASD. Any genes/pathways that they found that I also found would be considered a serendipitous discovery. Moreover, my goal was to use as many tools as I could from past work in my exploratory analysis of other projects. To find a “smoking gun” biomarker/s would be incredible, but I am skeptical that my junior analysis would find anything that hasn’t already been thoroughly researched (or is a false discovery).

This paper relied upon using brain cortex tissue that can only be acquired post-mortem. This presents a tremendous number of potential pitfalls and difficulties. The thanatotranscriptome (the transcriptome after death) faces intuitive issues, like the known issues of mRNA stability which have a ranged half-life of minutes to weeks [2]. Moreover, the environment of the body is a variable that can distinctly shape the thanatotranscriptome (in mice) [3]. The authors of this paper, and others, account for this by using controls that are age and sex-matched, but I haven’t yet seen anything about the environment of the body after death. Moreover, the condition of the body before death plays a role in the transcriptome. Someone who dies in hospice will likely have transcriptomic differences than those who experience traumatic death such as a car crash, or stroke. The organizations that maintain the database from which the samples were drawn are exemplary, so I’m sure that some, or maybe all, of the issues I’ve thought about are accounted for [4]. However, I couldn’t find specific answers to my questions.

Data Analysis

I imported and normalized the data using an R script that the primary paper’s authors provided [5]. I removed samples that were not specifically cortex tissue (21 samples) as previously mentioned. I next performed Significance Analysis of Microarrays (SAM) in R using the “samr” package and minimized for false discoveries with a 0.0 "med false pos". This gave me ~475 DEGs. 121 DEGs were upregulated in those with ASD and 354 were downregulated. In trying to understand how the patients clustered based on these DEGs (Fig. 1), I found that the clustering was not as “black and white” as one might have thought. However, this is consistent with what the primary paper’s authors discovered (see Fig 1A in primary paper).

To obtain a more diverse set of genes than those identified via SAMr, I employed a Random Forest (RF) to find the most important genes for ASD identification. When I ran

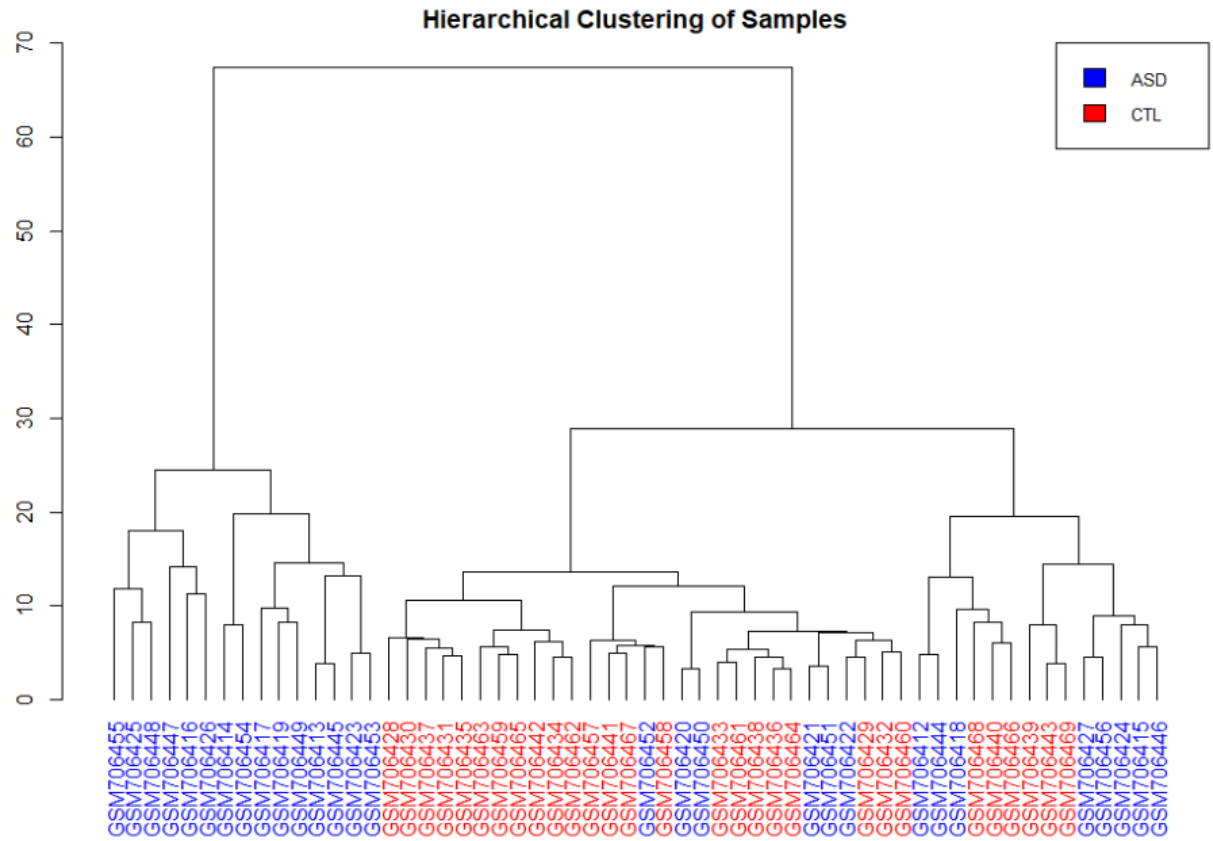

Figure 1. Hierarchical clustering of samples using the 475 DEGs identified via SAM. Ward.D2 and euclidean options were chosen. Note that two main clusters are observed, with one showing a cluster of just those with ASD. The other main cluster has two subclusters that indicate heterogeneity of gene expression, even within DEGS.

my RF on the entire dataset (9163 genes), my RF had an accuray of 81.03%. When I reduced the noise (and technically began to overfit the model) by running the model with the top 500 most important genes identified by the original RF model, the accuracy increased to 86.21%. It was possible to get accuracies as high as ~93% by further subsetting the gene pool, but I wanted to keep a gene pool of around 200-500 for final analysis.

I took the genes identified by the optimized RF model and found that 213 overlapped with the SAMr identified DEGs. These were the genes that I used for annotation and pathway analysis. I did this because I thought it struck a nice balance between genes that are explicitly statistically significant (SAMr) and those that are most important for classification accuracy (RF). I had to convert the gene names to their symbol names using the “AnnotationDbi” and “illuminaHumanv4.db” libraries. I began my gene set enrichment analysis (GSEA) in R by using the following MSigDB gene sets: H, C2, C2:CP:KEGG, C2:CP:REACTOME, C5, C5:GO:BP, C5:GO:CC, C5:GO:MF. I filtered results for only those pathways that had a padj < 0.05 (to avoid the problem of multiple hypothesis testing). This yielded 40 statistically significant enriched pathways. Of those 40, a few were more interesting than others (see table 1).

Pathway	Padj	Leading Edge Genes
REACTOME_NERVOUS_SYSTEM_DEVELOPMENT	0.03520996	RPS2, FYN, RPLP1, DPYSL3, RPL12, RPS15A
HALLMARK_TNFA_SIGNALING_VIA_NFKB	0.0003799	JUN, CEBPD, MCL1, ABCA1, IER3, CFLAR
GOBP_CELLULAR_RESPONSE_TO_STRESS	0.01649157	JUN, SDF2L1, YBX1, GADD45G, HSPB1, HSPA1A

Table 1. Three relatively arbitrarily chosen “interesting” significantly enriched pathways and the first 6 genes involved in the pathways.

REACTOME_NERVOUS_SYSTEM_DEVELOPMENT: This pathway is directly related to the development of the nervous system and could be crucial in understanding the developmental aspects of ASD.
HALLMARK_TNFA_SIGNALING_VIA_NFKB: TNF-alpha signaling via NFkB is involved in inflammatory responses. Inflammation and immune system dysregulation have been implicated in ASD [6].
GOBP_CELLULAR_RESPONSE_TO_STRESS: Stress responses, including oxidative stress, have been studied in the context of ASD, as they can affect neural function and development [7].

Next, I performed weighted correlation network analysis (WGCNA) on the same set of 213 genes via R using the WGCNA library. This yielded 5 distinct modules (Fig. 2). All five modules (yellow, brown, green, turquoise, blue) are significantly associated with the ASD phenotype. Modules Green, blue, yellow, and turquoise are negatively correlated with the ASD phenotype. Brown was the only one positively associated with the phenotype (code not shown but available). In exporting the network data to Cytoscape, I had to use a rather low threshold value of 0.21 to obtain a network that was small enough to manage on my computer, but not so small that there was no useful networking to analyze (Fig. 3).

My analysis with Cytospace revealed that the most connected genes, or hub genes, were INA, RGS7, GPRASP2, NSG1 (table 2). According to ncbi.nlm, all but NSG1 are preferentially expressed in brain tissue. Moreover, Cytoscape’s built-in enrichment feature (defaults used) showed that GPRASP2 was involved in a single pathway (HPA:0100211, p-value 0.022530285) that is involved with cerebral cortex tissue. INA is involved in 14/30 pathways and a majority relate specifically to brain cell components and the cerebral cortex. RGS7 is involved in 6/30 pathways and a majority of which relate to the same tissue and cell type as INA. The one outlier (and maybe a our smoking gun?) was NSG1, which is not preferentially expressed in brain tissue [8].

Probe ID	Common Gene Symbol
ILMN_1673704	INA
ILMN_1685496	RGS7
ILMN_1754727	GPRASP2
ILMN_1772627	NSG1

Table 2. Conversion of PROBE IDs to common gene symbols in four most networked hub genes.

NSG1 is involved 3 pathways according to our enrichment via Cytoscape. One relates to “The junction between an axon of one neuron and a dendrite of another neuron”. The second relates to “A cellular component that forms a specialized region of connection between two or more cells.” The third relates to "A synapse that uses glutamate as a neurotransmitter”. Network analysis by STRING [9] (not shown) helps to illustrate the importance this protein plays in brain development and activity. From the description of STRING [9]: *[NSG1 plays] a role in the recycling mechanism in neurons of multiple receptors, including AMPAR, APP and L1CAM and acts at the level of early endosomes to promote sorting of receptors toward a recycling pathway. Regulates sorting and recycling of GRIA2 through interaction with GRIP1 and then contributes to the regulation of synaptic transmission and plasticity by affecting the recycling and targeting of AMPA receptors to the synapse.*

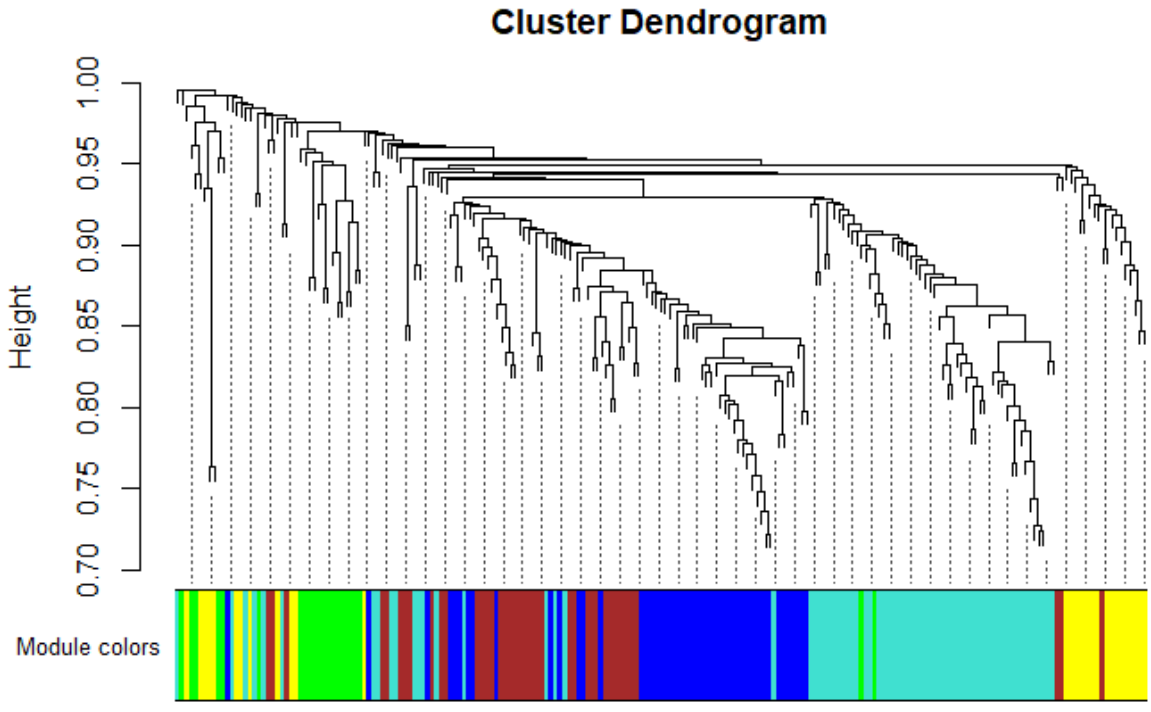

Figure 2. Dendrogram of WGCNA modules of the 213 subset of genes. I was elated to find a dendrogram that wasn’t completely dominated by turquoise (like last time).

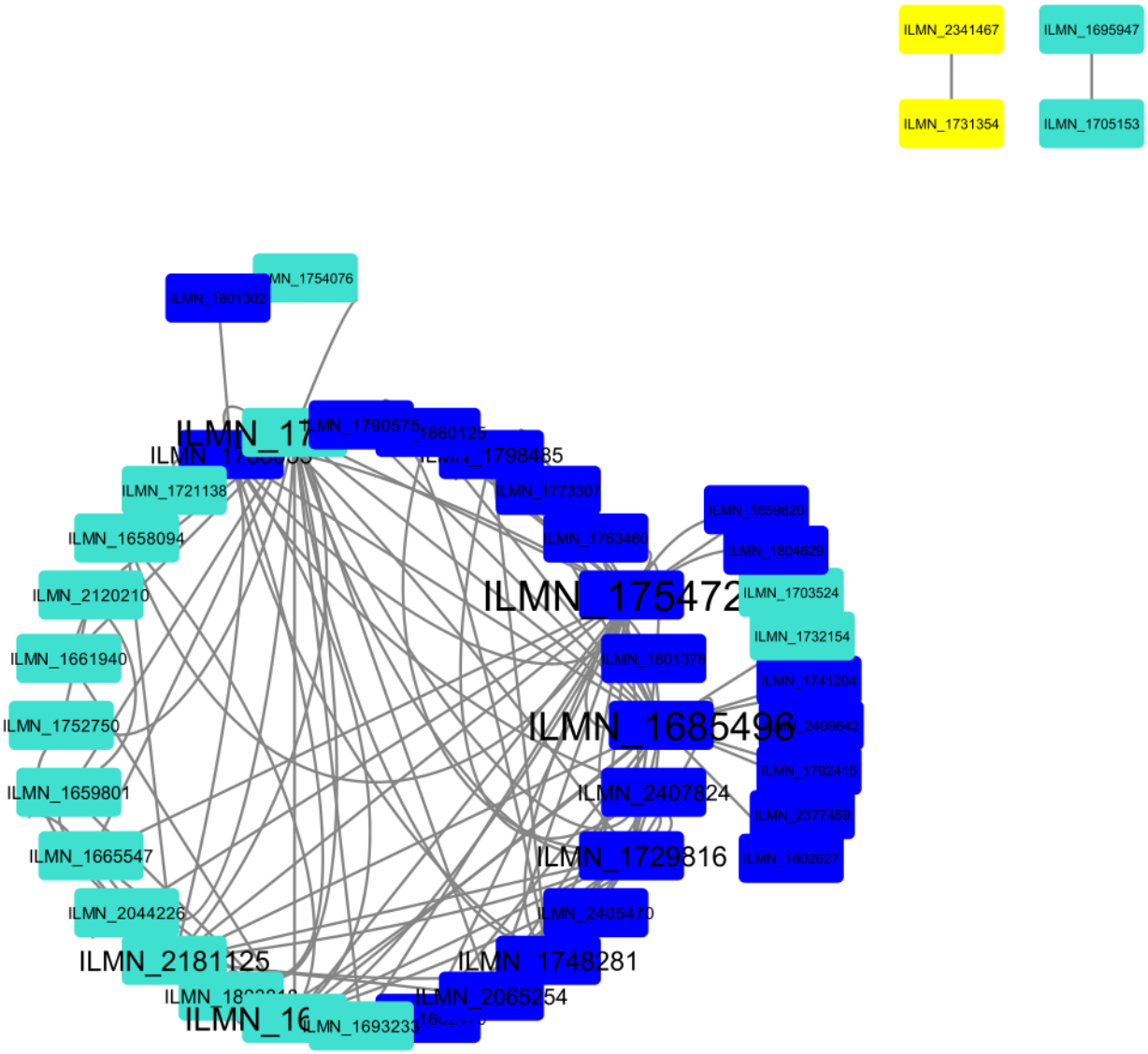

Figure 3. Circular layout of the 213 genes that exceeded the threshold value for WGCNA. Genes with larger label text are indicative of more connections (“hub genes”). Colors indicate module. All genes shown are negatively correlated with ASD phenotype. NSG1 is hidden near the top left of the main circular network.

Conclusion

The fact that NSG1 was the only hub gene that wasn’t preferentially expressed in brain tissue and that it is negatively correlated with ASD is intriguing as a biomarker. I cannot find any published evidence that NSG1 expression is associated with ASD. It wasn’t mentioned in the primary paper either. In fact, the only evidence I can find is on Genebass. When I search the phenotypes for “mental” for the NSG1 gene, I can find non-statistically significant associations with “Mental health problems ever diagnosed by a professional” and a primary loss of function, missense, or synonymous mutations in this gene [10]. Autism is included in this phenotype, but

not the only diagnosed mental health problem in this category. It was not possible for me to filter any SNPs of any specific mutation type in Genebass for autism specifically.

I searched my GSEA pathways and not one contained NGS1. I don't think that especially means anything, but I just thought it was interesting it wasn't in a single pathway. It only became a gene in a pathway after WGCNA and Cytoscape enrichment analysis.

To follow up this research, I would like to analyze transcription factor activity that directly affects NSG1 expression. I want to understand if NSG1 expression is causal in ASD or a result of downstream effects. I also do not have a strong enough understanding of molecular biology and molecular mechanisms to appreciate how decreased expression of NSG1 might affect brain development or function. I would like to reach out to experts to get their understanding of how they might imagine NSG1 under expression in those with ASD might make, or not make, intuitive sense.

Lastly, I think this attempt at analyzing transcriptome data was much more successful than past attempts not because the data set was richer or higher quality, but because I was able to learn from my past attempts. Using a hybrid approach of RF and SAMr helped to reduce some noise (and maybe some important data - I honestly don't know) helped make WGCNA much easier to interpret/handle and I think I found some interesting biomarkers. I don't believe that 58 total samples are enough to make population-level inferences, but post-mortem brain tissue in those with ASD is not as easy to acquire as a blood sample at a clinic. I think this study is ultimately underpowered but with the right meta-analysis, and possibly the right biobank access (GWAS, for additional confirmation), there might be enough statistical power to more confidently claim that NSG1 expression is negatively correlated with ASD.

Materials and Methods:
R version 4.3.2 (2023-10-31 ucrt)
R studio version 2023.6.0.421
Cytoscape Version: 3.10.1

Sources:

1. Voineagu, I., Wang, X., Johnston, P. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011). https://doi.org/10.1038/nature10110
2. Scott L, Finley SJ, Watson C, Javan GT. Life and death: A systematic comparison of antemortem and postmortem gene expression. Gene. 2020 Mar 20;731:144349. doi: 10.1016/j.gene.2020.144349. Epub 2020 Jan 11. PMID: 31935499.
3. Bonadio RS, Nunes LB, Moretti PNS, Mazzeu JF, Cagnin S, Pic-Taylor A, de Oliveira SF. Insights into how environment shapes post-mortem RNA transcription in mouse brain. Sci Rep. 2021 Jun 21;11(1):13008. doi: 10.1038/s41598-021-92268-y. PMID: 34155272; PMCID: PMC8217559.
4. Overview - Autism BrainNet. (2023, December 12). Retrieved from https://www.autismbrainnet.org/overview

5. https://www.ncbi.nlm.nih.gov/geo/geo2r/?acc=GSE28521
6. Siniscalco D, Schultz S, Brigida AL, Antonucci N. Inflammation and Neuro-Immune Dysregulations in Autism Spectrum Disorders. Pharmaceuticals (Basel). 2018 Jun 4;11(2):56. doi: 10.3390/ph11020056. PMID: 29867038; PMCID: PMC6027314.
7. Dong D, Zielke HR, Yeh D, Yang P. Cellular stress and apoptosis contribute to the pathogenesis of autism spectrum disorder. Autism Res. 2018 Jul;11(7):1076-1090. doi: 10.1002/aur.1966. Epub 2018 May 15. PMID: 29761862; PMCID: PMC6107407.
8. NSG1 neuronal vesicle trafficking associated 1 [Homo sapiens (human)] - Gene - NCBI. (2023, December 13). Retrieved from https://www.ncbi.nlm.nih.gov/gene/27065
9. NSG1 protein (human) - STRING interaction network. (2023, December 13). Retrieved from https://version-12-0.string-db.org/cgi/network?networkId=bWhwfRBzUryV
10. Genebass. (2023, March 23). Retrieved from https://app.genebass.org/gene/ENSG00000168824?burdenSet=synonymous&phewasOpts=1&resultLayout=full