Paradigm Academic Press
Journal of Innovations in Medical Research
ISSN 2788-7022
SEP. 2023 VOL.2, NO.9

Predicting Autism Spectrum Disorder Using Pluripotent Stem Cell RNA-Seq Data and Machine Learning

Richard Li[1]

[1] Highland Park High School, TX 75205, USA

Correspondence: Richard Li, Highland Park High School, TX 75205, USA.

doi:10.56397/JIMR/2023.09.07

Abstract

In this work, datasets of gene expression in Autism Spectrum Disorder (ASD) were analyzed with the goal of selecting the most attributed genes and performing classification with machine learning algorithms. The publicly published datasets (GSE129806 and GSE214323) from the Gene Expression Omnibus database, which are both RNA-seq gene count data of humans, were downloaded. Then the workflows with differential expression analysis, principal component analysis (PCA), gene set enrichment analysis (GSEA) (Subramanian et al., 2005) and gene expression Meta-Analysis (Toro-Domínguez et al., 2020) were developed. The datasets were following pipelines which used machine learning algorithms to develop prediction models for classification. The results of this exploratory study suggest that the gene expression profiles identified from the pluripotent stem cell samples with ASD can be used to identify a biological signature for ASD with machine learning techniques. And especially, the gene expression Meta-Analysis of multiple datasets and larger numbers of samples could lead to more practical tools, such as Machine Learning models and workflows, to detect ASD at an early age in the general population.

Keywords: Autism Spectrum Disorder (ASD), machine learning algorithms, workflows, tools

1. Introduction

According to WHO (World Health Organization), ASD, also referred to as Autism Spectrum Disorder, constitutes a diverse group of conditions related to development of the brain (World Health Organization: WHO, 2023). Autism is characterized by some degree of difficulty with social interaction and communication (WHO, 2023). Other characteristics are atypical patterns of activities and behaviors, such as difficulty with transition from one activity to another, a focus on details, and unusual reactions to sensations. According to CDC (Centers for Disease Control & Prevention), about 1 in 36 children has been identified with autism spectrum disorder (ASD) according to estimates from CDC's Autism and Developmental Disabilities Monitoring (ADDM) Network (Basics About Autism Spectrum Disorder (ASD) | NCBDDD | CDC, 2022). About 1 in 6 (17%) children aged 3–17 years were diagnosed with a developmental disability, as reported by parents, during a study period of 2009-2017. These included autism, attention-deficit/hyperactivity disorder, blindness, and cerebral palsy, among others (Data and Statistics on Autism Spectrum Disorder | CDC, 2023). Identifying autism is difficult before the age of about 12 months but diagnosis is generally possible by the age of 2 years (WHO Autism Q&A). Diagnosing ASD can be difficult since there is no medical test, like a blood test, to diagnose the disorder. Doctors look at the child's behavior and development to make a diagnosis. By age 2, a diagnosis by an experienced professional can be considered reliable. However, many children do not receive a final diagnosis until they are much older (Basics About Autism Spectrum Disorder (ASD) | NCBDDD | CDC, 2022). Some people are not diagnosed until they are adolescents or adults. This delay means that people with ASD might not get the early help they need.

So, early diagnosis of ASD can lead to increased benefits in therapy, personalized treatment, social accommodations and communication. Different types of biomarkers including prenatal history, genetics, neurological, metabolic and nutritional were used for diagnosis (Jensen et al., 2022). While its etiology is complex, ASD has a strong genetic basis (Hallmayer et al., 2011; Jeste & Geschwind, 2014; Colvert et al., 2015). So far there are many different genetic data that have been collected for ASD to be analyzed and used for the purpose of diagnosis. This study aims to find methods to speed and simplify diagnosis. As already been applied in various fields, including image and speech recognition, natural language processing, recommendation systems etc. Machine Learning techniques benefit our lives on a daily basis. Machine Learning is a subset of artificial intelligence (AI) that focuses on developing algorithms and models that allow computers to learn and make predictions or decisions based on data through statistical analysis. It has immense potential to enhance diagnostic and intervention research in the behavioral sciences (Bone et al., 2014), and may be especially useful in investigations involving the highly prevalent and heterogeneous syndrome of ASD. This study used its methodologies and tools aiming to build models to apply a transcriptomic approach using RNA-seq datasets to identify a gene expression signature with promising performances in the diagnostic prediction of ASD.

There are a few challenges in the gene expression analysis of ASD. The first challenge is that there is lots of noise in gene expression level data (Parab et al., 2022), which in general usually occurs due to variations associated with the experiments or the existence of alterations in the genes (Ansel et al., 2017). In the case of autism, the extra variance may be linked to the presence of alterations in many genes. Another challenge is the difficulty in selection and identification of the genes that are most relevant to autism (Selection of Gene) (Rahman et al., 2020). This problem exists because the gene expression levels in ASD show considerable diversity among individuals and because the sequences of several of these genes are highly variable. Another challenge is the limited number of samples (in the range of dozens or hundreds) that have been made in comparison to the very large number of genes (in the range of tens of thousands). In machine learning, this term is known as “high dimensionality”, and sophisticated methods are required to handle it properly.

2. Methods

2.1 Data Source

The publicly available datasets (GSE129806 and GSE214323) were downloaded from the Gene Expression Omnibus database (GEO DataSets — NCBI, n.d.). The two datasets were chosen for this study because they are relatively recent data generated in 2020 and 2023 respectively; and both of them use expression profiling by high throughput sequencing equipments of Illumina HiSeq 3000 and Illumina NovaSeq 6000 respectively; at the same time both of them were generated from stem cells of humans. In addition, they contain relative higher number of samples, which is crucial for machine learning training, compared to several other RNA-seq datasets from Gene Expression Omnibus, e.g., GSE105046 only contains 6 samples, GSE125020 has 15 samples while GSE221923 contains 18 samples.

GSE129806, claimed by the authors that it was the first attempt to model multiplex autism using patient-derived induced pluripotent stem cells (iPSCs), aiming on providing evidence of morphological, physiological, and transcriptomic signatures of polygenic liability to ASD. The dataset is RNA-sequencing of humans induced pluripotent stem cell-derived cortical inhibitory and excitatory neural progenitors for four cell lines from four different individuals with varying autism affectation; four biological replicates per cell line. It analyzes cellular and molecular characterization of multiplex autism in humans induced pluripotent stem cell-derived neurons. The samples totally contain ASDs (n=16) and Controls (n=16).

GSE214323. Alterations in cortical neurogenesis are implicated in neurodevelopmental disorders including Autism Spectrum Disorders. The study aims to provide experimental evidence for the understudied cortical neurogenesis in addition to ASD risk genes. The dataset is RNA-sequencing of humans using isogenic induced pluripotent stem cell (iPSC)-derived neural progenitor cells (NPCs) and cortical organoid models. It reports that a heterozygous PTEN p.I135L mutation found in an ASD patient with macrocephaly dysregulates cortical neurogenesis in an ASD genetic background-dependent fashion. Libraries for each genotype include three independent cell culture replicates and three separate passages. The samples totally contain ASDs (n=36) and Controls (n=18).

2.2 Workflows

In this study two workflows were developed: Single Dataset vs Multi-Dataset with Meta-Analysis as illustrated in **Supplementary Figure S1**. The processing blocks in light yellow were manually performed while the processing blocks in lime and light red were programmatically performed.

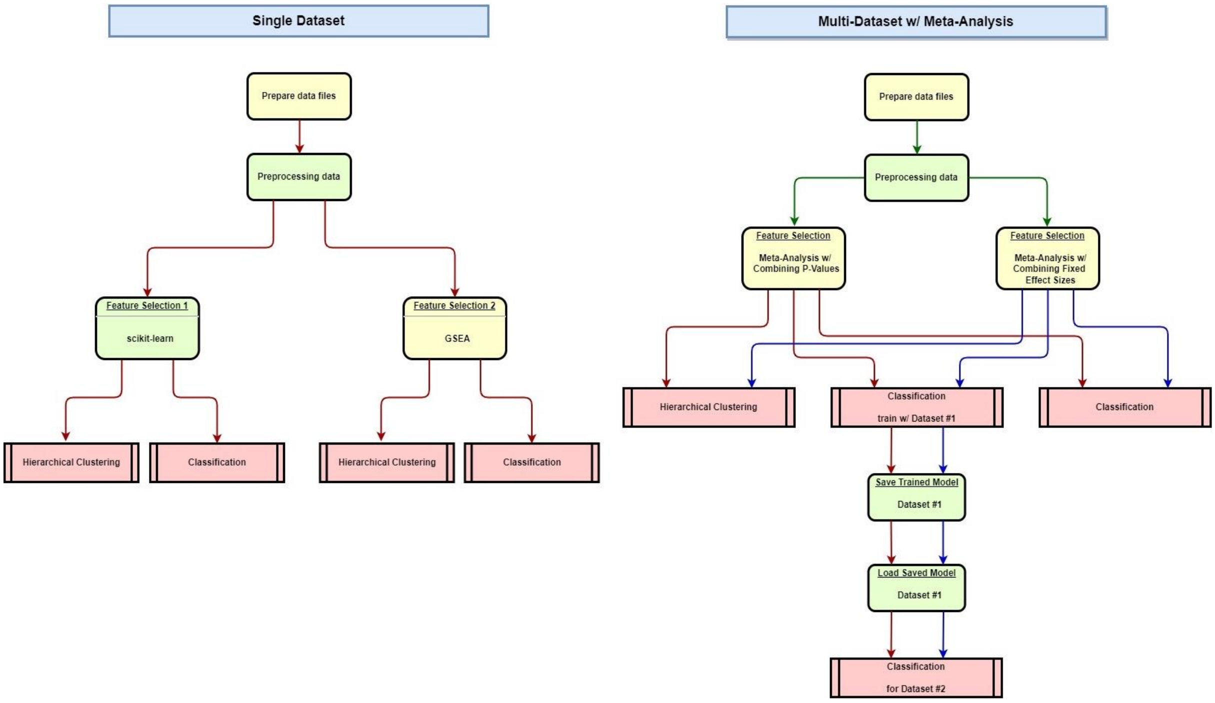

Figure S1.

Each of the workflow shares the similar steps in their processing pipeline. First, the RNA-seq gene count and metadata files were prepared manually from original GSE datasets in the “Prepare date files” step. Second, the preprocessing module of “scikit-learn” was used to scale per gene cross samples programmatically in the “Preprocessing data” step. The “Feature (Gene) Selection” step aimed to handle the “high dimensionality” challenge. There are two approaches were taken: using “scikit-learn” programs in Python and using GSEA (Gene Set Enrichment Analysis) in the “Single Dataset” workflow; and using gene expression Meta-Analysis to obtain the top genes in the “Multi-Dataset with Meta-Analysis” workflow. With the selected features (genes) from the “Feature Selection” step, the “Hierarchical Clustering” analysis and visualization of all samples in each dataset were performed using the heatmap function in the “bioinfokit” package in Python (Reneshbedre, n.d.). For “Classification”, the programs in Python utilizing the classifiers in “scikit-learn” were used to train the datasets for development of prediction models. The classifiers include K-Nearest Neighbors, Stochastic Gradient Descent, AdaBoost and Quadratic Discriminant Analysis. Each dataset was randomly divided into 70% training data and 30% test data for 20 runs. In each run, different classifiers/algorithms were trained and results were evaluated.

2.3 Single Dataset Workflow

The Single Dataset Workflow is illustrated in **Supplementary Figure S1**. In this workflow, a single dataset was used (GSE129806 and GSE214323 were independent of one another). There were two methods used for feature/gene selection: “scikit-learn” programs in Python vs GSEA (Gene Set Enrichment Analysis).

The approach of “scikit-learn” programs preprocesses data by normalizing the data to a unit norm. The preprocessed data was followed by FDR (False discovery rate) correction to ensure only statistically-significant ($p < 0.05$) genes/features remain. Then, the dimension reduction with PCA (Principal component analysis) was applied to determine how many genes (features) would be needed for the downstream processing of Hierarchical Clustering and Classification. PCA was performed using the “Incremental PCA” class (Sklearn. Decomposition. Incremental PCA, n.d.) of the “scikit-learn” package. Its linear dimensionality reduction uses Singular Value Decomposition (SVD) of the data, keeping only the most significant singular vectors to project the data to a lower dimensional space. And the PCA two-dimension visualization was plotted using the “Seaborn” (Seaborn: Statistical Data Visualization, n.d.) packages in Python. To finally select the top genes/features from the PCA, ANOVA (Analysis of variance) which uses the F-test was executed. All the steps were programmed and done utilizing the pipeline feature of “scikit-learn”.

For the approach of GSEA (Gene Set Enrichment Analysis), which is a computational method that determines whether an a priori defined set of genes shows statistically significant, concordant differences between two biological states (e.g., phenotypes). GSEA reveals many biological pathways in common. In this study specifically, GSEA was used to obtain the gene markers ranked by the enrichment scores. GSEA is available for download at https://www.gsea-msigdb.org/ (GSEA, n.d.). Both GSE datasets (GSE129806 and GSE214323) were processed against “GSEA C2-KEGG” Gene Sets (GSEA | MSIGDB | Browse Human Gene Sets, n.d.) manually GSEA and the top 100 genes/features were obtained based on enrichment scores with a 0.05 statistical significance threshold.

These selected top genes/features were then fed to the downstream processing steps of Hierarchical Clustering and Classification. A Hierarchical Clustering analysis and visualization was performed programmatically using the heatmap function in the “bioinfokit” package in Python and the end result was evaluated. In the Classification, the “scikit-learn” programs utilizing the classifiers, K-Nearest Neighbors, Stochastic Gradient Descent, AdaBoost and Quadratic Discriminant Analysis, were used to train the dataset and the development of a prediction model. Each dataset was randomly divided into 70% training data and 30% test data for 20 runs, then different classifiers/algorithms were trained and results were evaluated.

2.4 Multi-Dataset with Meta-Analysis Workflow

The Multi-Dataset Workflow is illustrated in **Supplementary Figure S1**. In this workflow, the two datasets of GSE129806 and GSE214323 were processed manually through gene expression Meta-Analysis for Feature (gene) Selection, and the top 100 genes (features) were obtained with a 0.05 statistical significance then fed to the downstream processing steps of Hierarchical Clustering and Classification. In this study, two methods were used when performing gene expression meta-analysis. The first method is based on combining P values with the “Fisher’s method” (Yoon et al., 2021). The second method is combining fixed effect size, which is a linear model that considers that the different studies share a common effect size called true effect. The **Supplementary Table S12** shows the detailed formulas for each of them. A major tool utilized in this study is “NetworkAnalyst” (NetworkAnalyst, n.d.).

The Hierarchical Clustering and Classification were conducted using the same procedures as the “Single Dataset” workflow. In addition, in this workflow the dataset GSE129806, which shares the same features/genes with GSE214323 from Feature Selection, was first trained using the different classifiers/algorithms. The trained model was then saved into a file using the “Joblib” package (Joblib: Running Python Functions as Pipeline Jobs, n.d.) and later loaded back into the program to predict the samples in dataset GSE214323 which were previously unseen by the model. Finally, the results of the models’ predictions were recorded.

3. Results

3.1 Pathways and Biomarkers

in “Feature (Gene) Selection” step, there were three different approaches were taken: “scikit-learn” programs including feature selection and dimension deduction (written in Python), using GSEA (Gene Set Enrichment Analysis) to figure out the top genes; and using gene expression Meta-Analysis to obtain the top genes. All the selected genes with each method satisfied the statistical significance threshold of 0.05.

There are genes selected from the datasets in this study that were also previously identified in the pathways of KEGG Autism - KEGG DISEASE (KEGG DISEASE: Autism, n.d.) As shown in the **Supplementary Table S1**, *HLA-C, SLITRK2, NRXN1, CADM3 and CNTN1* are part of the “Cell adhesion molecules” pathway; And *GRIK2, GRIA2, GRIA4, PRKCB, PRKCG and GRM8* are part of the “Glutamatergic synapse” pathway. The selected *PRR5* is part of the “mTOR signaling” pathway. Meanwhile *PRKCB*, *PRKCG*, *FZD8* and *FZD6* are part of both “Wnt signaling” and “mTOR signaling” pathways. This overlap affirmed that the biomarkers are informative for detecting ASD.

Table S1. KEGG ASD suspected genes overlapping with selected genes in this study

Gene Symbol	Gene Name	KEGG ASD Pathway(s)	Dataset w/ Feature Selection
HLA-C	major histocompatibility complex, class I, C	Cell adhesion molecules (hsa04514)	GSE129806 - GSEA
SLITRK2	SLIT and NTRK like family member 2	Cell adhesion molecules (hsa04514)	Meta-Analysis w/ Combining Effect Sizes
NRXN1	neurexin 1	Cell adhesion molecules (hsa04514)	GSE214323 - scikit-learn
CADM3	cell adhesion molecule 3	Cell adhesion molecules (hsa04514)	Meta-Analysis w/ Combining P-Values
CNTN1	contactin 1	Cell adhesion molecules (hsa04514)	Meta-Analysis w/ Combining Effect Sizes

GRIK2	glutamate ionotropic receptor kainate type subunit 2	Glutamatergic synapse (hsa04724)	Meta-Analysis w/ Combining Effect Sizes
GRIA2	glutamate ionotropic receptor AMPA type subunit 2	Glutamatergic synapse (hsa04724)	GSE214323 - GSEA Meta-Analysis w/ Combining Effect Sizes
GRIA4	glutamate ionotropic receptor AMPA type subunit 4	Glutamatergic synapse (hsa04724)	Meta-Analysis w/ Combining Effect Sizes
PRKCB	protein kinase C beta	Glutamatergic synapse (hsa04724) Wnt signaling (hsa04310) mTOR signaling (hsa04150)	GSE214323 - GSEA
PRKCG	protein kinase C gamma	Glutamatergic synapse (hsa04724) Wnt signaling (hsa04310) mTOR signaling (hsa04150)	GSE214323 - GSEA
GRM8	glutamate metabotropic receptor 8	Glutamatergic synapse (hsa04724)	GSE214323 - GSEA Meta-Analysis w/ Combining Effect Sizes Meta-Analysis w/ Combining P-Values
FZD8	frizzled class receptor 8	Wnt signaling (hsa04310) mTOR signaling (hsa04150)	GSE214323 - scikit-learn Meta-Analysis w/ Combining P-Values
FZD6	frizzled class receptor 6	Wnt signaling (hsa04310) mTOR signaling (hsa04150)	GSE214323 - GSEA
PRR5	proline rich 5	mTOR signaling (hsa04150)	GSE214323 - scikit-learn Meta-Analysis w/ Combining P-Values

3.2 Single Dataset Workflow

3.2.1 Primary Component Analysis

As shown in **Supplementary Figure S2**, for the dataset of GSE129806, it would need 27 components to reach 100% of cumulative explained variances. And **Supplementary Figure S3** shows for two-dimension PCA, component 1 stands for 76.74% of explained variances and component 2 stands for 4.48% of explained variances.

For the dataset of GSE214323, it would need 49 components to reach 100% of cumulative explained variances as shown in **Supplementary Figure S4**. And **Supplementary Figure S5** shows two-dimension PCA component 1 stands for 69.75% of explained variances and component 2 stands for 6.01% of explained variances.

3.2.2 Hierarchical Clustering

Using dataset GSE129806 the Hierarchical Clustering with “scikit-learn” programs and GSEA (Gene Set Enrichment Analysis) both showed that 16 ASDs and 16 Controls were correctly clustered, suggesting that these selected features could be helpful for differentiating between ASDs and Controls. The detailed results are presented in the **Supplementary Figure S6** and **Supplementary Figure S7**.

Using the dataset GSE214323 the Hierarchical Clustering with “scikit-learn” programs and GSEA (Gene Set Enrichment Analysis) both showed that 36 ASDs and 18 Controls were correctly clustered, suggesting that these selected features could be helpful for differentiating between ASDs and Controls. The detailed results are presented in the **Supplementary Figure S8** and **Supplementary Figure S9**.

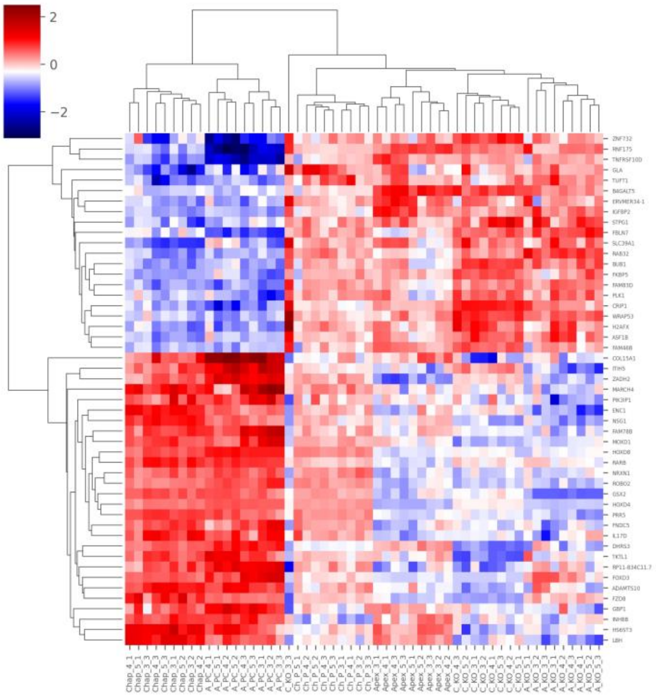

Figure S8. Scikit-learn hierarchical clustering

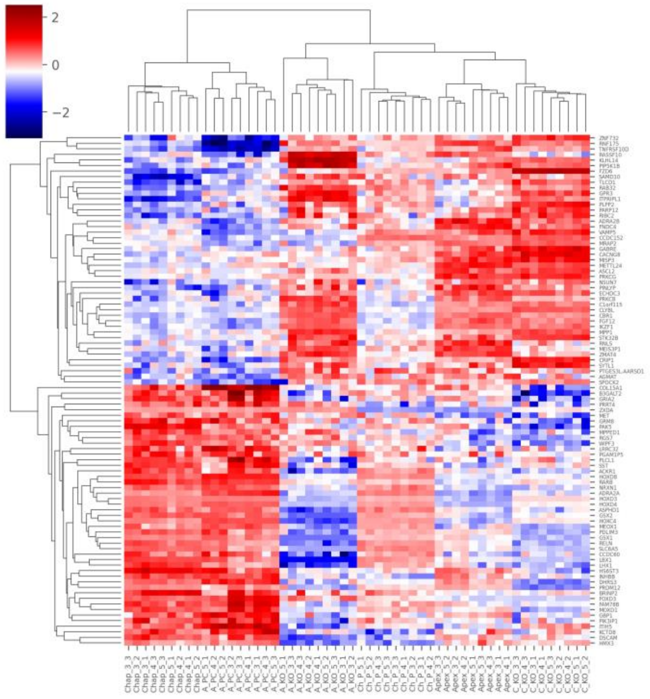

Figure S9. GSEA hierarchical clustering

3.2.3 Classification

As shown in **Supplementary Table S2** and **Table S3,** for the dataset of GSE129806, classifiers of "K-Nearest Neighbors" and "Stochastic Gradient Descent" showed 100% accuracy during the process using both scikit-learn and GSEA feature selection. And the classifier of "AdaBoost" showed 96.50% and 96.00% accuracy for scikit-learn and GSEA feature selection respectively. Meanwhile the classifier of "Quadratic Discriminant Analysis" showed lower 78.00% accuracy for scikit-learn feature selection and 73.50% accuracy for GSEA feature selection.

For the dataset of GSE214323, as shown in **Supplementary Table S4** and **Table S5,** the classifier of "K-Nearest Neighbors" showed 97.35% and 100.00% accuracy for scikit-learn and GSEA feature selection respectively. The classifier of "Stochastic Gradient Descent" showed 97.06% accuracy using scikit-learn and 99.12% accuracy using

GSEA feature selection. And the classifier of “AdaBoost” showed accuracy of 91.47% and 92.94% for scikit-learn and GSEA feature selection respectively. Meanwhile the classifier of “Quadratic Discriminant Analysis” showed lower 70.88% accuracy for scikit-learn feature selection and 75.29% accuracy for GSEA feature selection. The prediction models distinguished between the individuals with ASDs and Controls yielding promising results with relatively lower accuracy for the “Quadratic Discriminant Analysis” classifier.

3.3 Multi-Dataset with Meta-Analysis Workflow

3.3.1 Hierarchical Clustering

As shown in the figures **Supplementary Figure S12** and **Figure S13**, 36 ASDs and 18 Controls were entirely correctly clustered for GSE214323 for both meta-analysis methods of combining P-values and combining fixed effect sizes.

The Hierarchical Clustering using the dataset GSE214323 was relatively well discriminated from the dataset GSE129806. As shown in the figures **Supplementary Figure S10** and **Figure S11** the clustering was not entirely correct for GSE129806. For meta-analysis with combining P-values method, 8 out of 16 ASDs and 8 out of 16 Controls were correctly clustered, while for meta-analysis with combining fixed effect sizes method, 8 out of 16 ASDs and 9 out of 16 Controls were correctly clustered. Overall, Hierarchical Clustering results suggest these selected features could be helpful for differentiating between ASDs and Controls.

3.3.2 Classification

As shown in **Supplementary Table S6** and **Supplementary Table S7,** for the dataset of GSE129806 with meta-analysis using combining P-values method the classifier “Stochastic Gradient Descent” gave the highest accuracy of 98.50% while the classifier of “AdaBoost” showed the highest accuracy of 99.50% with meta-analysis using combining fixed effective size method. The classifier of “K-Nearest Neighbors” showed accuracy of 82.00% and 89.50% with meta-analysis using combining P-values and combining fixed effective size methods respectively. And the classifier of “Quadratic Discriminant Analysis”, similar to the performance in “Single Dataset” workflow, showed the lowest accuracy: 77.00% for meta-analysis using combining P-values and 70.00% for meta-analysis using combining fixed effective size.

For the dataset of GSE214323, as shown in **Supplementary Table S8** and **Supplementary Table S9,** the classifier of “K-Nearest Neighbors” gave 100% accuracy for both meta-analysis methods of combining P-values and combining fixed effect sizes. “Stochastic Gradient Descent” also showed high accuracy of 98.53% and 99.12% for the two meta-analysis methods respectively. The third best performing classifier of “AdaBoost” showed accuracy of 94.41% and 98.24% respectively for the two meta-analysis methods. And the classifier of “Quadratic Discriminant Analysis” gave relatively poorer performance for 72.94% accuracy using the meta-analysis method of combining P-values and 73.24% accuracy using the method of combining fixed effect sizes.

In this “Multi-Dataset with Meta-Analysis” workflow, dataset GSE129806, which shares the same features (genes) with GSE214323 from Feature Selection, was first trained using the different classifiers/ algorithms already mentioned. The trained models were saved into files (e.g., “GSE129806_MA-CombinePV_Stochastic Gradient Descent.joblib”) using the “Joblib” package. Then, at a later time, the saved model files were loaded back into the program and used to train the dataset GSE214323, for which the model had never seen before. GSE214323 was then randomly divided into 70% training data and 30% test data, and the loaded trained models used the test data to validate the prediction of ASDs vs Controls. This validation revealed that the prediction models, with the unseen test data, distinguished between the individuals with ASDs and Controls with relatively accurate results for classifiers such as “K-Nearest Neighbors” (85.88% accuracy) and “Stochastic Gradient Descent” (83.82% accuracy), which is shown in **Supplementary Table S10**.

Table S10. Classification w/ combining P-values trained models after 20 run(s) for GSE214323

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
K-Nearest Neighbors	85.88	100.00	78.52	87.68
Stochastic Gradient Descent	83.82	100.00	75.34	85.51
AdaBoost	61.18	75.90	59.85	65.77
Quadratic Discriminant Analysis	56.47	67.63	60.86	63.22

Notably, when the meta-analysis was performed using the “combining fixed effect sizes” method, the prediction results were better for every single classifier compared to using the “combining P-values” method, which is shown in **Supplementary Table S11**. For example, the accuracy of the “Stochastic Gradient Descent” classifier had increased from 83.82% to 90.59% when using the combined fixed effect sizes method. This behavior was expected for the method, which usually gives more conservative results (less DE/differential expressed features but more confident).

4. Discussion

Many people with ASD require life-long care and support. Even if some of them can live independently, they will face disadvantages for their education and employment opportunities. In addition, the demands on families providing care and support can be significant. Considering the prevalence of ASD in the United States and around the globe (Maenner et al., 2023), it is actually a pressing requirement for the academic research community, healthcare and technology industries to come up with solutions to help those individuals and families. To be able to detect the potential ASD development at the earlier age of people, like newborns or young children, would play a key role to apply appropriate treatments. Some people are not diagnosed until they are adolescents or adults. This delay means that people with ASD might not get the early help they need (Screening and Diagnosis | Autism Spectrum Disorder (ASD) | NCBDDD, 2022). Utilizing machine learning techniques to analyze available data could have the potential to speed and simplify diagnosis (Kassraian-Fard et al., 2016). According to CDC, there are many different factors that have been identified that may make a child more likely to have ASD, notably among the factors that may put children at greater risk for developing ASD (Basics About Autism Spectrum Disorder (ASD) | NCBDDD | CDC, 2022), several are genetic related such as:

- Having a sibling with ASD
- Having certain genetic or chromosomal conditions, such as fragile X syndrome or tuberous sclerosis

Using genetic data would be more effective for the purpose of detecting ASD as early as possible, compared to some other efforts using data like “set of behaviors” (Kosmicki et al., 2015), when the data is made available at much later time for people developing ASD. In this study, different approaches for Feature Selection were used: Python programs using “scikit-learn”, using GSEA manually, and running gene expression Meta-Analysis manually. A result of importance is that the obtained genes of significance list overlapped with previously reported candidate genes and pathway associations for ASD from the KEGG Autism - KEGG DISEASE (KEGG DISEASE: Autism, n.d.) database, confirming that the machine learning prediction models could be effective at determining biomarkers.

The machine learning classification results in the study were consistent with the findings of previous studies that reported on gene expression signatures with a high diagnostic accuracy for ASD (Pramparo et al., 2015; Hu & Lai, 2013). The classifiers used in this study, which are “K-Nearest Neighbors”, “Stochastic Gradient Descent”, “AdaBoost” and “Quadratic Discriminant Analysis”, had different performances in different workflows. Overall “Quadratic Discriminant Analysis” yielded the lowest prediction accuracy scores. While other three classifiers resulted in promising results of accuracy and other metrics. Utilizing multiple classifiers enabled the cross reference for the same sample(s) to yield higher confidence for the prediction results. One of the study’s goals was to develop a more generic, widespread application of machine learning and use them to perform prediction on unseen samples in the biogenetic field. Instead of only focusing on a single dataset for machine learning prediction like other studies (Oh et al., 2017; Lin et al., 2021), this exploratory study used meta-analysis methodologies and tools to create “common” feature/gene sets for multiple datasets. With the “common” feature/gene sets, data from one dataset was trained by machine learning algorithms then the trained models were used to perform prediction on unseen dataset. This approach revealed the encouraging results that gene expression meta-analysis with machine learning may offer the stepping-stones to achieve the original goal eventually.

There are some limitations that come with this study. First, given that our study applied analysis on archival pre-existing datasets, which are pluripotent stem cell RNA-seq data. Further validation of the effectiveness for the developed workflows might be needed for different cell types, such as brain cells and blood cells. Second, the datasets have limited sample sizes, which are common in gene expression level data. From a machine learning point of view, many more samples would definitely be helpful for building more general and accurate prediction models. Third, the meta-analysis tools used in this study do not support some of the formulas for combining P-values, including Pearson’s method, Tippet’s method (minimum of P values) and Wilkinson’s method (maximum of P values) (Toro-Domínguez et al., 2020). So, the more comprehensive prediction performance comparison could not be done with the different selected genes/features. Due to the limitations mentioned above the results of this study should be cautiously interpreted.

5. Conclusions

In conclusion, this study suggests that machine learning techniques can be used to analyze RNA-seq genetic data and use it to distinguish between ASD and control samples with promising accuracy. If further analysis is performed on more datasets and validated in a larger cohort of cases and controls, the more general the workflows and machine learning models would be, which would increase the accuracy and expedite future diagnoses of ASD. Subsequently, individualized treatment options for patients could be made earlier which would have a significant positive impact on both the patients and their families.

Acknowledgments

Special thanks to the following two gentlemen who gave invaluable advice to this study and helped with the peer-review of the research paper:

Shuangge Steven Ma, PhD, Department Interim Chair and Professor of Biostatistics; Affiliated Faculty, Yale Institute for Global Health; Director, Biostatistics and Bioinformatics Shared Resource, Yale University.

Frederick Zhao, MS in Neural Science, Northwestern University.

Reference

Ansel, A., Rosenzweig, J. P., Zisman, P. D., Melamed, M. L., & Gesundheit, B., (2017). Variation in gene expression in autism Spectrum Disorders: An Extensive Review of Transcriptomic studies. *Frontiers in Neuroscience*, *10*. https://doi.org/10.3389/fnins.2016.00601.

Autism,(2023, March 29). https://www.who.int/news-room/questions-and-answers/item/autism-spectrum-disorders-(asd).

Bedre, R., (2020, May 24). Reneshbedre/bioinfokit: Bioinformatics Data Analysis and visualization toolkit. Zenodo. https://zenodo.org/record/3841708#.XyCfi-dOmUk.

Bedre, R., (2022, September 4). Reneshbedre/bioinfokit: Bioinformatics Data Analysis and visualization toolkit. Zenodo. http://doi.org/10.5281/zenodo.3698145.

Bone, D., Goodwin, M. S., Black, M., Lee, C., Audhkhasi, K., & Narayanan, S. S., (2014). Applying Machine Learning to Facilitate Autism Diagnostics: Pitfalls and promises. *Journal of Autism and Developmental Disorders*, *45*(5), 1121-1136. https://doi.org/10.1007/s10803-014-2268-6.

Borenstein, M., Hedges, L.V., Higgins, J.P.T. and Rothstein, H.R., (2009). Front Matter. In *Introduction to Meta-Analysis* (eds M. Borenstein, L.V. Hedges, J.P.T. Higgins and H.R. Rothstein). https://doi.org/10.1002/9780470743386.fmatter.

Centers for Disease Control and Prevention, (2022, December 9). What is autism spectrum disorder?. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/facts.html.

Centers for Disease Control and Prevention, (2022, March 31). Screening and diagnosis of autism spectrum disorder. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/screening.html.

Centers for Disease Control and Prevention, (2023, April 4). Data & statistics on autism spectrum disorder. Centers for Disease Control and Prevention. https://www.cdc.gov/ncbddd/autism/data.html.

Colvert, E., Tick, B., McEwen, F., Stewart, C., Curran, S. R., Woodhouse, E., Gillan, N., Hallett, V., Lietz, S., Garnett, T., Ronald, A., Plomin, R., Rijsdijk, F., Happé, F., & Bolton, P., (2015). Heritability of Autism Spectrum Disorder in a UK Population-Based Twin Sample. *JAMA psychiatry*, *72*(5), 415-423. https://doi.org/10.1001/jamapsychiatry.2014.3028.

GSEA | MSIGDB | Browse Human Gene Sets, (n.d.). https://www.gsea-msigdb.org/gsea/msigdb/human/genesets.jsp?collection=CP:KEGG.

GSEA Tool, (n.d.). https://www.gsea-msigdb.org/gsea/index.jsp.

GSEA, (n.d.). https://www.gsea-msigdb.org/gsea/index.jsp

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J. M., Cohen, B., Torigoe, T., Miller, J. E., Fedele, A., Collins, J., Smith, K., Lotspeich, L., Croen, L., Ozonoff, S. J., Lajonchere, C., Grether, J. K., & Risch, N., (2011b). Genetic heritability and shared environmental factors among twin pairs with autism. *Archives of General Psychiatry*, *68*(11), 1095. https://doi.org/10.1001/archgenpsychiatry.2011.76.

Hameed, S. S., Hassan, R., & Muhammad, F. F., (2017, November 2). Selection and classification of gene expression in autism disorder: Use of a combination of statistical filters and a GBPSO-SVM algorithm. *PloS one.* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5667738/

Hu, V. W., & Lai, Y., (2013). Developing a Predictive Gene Classifier for Autism Spectrum Disorders Based upon Differential Gene Expression Profiles of Phenotypic Subgroups. *North American journal of medicine & science, 6*(3). doi: 10.7156/najms.2013.0603107. https://doi.org/10.7156/najms.2013.0603107.

Jensen, A.R., Lane, A.L., Werner, B.A. et al., (2022). Modern Biomarkers for Autism Spectrum Disorder: Future Directions. *Mol Diagn Ther* 26, 483-495. https://doi.org/10.1007/s40291-022-00600-7.

Jeste SS, Geschwind DH., (2014 February). Disentangling the heterogeneity of autism spectrum disorder through genetic findings. *Nat Rev Neurol., 10*(2), 74-81. doi: 10.1038/nrneurol.2013.278. Epub 2014 Jan 28. PMID: 24468882; PMCID: PMC4125617.

Joblib: running Python functions as pipeline jobs — joblib 1.3.2 documentation, (n.d.). https://joblib.readthedocs.io/en/stable/

Kassraian-Fard P, Matthis C, Balsters JH, Maathuis MH, Wenderoth N., (2016, December 1). Promises, Pitfalls, and Basic Guidelines for Applying Machine Learning Classifiers to Psychiatric Imaging Data, with Autism as an Example. *Front Psychiatry, 7,* 177. doi: 10.3389/fpsyt.2016.00177. PMID: 27990125; PMCID: PMC5133050.

KEGG DISEASE: Autism, (n.d.). https://www.genome.jp/entry/H02111.

Kosmicki, J. A., Sochat, V., Duda, M., & Wall, D. P., (2015). Searching for a minimal set of behaviors for autism detection through feature selection-based machine learning. *Translational Psychiatry*, *5*(2), e514. https://doi.org/10.1038/tp.2015.7.

Liberzon A, et al., (2011). Mesirov, Molecular signatures database (MSigDB) 3.0, *Bioinformatics, 27*(12), pp. 1739-1740, https://doi.org/10.1093/bioinformatics/btr260.

Liberzon A, et al., (2015). The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell systems, 1*(6), 417-425. https://doi.org/10.1016/j.cels.2015.12.004.

Lin Y, Afshar S, Rajadhyaksha AM, Potash JB and Han S., (2020). A Machine Learning Approach to Predicting Autism Risk Genes: Validation of Known Genes and Discovery of New Candidates. *Front Genet, 11*, 500064. doi: 10.3389/fgene.2020.500064.

Lin, P.-I., Moni, M. A., Gau, S. S.-F., & Eapen, V., (2021, May 12). Identifying subgroups of patients with autism by gene expression profiles using machine learning algorithms. *Frontiers in psychiatry*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8149626/.

Maenner, M. J., Warren, Z., Williams, A. R., Amoakohene, E., Bakian, A. V., Bilder, D. A., Durkin, M. S., Fitzgerald, R. T., Furnier, S. M., Hughes, M. M., Ladd-Acosta, C., McArthur, D., Pas, E. T., Salinas, A., Vehorn, A., Williams, S. P., Esler, A., Grzybowski, A., Hall-Lande, J., Shaw, K. A., (2023e). Prevalence and characteristics of autism spectrum disorder among children aged 8 years — Autism and Developmental Disabilities Monitoring Network, 11 sites, United States, 2020. *Morbidity and Mortality Weekly Report, 72*(2), 1-14. https://doi.org/10.15585/mmwr.ss7202a1.

Mootha, V. K., Lindgren, C. M., et al., (2003). PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet 34*, 267–273 (2003). https://doi.org/10.1038/ng1180

Networkanalyst, (n.d.). https://www.networkanalyst.ca/NetworkAnalyst/home.xhtml.

NumPy, (n.d.). https://numpy.org/

Oh, D. H., Kim, I. B., Kim, S. H., & Ahn, D. H., (2017, February 28). Predicting autism spectrum disorder using blood-based gene expression signatures and machine learning. *Clinical psychopharmacology and neuroscience: the official scientific journal of the Korean College of Neuropsychopharmacology*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5290715/

Pandas, (n.d.). https://pandas.pydata.org/

Parab L, Pal S, Dhar R., (2022, December 12). Transcription factor binding process is the primary driver of noise in gene expression. *PLoS Genet.,18*(12), e1010535. doi: 10.1371/journal.pgen.1010535. PMID: 36508455; PMCID: PMC9779669.

Pramparo T, Pierce K, Lombardo MV, Carter Barnes C, Marinero S, Ahrens-Barbeau C, Murray SS, Lopez L, Xu R, Courchesne E., (2015, April). Prediction of autism by translation and immune/inflammation coexpressed genes in toddlers from pediatric community practices. *JAMA Psychiatry, 72*(4), 386-94. doi: 10.1001/jamapsychiatry.2014.3008. PMID: 25739104.

Rahman, M. M., Usman, O. L., Muniyandi, R. C., Sahran, S., Mohamed, S., & Razak, R. A., (2020, December 7). A review of machine learning methods of feature selection and classification for autism spectrum disorder. *Brain sciences*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7762227/

Renesh Bedre, (2020, March 5). Reneshbedre/bioinfokit: Bioinformatics data analysis and visualization toolkit. Zenodo. http://doi.org/10.5281/zenodo.3698145.

Reneshbedre, (n.d.). Reneshbedre/bioinfokit: Bioinformatics Data Analysis and visualization toolkit. GitHub. https://github.com/reneshbedre/bioinfokit.

SKLEARN. DECOMPOSITION. INCREMENTALPCA. Scikit, (n.d.). https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.IncrementalPCA.html.

Statistical Data Visualization#. Seaborn, (n.d.). https://seaborn.pydata.org/

Subramanian, A., Tamayo, P., et al., (2005, PNAS). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545–15550. https://doi.org/10.1073/pnas.0506580102.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A. G., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P., (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545-15550. https://doi.org/10.1073/pnas.0506580102.

Thudumu, S., Branch, P., Jin, J., & Singh, J. (Jack)., (2020, July 2). A comprehensive survey of anomaly detection techniques for high dimensional big data — journal of big data. *SpringerOpen*. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00320-x#:~:text=High%20dimensionality%20refers%20to%20data,methods%20to%20process%20the%20data

Toro-Domínguez, D., Villatoro-García, J. A., Martorell-Marugán, J., Román-Montoya, Y., Alarcón-Riquelme, M. E., & Carmona-Sáez, P., (2020). A survey of gene expression meta-analysis: methods and applications. *Briefings in Bioinformatics*, *22*(2), 1694-1705. https://doi.org/10.1093/bib/bbaa019.

U.S. National Library of Medicine, (n.d.-c). Home — Geo Datasets — NCBI. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gds.

World Health Organization: WHO, (2023). Autism. www.who.int. https://www.who.int/news-room/fact-sheets/detail/autism-spectrum-disorders.

Yoon, S., Baik, B., Park, T. et al., (2021). Powerful *p*-value combination methods to detect incomplete association. *Sci Rep*, 11, 6980. https://doi.org/10.1038/s41598-021-86465-y.

Supplementary Materials

Figures

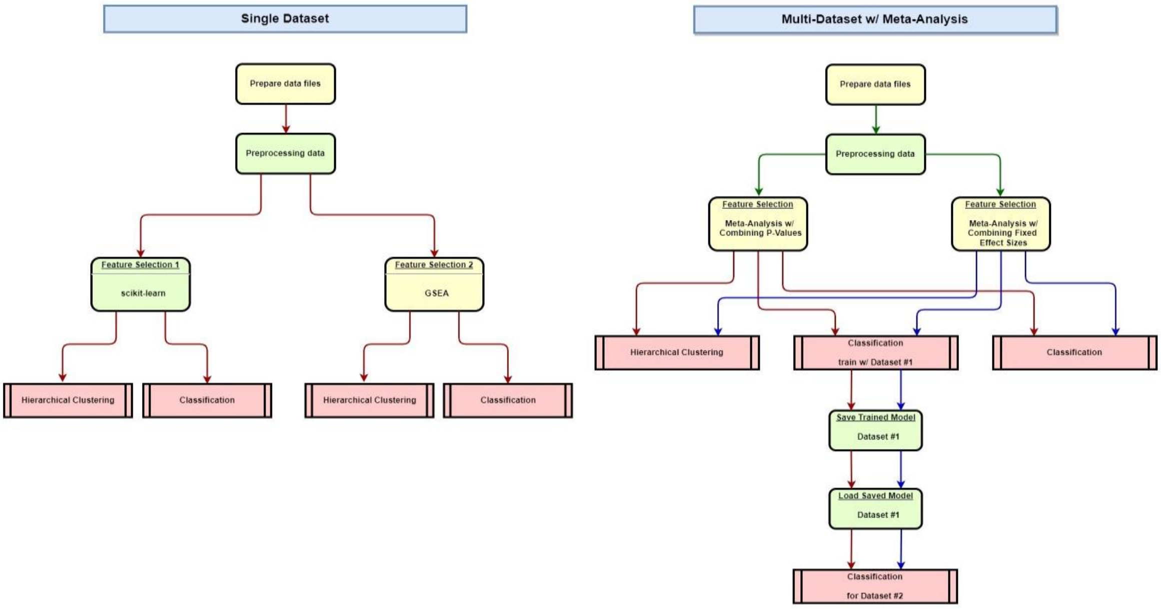

Figure S1. Workflows

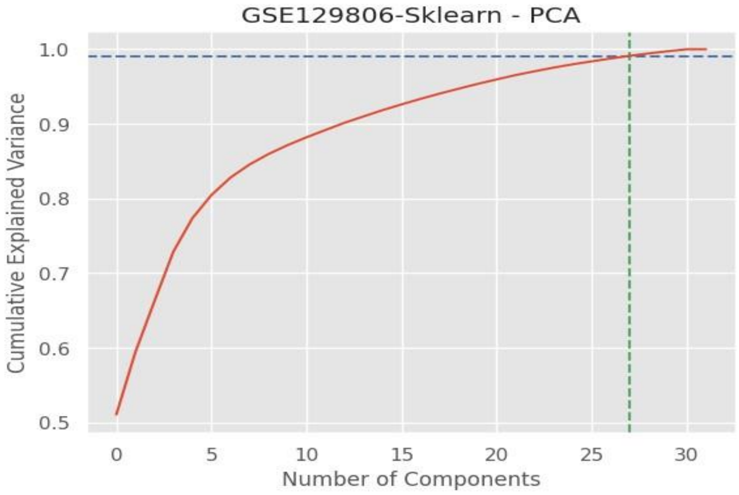

Figure S2. Number of Components needed vs Cumulative Explained Variance for dataset of GSE129806

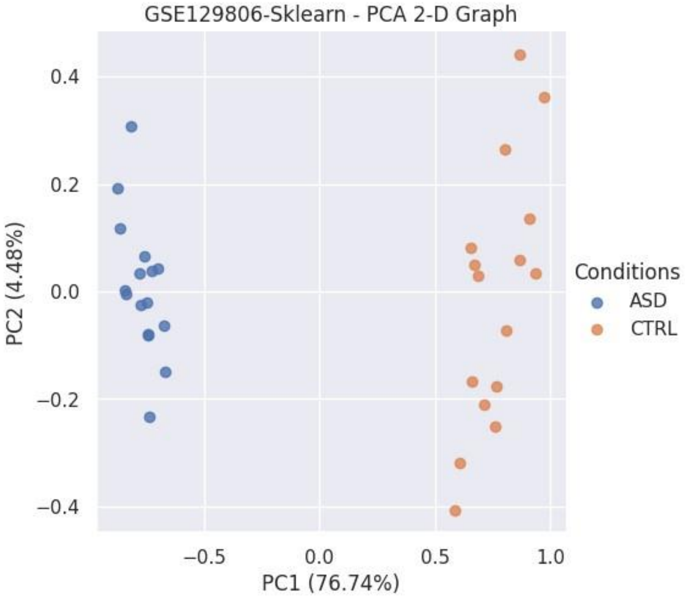

Figure S3. Two-component PCA plot showing ASD vs CTRL samples for dataset of GSE129806

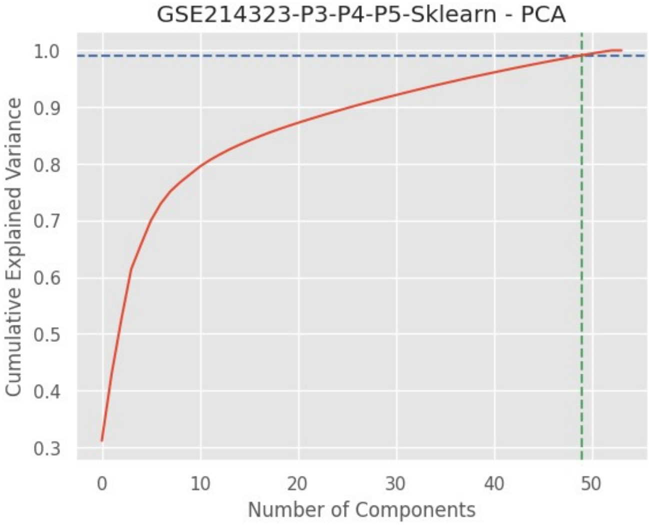

Figure S4. Number of Components needed vs Cumulative Explained Variance for dataset of GSE214323

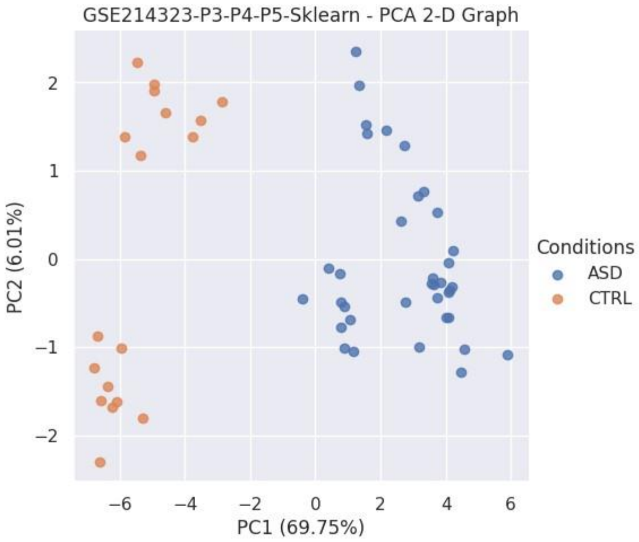

Figure S5. Two-component PCA plot showing ASD vs CTRL samples for dataset of GSE214323

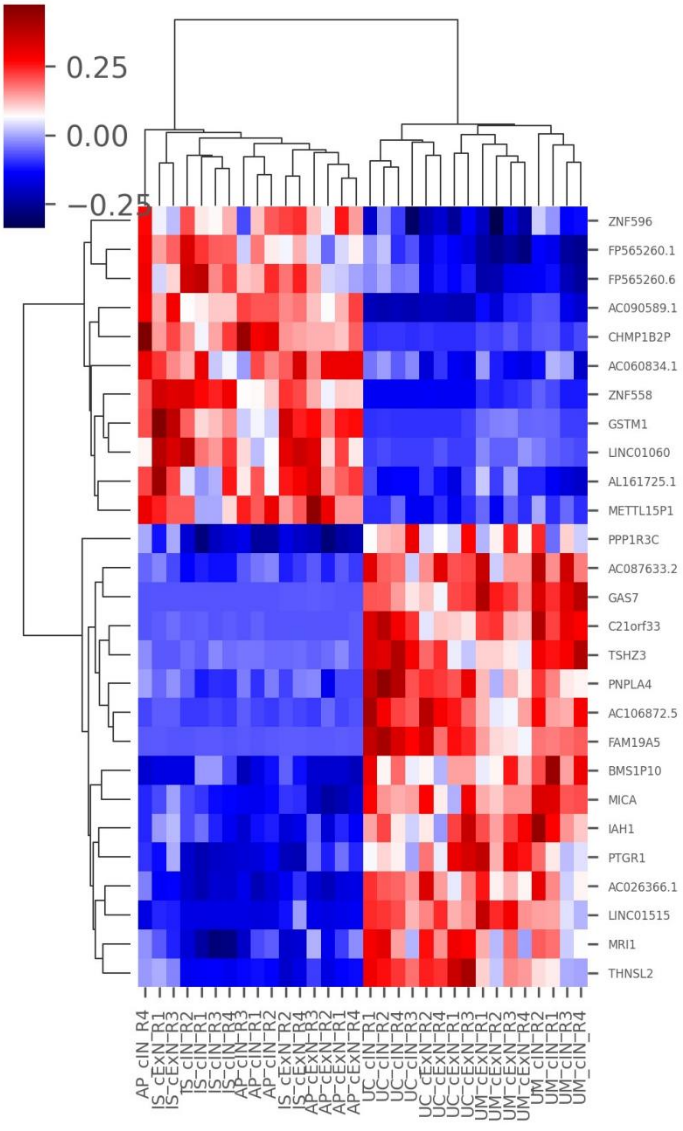

Figure S6. Scikit-learn hierarchical clustering heatmap for dataset of GSE129806

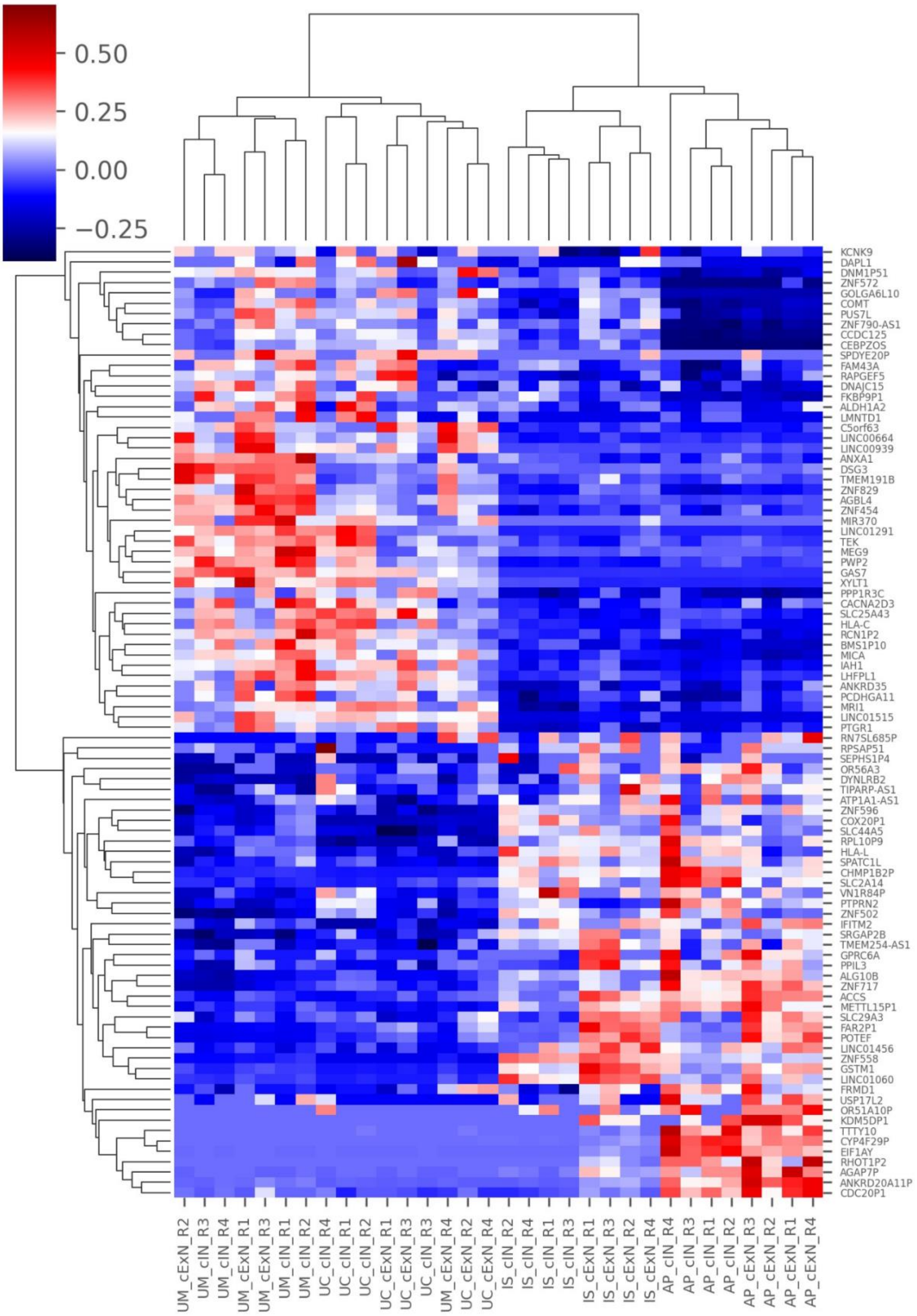

Figure S7. GSEA hierarchical clustering heatmap for dataset of GSE129806

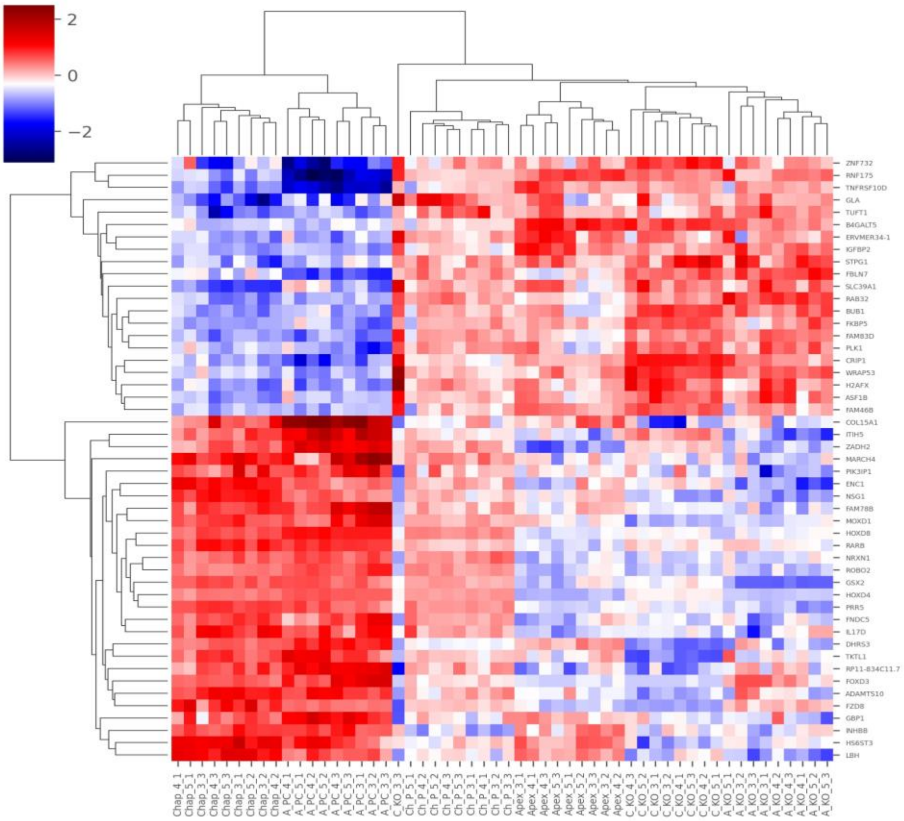

Figure S8. scikit-learn hierarchical clustering heatmap for dataset of GSE214323

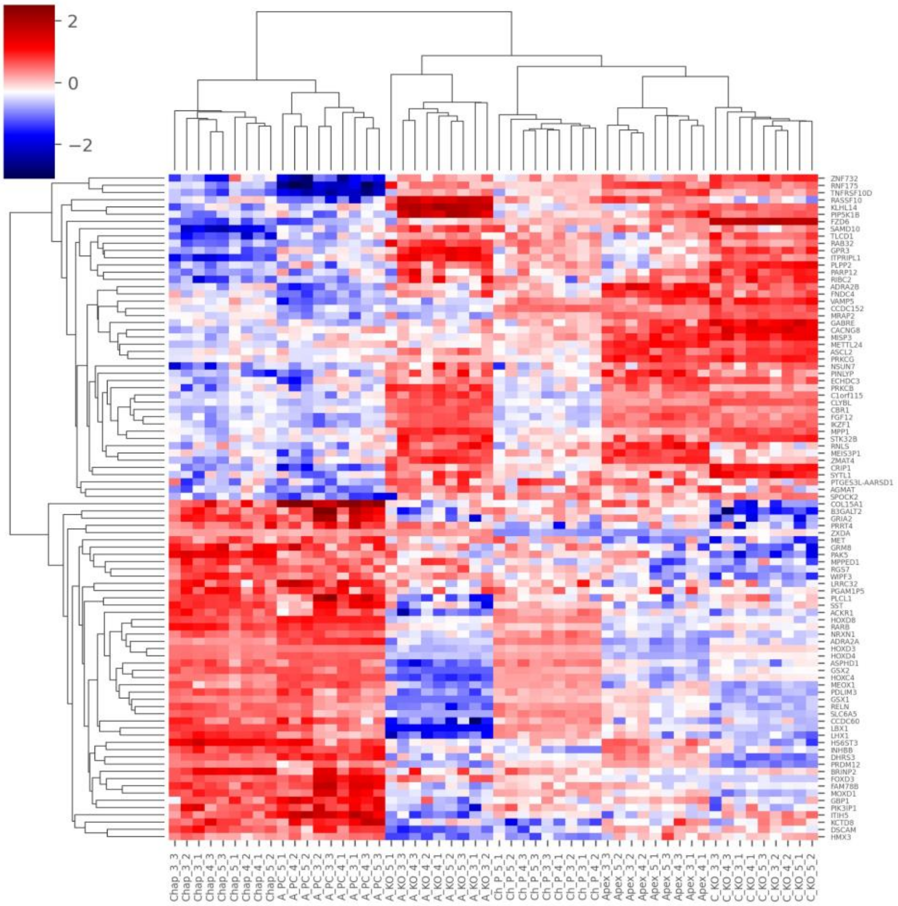

Figure S9. GSEA hierarchical clustering heatmap for dataset of GSE214323

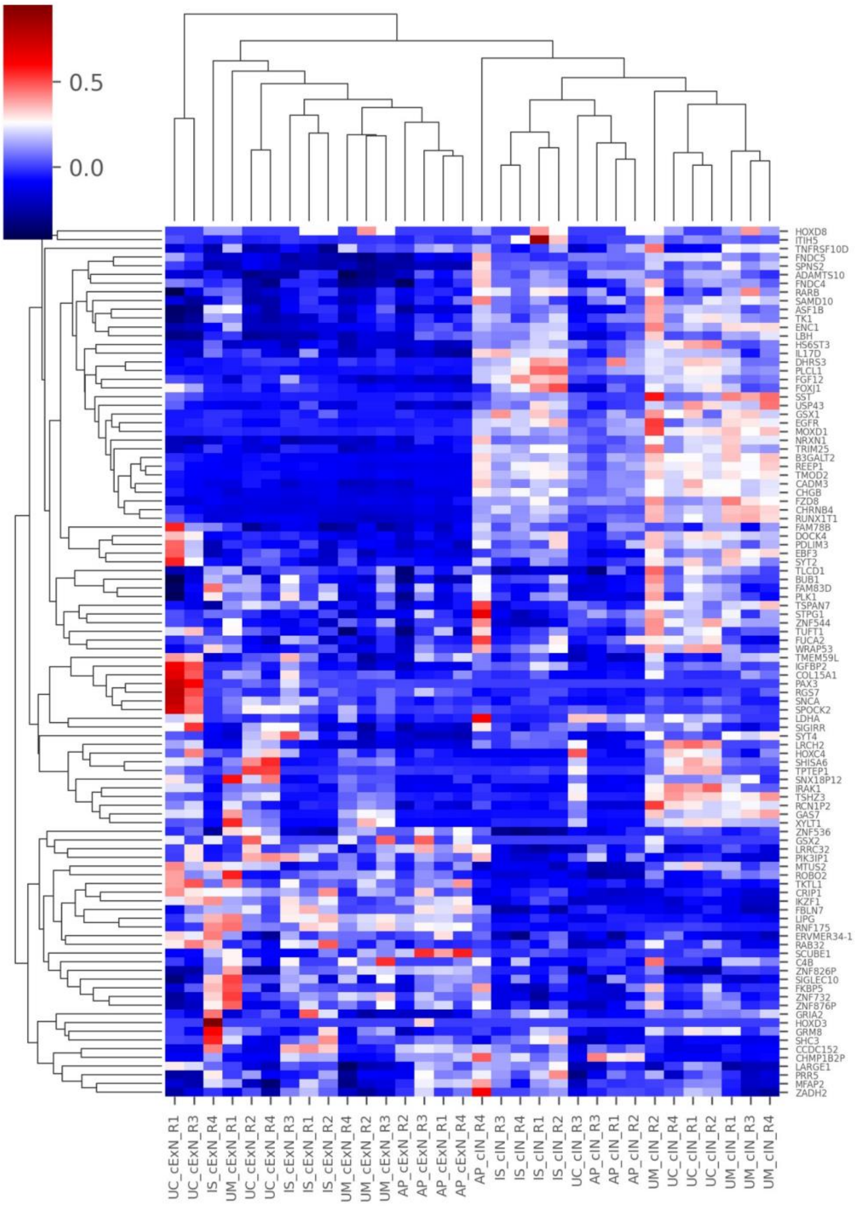

Figure S10. Meta-Analysis w/ combined P-values hierarchical clustering heatmap for dataset of GSE129806

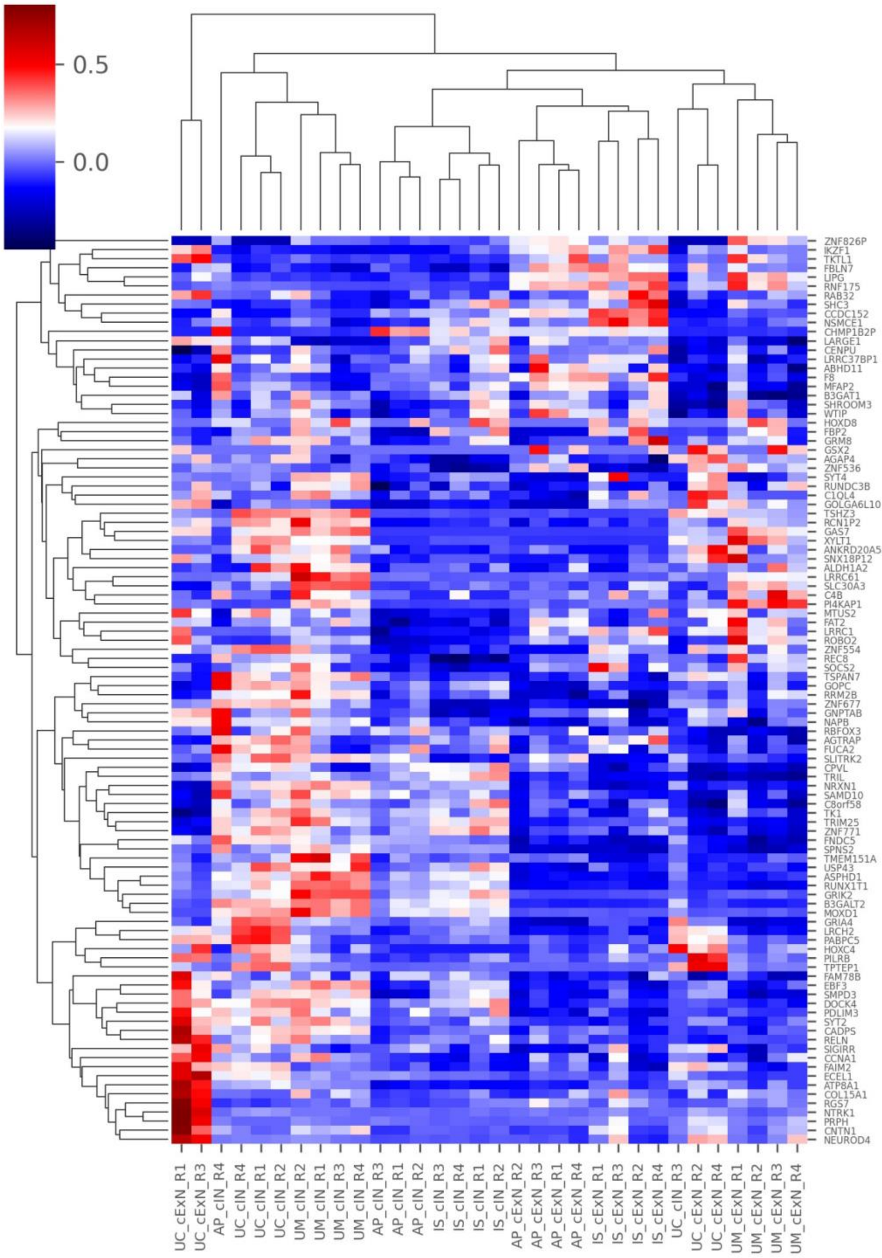

Figure S11. Meta-Analysis w/ combined effect sizes hierarchical clustering heatmap for dataset of GSE129806

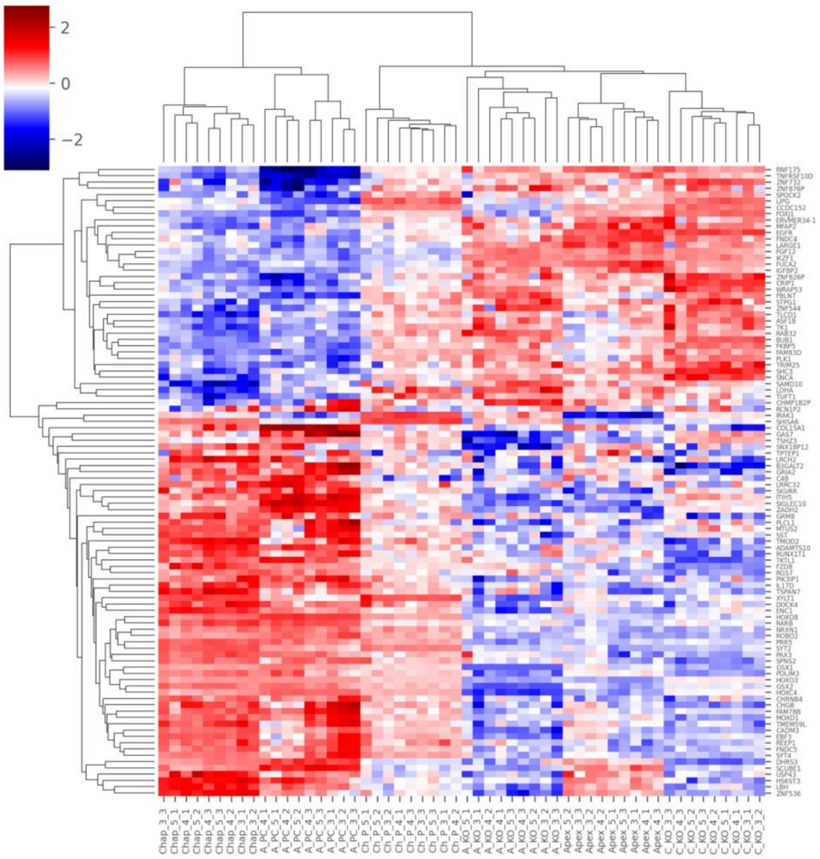

Figure S12. Meta-Analysis w/ combined P-values hierarchical clustering heatmap for dataset of GSE214323

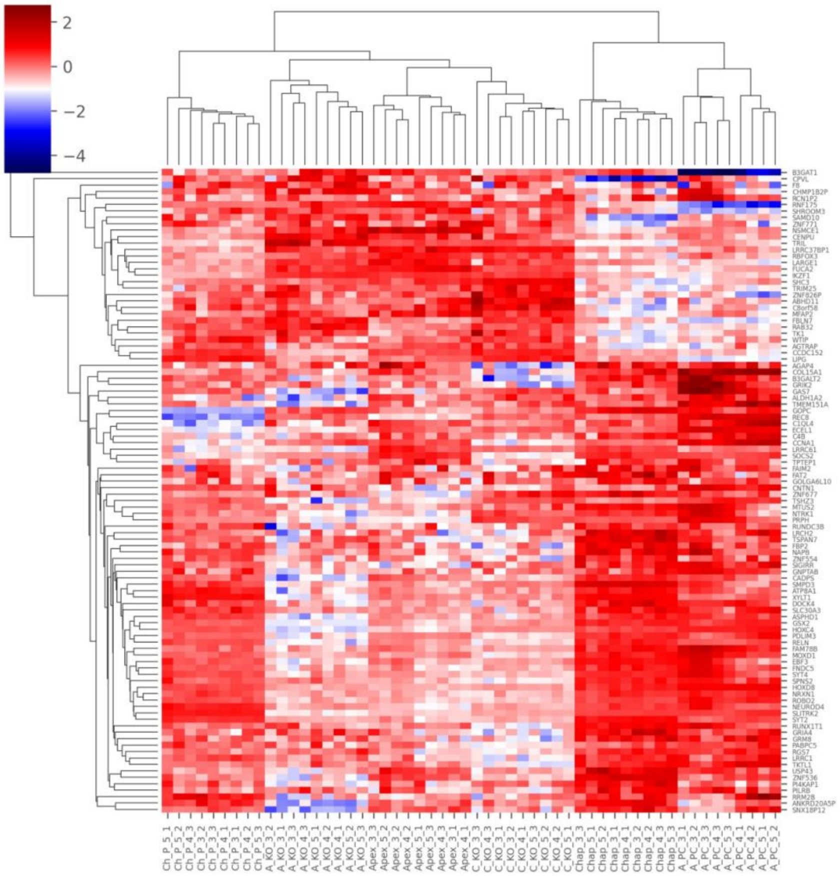

Figure S13. Meta-Analysis w/ combined effect sizes hierarchical clustering heatmap for dataset of GSE214323

Tables

Table S1. KEGG ASD suspected genes overlapping with selected genes in this study

Gene	**Gene**	**KEGG ASD**	**Dataset**
Symbol	**Name**	**Pathway(s)**	**w/ Feature Selection**
HLA-C	major histocompatibility complex, class I, C	Cell adhesion molecules (hsa04514)	GSE129806 - GSEA
SLITRK2	SLIT and NTRK like family member 2	Cell adhesion molecules (hsa04514)	Meta-Analysis w/ Combining Effect Sizes
NRXN1	neurexin 1	Cell adhesion molecules (hsa04514)	GSE214323 - scikit-learn
CADM3	cell adhesion molecule 3	Cell adhesion molecules (hsa04514)	Meta-Analysis w/ Combining P-Values
CNTN1	contactin 1	Cell adhesion molecules (hsa04514)	Meta-Analysis w/ Combining Effect Sizes
GRIK2	glutamate ionotropic receptor kainate type subunit 2	Glutamatergic synapse (hsa04724)	Meta-Analysis w/ Combining Effect Sizes
GRIA2	glutamate ionotropic receptor AMPA type subunit 2	Glutamatergic synapse (hsa04724)	GSE214323 - GSEA Meta-Analysis w/ Combining Effect Sizes
GRIA4	glutamate ionotropic receptor AMPA type subunit 4	Glutamatergic synapse (hsa04724)	Meta-Analysis w/ Combining Effect Sizes
PRKCB	protein kinase C beta	Glutamatergic synapse (hsa04724) Wnt signaling (hsa04310) mTOR signaling (hsa04150)	GSE214323 - GSEA
PRKCG	protein kinase C gamma	Glutamatergic synapse (hsa04724) Wnt signaling (hsa04310) mTOR signaling (hsa04150)	GSE214323 - GSEA
GRM8	glutamate metabotropic receptor 8	Glutamatergic synapse (hsa04724)	GSE214323 - GSEA Meta-Analysis w/ Combining Effect Sizes Meta-Analysis w/ Combining P-Values
FZD8	frizzled class receptor 8	Wnt signaling (hsa04310) mTOR signaling (hsa04150)	GSE214323 - scikit-learn Meta-Analysis w/ Combining P-Values
FZD6	frizzled class receptor 6	Wnt signaling (hsa04310) mTOR signaling (hsa04150)	GSE214323 - GSEA
PRR5	proline rich 5	mTOR signaling (hsa04150)	GSE214323 - scikit-learn Meta-Analysis w/ Combining P-Values

Table S2. Classification using **scikit-learn** feature selection after 20 run(s) for GSE129806

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
K-Nearest Neighbors	100.00	100.00	100.00	100.00
Stochastic Gradient Descent	100.00	100.00	100.00	100.00
AdaBoost	96.50	97.92	95.33	96.49
Quadratic Discriminant Analysis	78.00	81.99	73.31	73.47

Table S3. Classification using **GSEA** feature selection after 20 run(s) for GSE129806

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
K-Nearest Neighbors	100.00	100.00	100.00	100.00
Stochastic Gradient Descent	100.00	100.00	100.00	100.00
AdaBoost	96.00	95.74	97.17	96.21
Quadratic Discriminant	73.50	72.53	74.82	71.58

Table S4. Classification using **scikit-learn** feature selection after 20 run(s) for GSE214323

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
K-Nearest Neighbors	97.35	99.09	97.50	97.83
Stochastic Gradient Descent	97.06	99.44	96.29	97.75
AdaBoost	91.47	96.06	91.62	93.54
Quadratic Discriminant Analysis	70.88	91.94	66.21	73.17

Table S5. Classification using **GSEA** feature selection after 20 run(s) for GSE214323

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
K-Nearest Neighbors	100.00	100.00	100.00	100.00
Stochastic Gradient Descent	99.12	100.00	98.67	99.30
AdaBoost	92.94	94.05	94.90	94.20
Quadratic Discriminant Analysis	75.29	86.58	78.54	80.42

Table S6. Classification using Meta-Analysis **combining P-values** after 20 run(s) for GSE129806

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Stochastic Gradient Descent	98.50	98.33	97.50	97.57
AdaBoost	96.50	96.29	97.62	96.65
K-Nearest Neighbors	82.00	80.29	89.53	83.16
Quadratic Discriminant Analysis	77.00	79.24	65.05	68.21

Table S7. Classification using Meta-Analysis **combining effect sizes** after 20 run(s) for GSE129806

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
Stochastic Gradient Descent	96.00	96.62	96.57	96.12
AdaBoost	99.50	99.17	100.00	99.55
K-Nearest Neighbors	89.50	86.35	96.74	89.93
Quadratic Discriminant Analysis	70.00	72.11	64.19	62.70

Table S8. Classification using Meta-Analysis **combining P-values** after 20 run(s) for GSE214323

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
K-Nearest Neighbors	100.00	100.00	100.00	100.00
Stochastic Gradient Descent	98.53	100.00	97.70	98.77
AdaBoost	94.41	96.22	94.84	95.40
Quadratic Discriminant Analysis	72.94	86.42	74.20	76.42

Table S9. Classification using Meta-Analysis **combining effect sizes** after 20 run(s) for GSE214323

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
K-Nearest Neighbors	100.00	100.00	100.00	100.00
Stochastic Gradient Descent	99.12	100.00	98.83	99.38
AdaBoost	98.24	100.00	97.34	98.55
Quadratic Discriminant Analysis	73.24	88.06	72.42	76.51

Table S10. Classification w/ **combining P-values trained models** after 20 run(s) for GSE214323

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
K-Nearest Neighbors	85.88	100.00	78.52	87.68
Stochastic Gradient Descent	83.82	100.00	75.34	85.51
AdaBoost	61.18	75.90	59.85	65.77
Quadratic Discriminant Analysis	56.47	67.63	60.86	63.22

Table S11. Classification w/ **combining effect sizes trained models** after 20 run(s) for GSE214323

Classifier	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1-Score (%)**
K-Nearest Neighbors	96.18	100.00	94.14	96.91
Stochastic Gradient Descent	90.59	100.00	86.52	92.67
AdaBoost	65.00	79.94	61.23	69.01
Quadratic Discriminant Analysis	62.35	70.31	80.79	74.76

Table S12. Meta-Analysis methods of Combining P-Values vs Combining Fixed Effect Sizes

Method	**Formula**
Combining P-Values	"Fisher's method", uses the sum of the logarithms of the P values, that is to say: $-2 \times \sum_{i=1}^{k} ■ = ln\ (pi)$" where pi is each of the P values of the different studies. In this case, the null hypothesis is when there is no difference in gene expression between the different studies and it is distributed as a χ2 with 2k degrees of freedom (k being the number of studies).
Combining Fixed Effect Sizes	It is a linear model that considers that the different studies share a common effect size called true effect. The *combined effect*, $\underline{T.}$, is calculated as: $\underline{T.} = \frac{\sum \quad \omega_i T_i}{\sum ■ ■ \omega_i}$ where ω_i are the different weights assigned to each study, that is, the inverse within-study variance, $V\ (T_i)$: $\omega_i = \frac{1}{V\ (T_i)}$ The variance of the combined effect is defined as: $V\left(\underline{T.}\right) = \frac{1}{\sum \quad \omega_i}$ The *combined effect* value for a standard normal: $Z = \frac{\underline{T.}}{\sqrt{V\ (\underline{T.})}}$ Therefore, two-tailed *P* value can be calculated by: $P = 2\ [1 - (\Phi(\ \lvert Z \rvert\ ))]$ where Φ is the standard normal cumulative distribution function.

Associated Data

GSE129806 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129806)

GSE214323 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE214323)

Copyrights