OMICS RESEARCH SYMPOSIUM

Coding an App in R using scRNA-seq fetal brain samples for the identification of potential autism risk genes

Andrew Lee

To see the figures and text more clearly, please zoom in on the view tab on ppt, zoom in on browser or manually enlarge the figures. Or please go to this link to see a google doc with all the enlarged versions: https://docs.google.com/document/d/1w13uk4FSeQB1amX1KReQayXHqPmvz0maPBUUpxAlf6w/edit?usp=sharing

Introduction

Autism Spectrum Disorder is a neurological condition that causes limitations in social communication, repetitive and compulsive actions, and declines in cognitive ability appearing early in childhood. There is a high prevalence of autism with approximately 1% of the total population affected by the disorder.

The current method of diagnosis is through clinical evaluation. However, the similar clinical symptomatology of various neurological disorders leads to common misdiagnosis, which is a barrier for the creation of effective individualized treatment regimens. Inaccurate diagnosis can result in suboptimal education and occupational therapy programs and ineffiectve medication prescriptions or adverse effects. Delayed diagnosis has significant consequences, as early intervention may be more effective and durable as brain plasticity is more pronounced at the early stages of childhood.

Current Methods of Diagnosis: DSM-V and Genetic Testing

DSM-V

The Diagnostic and Statistical Manual of Mental Disorders fifth edition (DSM-V), the gold-standard diagnostic criteria used by physicians, is the main method of diagnosis. The DSM-V diagnostic criteria also is thought to oversimplify the disorder phenotypes with overly broad explanations for both cognitive ability and behavioral criteria. As the DSM-V is not optimal and clinical evaluation is particularly difficult for young children, genetic testing may enhance diagnostic accuracy.

Genetic Testing

Genetic testing depends on the screening of autism risk genes to diagnose the disoder. Recent studies estimate that a thousand or more genes cause autism when mutated, but there have been only around 100 genes that have been identified and confirmed to date.

Current research has attampted to identify individual genes that have mutations contributing to autism. Advances in whole exome and genome sequencing in the last decade have created the groundwork for data science analysis.

Network based appaorches have included gene interaction networks, cell specific gene expression profiles and human brain region expression patterns.

Machine learning is a more recent approach that is used to efficiently identify genes that have similar expression profiles to already discovered autism genes. This method incorporates previous gene expression analysis.

Purpose

This project's aim is two fold. First I seek to identify genes of interest from an unsupervised analysis approach. Then there can be a single gene approach to identify genes that have similar expression profiling among distinct cell clusters. This twofold approach will reduce some of the coincidental connections created through the unsupervised methods.
The next purpose is to identify the functional characteristics of the isolated genes in terms of their molecular functions and biological processes linked to the genes.
Additionally, I will facilliate my reasearch and the research of others by creating a publicly available app with a user friendly interface such that the input of the name of a gene of interest results in clear expression levels in more than 20 types of human cells.

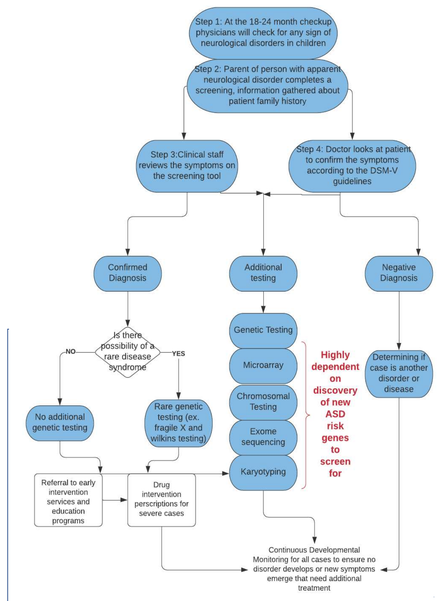


Fig 1: Steps to diagnose autism spectrum disorder.

Methods

Datasets

Arisk scRNA-seq dataset (Chen et al., 2020)

The dataset was composed of two groups: one was of human fetal brain sample expression readings between 6-11 weeks of age, while the second was from prefrontal cortex sample expression readings between 8-26 weeks.
The datasets were integrated to create an expression matrix (expression levels at distinct time points) with 95 total features.
Preprocessing and exclusion criteria: cell clusters ≥10 cells were removed

Identification of possible autism related genes

A heatmap and principal component analysis (PCA) with the Arisk dataset was performed. Several plots were created for single gene user inputs.
Both plots were annotated with 202 genes, 121 from the SPARK database of known autism risk genes, while the rest were graphed to investigate their clustering on PC1 and PC2 in relation to the autism genes.

Heatmap: The heatmap was colored red for the SPARK.risk genes, and blue for the control.LGD genes for the potential autism risk genes with hierarchical clustering, while the functional scores are color coded on the actual plot.

Several genes were selected for single gene plot analysis to determine if the gene shares simliar expression pattern sot the known autism risk genes based on their close relationship to the autism risk genes on PC1 and on the heatmap.
.
App Creation
The app was coded in R using the Shiny and Shiny dashboard packages. Some of the code was modified from the OmicsLogic/BioTech Transcriptomics course, while several of the plots were created using the Seurat and associated packages.

Further analysis of genes using single gene plots in app

t-distributed stochastic neighbor embedding (t-SNE) plot. Clusters that have the expressed gene are highlighted in blue depending on the expression levels.
Violin Plots. This plot uses the single gene input to immediately generate. Expression levels for the gene name input are mapped among more than 20 cell types/clusters.
Ridge Plot: This was an alternative to the single gene t-SNE.

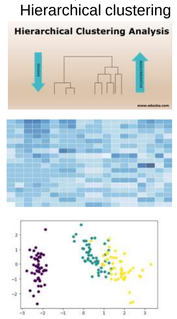
Annotations: All of the above plots were annotated through the metadata file with distinct cell clusters and the standard color R code.

Expression levels were compared to those of known autism risk genes. Genes with similar expression profiling in identical cell types were identified as possible candidate autism risk genes.

a. Expression matrix input

	hDA0:week_10	hDA0:week_9
DDX11L1	0.0	0.22944216
WASH7P_p1	0.3836881	0.0
LINC01002_loc4	0.0	0.0
LOC100133331_loc1	0.18377306	0.0
LOC100132287_loc2	0.0	0.0
LOC101928626	0.0	0.09586797
MIR6723	0.67272115	0.15172525
LOC100133331_loc2	0.0	0.0
LOC100288069_p1	0.0	0.09586797
LINC00115	0.068216935	0.0
LINC01128	0.3038853	0.14457226
FAM41C	0.14606576	0.0
SAMD11	0.0	0.0
NOC2L	0.11181325	0.34196487

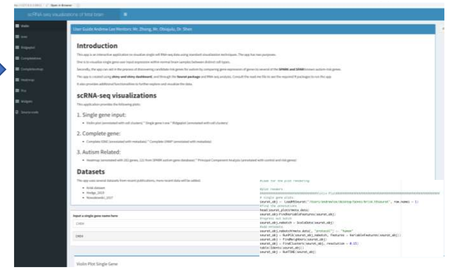
b. Heatmap, PCA and Hierarchical clustering



c. List of known and possible autism genes + known genes

HGNC	conditions
ADNP	SPARK.risk
ANK2	SPARK.risk
ANKRD11	SPARK.risk
ARID1B	SPARK.risk
ASH1L	SPARK.risk
ASXL3	SPARK.risk
BAZ2B	SPARK.risk
...	

d. Coded app in R, analyzed possible autism genes using t-SNE, violin plot, ridge plot for similar expression patterns to known, eliminating genes that are not expressed in similar cell types



e. Final revised list of possible candidate genes + known genes that cause autism

HGNC	conditions
ADNP	SPARK.risk
ANK2	SPARK.risk
ANKRD11	SPARK.risk
ARID1B	SPARK.risk
ASH1L	SPARK.risk
ASXL3	SPARK.risk
BAZ2B	SPARK.risk
...	

Fig 2. Schematic of the stpes in this project. Input: gene matrix. Data undergoes initial analysis on complete data to isolate individual genes. Second set is done manually using the gene input name into the app that was coded in R to find similar expression patterns with autism genes

Results

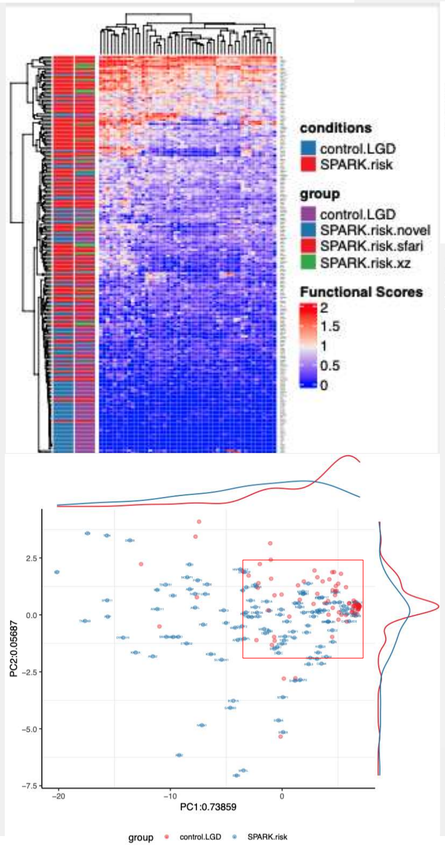
For the initial step to identify autism risk genes, unsupervised methods were performed. Both plots were annotated with SPARK risk genes, which were the genes that were known to cause autism from previous studies with verification. Next, genes with expression levels from the A Risk dataset were plotted.

The hierarchical clustering of autism risk genes with control genes (genes that are being tested for their autism correlation) revealed several genes that may be implicated.

On the Principal Component Analysis, it is apparent that the region boxed in red is enriched for autism genes. This region was searched for other genes that could be linked to autism.

All of the genes were mapped based on their expression levels and functional scores. The genes clustered on the heatmap and the PCA have similar expression to the autism genes, and therefore are likely to be autism risk genes themselves.

This method of identifying genes based on the cell clusters with other autism risk genes has been used in previous studies to circumvent the labor-intensive direct functional analysis and specific gene studies that are traditionally used to identify risk genes and genes that cause the neurological disorder.



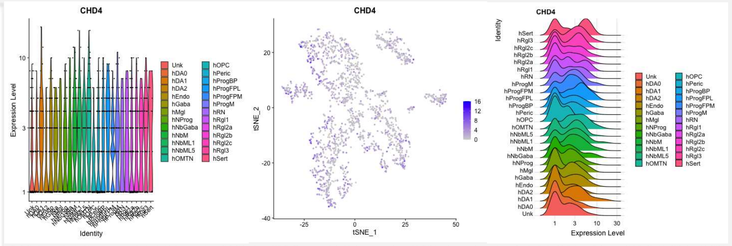


Fig 3: Single gene plots in shiny app generated using A risk dataset option. t-SNE, Ridgeplot and Violin Plot were coded with the seurat package with CHD4 as the example gene input. Graphs display 26 distinct cell clusters. a) Violin Plot b) t-SNE plot with CHD4 cells colored in blue c) Ridge Plot

Single gene plots were created for the genes that were clustered next to autism risk genes from the heatmap and the Principal Component Analysis.
Genes that have similar expression patterns based on the expression level across the color coded cell clusters were identified as potential autism risk genes. The plots above show what the expression profile for a single gene using CHD4 as an example would look like. This process was repeated for each of the genes from the heatmap and PCA.

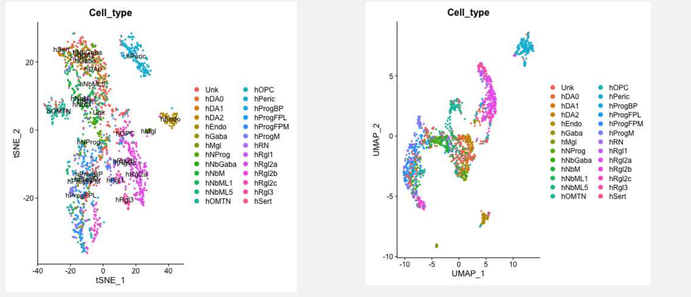


Fig 4: Additional plots that were used purely for visualization, not related to autism gene discovery. UMAP plot as an alternative to the t-SNE on the right, the complete gene t-SNE is shown on the right

These are additional plots that were created in the app with some additional modifications from the Omics Logic course. Note the plots are added for visualization of data and do not have a direct link to the discovery of autism risk genes. There are clear clusters based on cell type, indicating that the genes are expressed in 26 distinct cell types. The UMAP is another alternative to the t-SNE plot.

genes	HGNC	conditions	group	labels
PDS5A	PDS5A	control.LGD	Potential autism risk gene	
IARS	IARS	control.LGD	Potential autism risk gene	
SLIT1	SLIT1	control.LGD	Potential autism risk gene	
APPBP2	APPBP2	control.LGD	Potential autism risk gene	
SNAP91	SNAP91	control.LGD	Potential autism risk gene	

After removing genes that are not highly expressed or are not expressed in autism related cell types in the brain, **5 genes were identified as potential autism genes and will be further studied.**

These genes share similar expression profiling and are highly expressed in the brain samples.121 other genes were also highly expressed in the dataset, but these were already confirmed autism genes.

Conclusions

Research Conclusions

The coded app provides ease of use with automatic plot rendering and image downloads so that the user can visualize the single gene expression data across multiple cell types. The user interface only requires the input of the name of a single brain-specific gene. The app is useful for other researchers seeking to correlate genes of interest from their own studies to multiple cell clusters and cell types.

In this study, the application was used to search for genes correlated with autism.

On the heatmap and Principal Component Analysis, there was apparent clustering of autism risk genes into distinct regions of the PC1 and on the hierarchical heatmap clustering.

Five genes were identified as potential autism genes that may be markers of autism or directly cause autism when mutated.

Impact

The discovery of autism risk genes can improve the genetic testing available for autism, which in turn provides more accurate and clear diagnosis. Early intervention is one of the benefits of genetic testing diagnosis, as among the benefits are more accurate learning program designation and better drug treatment options.

Genetic testing has been limited by the lack of autism risk gene discovery and the amount of datasets available, but this study has provided several genes for further study

Limitations

Determination of autism genes based on gene expression clustering and unsupervised methods has its pitfalls.

In some cases, brain expression similarities can be coincidental, and functional analysis is needed to verify any of the genes identified. Supervised machine learning approaches and neural networks may be better for gene discovery to reduce the risk of false positives.

Future Research

This study has proved useful in identifying five candidate autism genes for further study regarding their biological and molecular functions.

Extensions to this project involve continuing adding more recent and larger datasets to use in the unsupervised machine learning methods and in the app.

Lastly, improving the user interface of the app such as adding new interactive data tables and multiple gene file inputs would allow for other researchers to visualize the single and multiple gene expression profiles for their own studies in more detail.

References

[1] D. Walzl, A. J. Carson, and J. Stone, "The misdiagnosis of functional disorders as other neurological conditions," *Journal of Neurology*, vol. 266, no. 8, pp. 2018–2026, 2019.
[2] S. Aggarwal and B. Angus, "Misdiagnosis versus missed diagnosis: diagnosing autism spectrum disorder in adolescents," *Australasian Psychiatry*, vol. 23, no. 2, pp. 120–123, 2015.
[3] "Pharmacologic treatment options for children and adolescents with autism spectrum disorder," *Clinical Pharmacist*, 2017.
[4] M. A. Waldrop and S. J. Kolb, "Current Treatment Options in Neurology—SMA Therapeutics," *Current Treatment Options in Neurology*, vol. 21, no. 6, 2019.
[5] J. M. Berger, T. T. Rohn, and J. T. Oxford, "Autism as the Early Closure of a Neuroplastic Critical Period Normally Seen in Adolescence," *Biological Systems: Open Access*, vol. 02, no. 03, 2012.
[6] K. Ian, W. Jessica, C. Danielle, L. Veema, and H. Jeff, "Evidence of Hyperplasticity in adults with Autism Spectrum Disorder," *Frontiers in Human Neuroscience*, vol. 7, 2013.
[7] S. Calderoni, L. Billeci, A. Narzisi, P. Brambilla, A. Retico, and F. Muratori, "Rehabilitative Interventions and Brain Plasticity in Autism Spectrum Disorders: Focus on MRI-Based Studies," *Frontiers in Neuroscience*, vol. 10, 2016.
[8] H. Hodges, C. Fealko, and N. Soares, "Autism spectrum disorder: definition, epidemiology, causes, and clinical evaluation," *Translational Pediatrics*, vol. 9, no. S1, 2020.
[9] G. Xu, L. Strathearn, B. Liu, and W. Bao, "Corrected Prevalence of Autism Spectrum Disorder Among US Children and Adolescents," *JAMA*, vol. 319, no. 5, p. 505, 2018.
[10] L. E. Vissers, C. Gilissen, and J. A. Veltman, "Genetic studies in intellectual disability and related disorders," *Nature Reviews Genetics*, vol. 17, no. 1, pp. 9–18, 2015.
[11] I. Panigrahi, P. Jain, S. Didel, S. Agarwal, S. Muthuswamy, A. Saha, and V. Kulkarni, "Identification of microdeletion and microduplication syndromes by chromosomal microarray in patients with intellectual disability with dysmorphism," *Neurology India*, vol. 66, no. 5, p. 1370, 2018.
[12] L. S. Weaving, "Rett syndrome: clinical review and genetic update," *Journal of Medical Genetics*, vol. 42, no. 1, pp. 1–7, 2005.
[13] J. P. Ip, N. Mellios, and M. Sur, "Rett syndrome: insights into genetic, molecular and circuit mechanisms," *Nature Reviews Neuroscience*, vol. 19, no. 6, pp. 368–382, 2018.
[14] S. M. Kyle, N. Vashi, and M. J. Justice, "Rett syndrome: a neurological disorder with metabolic components," *Open Biology*, vol. 8, no. 2, p. 170216, 2018.
[15] I. C. Smith, "DSM-5 and Autism Spectrum Disorder," *Encyclopedia of Autism Spectrum Disorders*, pp. 1–6, 2017
[16] M. Solerdelcoll Arimany, "Diagnostic stability of autism spectrum disorders with the DSM-5 diagnostic criteria," 2017.
[17] D. A. Regier, E. A. Kuhl, and D. J. Kupfer, "The DSM-5: Classification and criteria changes," *World Psychiatry*, vol. 12, no. 2, pp. 92–98, 2013.
[18] Y. Liu and M. R. Chance, "Pathway Analyses and Understanding Disease Associations," *Current Genetic Medicine Reports*, vol. 1, no. 4, pp. 230–238, 2013.
[19] A. Alonso-Gonzalez, C. Rodriguez-Fontenla, and A. Carracedo, "De novo Mutations (DNMs) in Autism Spectrum Disorder (ASD): Pathway and Network Analysis," *Frontiers in Genetics*, vol. 9, 2018.
[20] J. A. Veltman and H. G. Brunner, "De novo mutations in human genetic disease," *Nature Reviews Genetics*, vol. 13, no. 8, pp. 565–575, 2012.
[21] G. M. Anderson, "Autism Biomarkers: Challenges, Pitfalls and Possibilities," *Journal of Autism and Developmental Disorders*, vol. 45, no. 4, pp. 1103–1113, 2014
[22] I. Voineagu and H. J. Yoo, "Current Progress and Challenges in the Search for Autism Biomarkers," *Disease Markers*, vol. 35, pp. 55–65, 2013.

I would like to thank Mr. Elia Brodsky, Dr. Harpreet Kaur and my mentors at my Columbia University internship: Mr. Zhong, Mr. Obiajulu, and Dr. Shen