# Genetic Biomarkers and Biological Pathways for Autism Spectrum Disorders

Kashvi Srinivasan

Oakridge International School Bangalore

srinivasankashvi@gmail.com

# Contents

# Autism Spectrum Disorders (ASD)

- 1 in every 100 people are affected by ASD (Zeidan et al., 2022).
- Symptoms Include (Matson et al., 2008; Nadeem et al., 2021):
  - Challenges with social interactions
  - Restricted interests
  - Repetitive behavior
  - Gastrointestinal dysfunction
  - Weak immunity
- Early diagnosis can improve later outcomes through reductions in symptom severity (Charman and Baird, 2002).

# Background

- **Current Diagnostic Methods**:

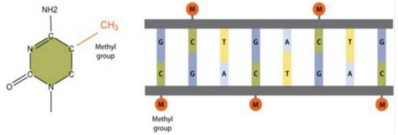
  Expert Clinical Observation  (Charman and Baird, 2002)

- **Challenges in Diagnosis** (McCarty and Frye, 2020):
  - Symptoms are considered normal at young age.
  - Limited peer interaction may obscure early symptoms.
  - Overlap with other neuropsychiatric disorders.
  - There is a 33% accuracy when reporting a positive case.

# Background

- Compared to similar disorders, ASD has the strongest genetic component (Klin, 2018).
- Several genes have been linked to ASD (PTEN, CHD8, SCN2A, and ADNP) (Vogt et al., 2015; Yu et al., 2022; Sanders et al., 2018; Arnett et al., 2018). However, these have not been applied to diagnosis or treatment.

# Background

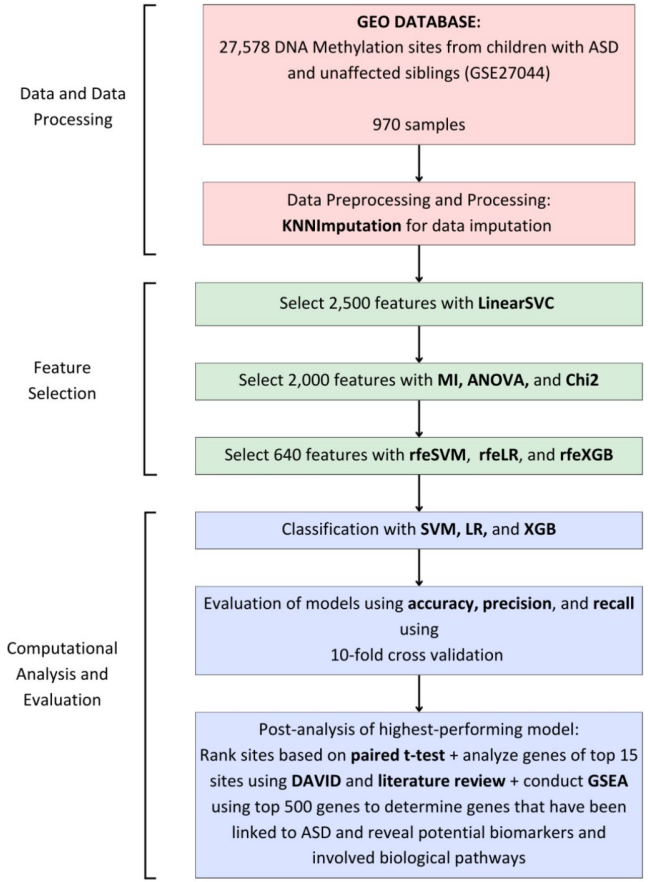



- DNA methylation involves adding a methyl group to a cytosine nucleotide at the C5 position to form 5-methylcytosine.
- Regulate gene expression without affecting nucleotide sequences.
- Prevent the binding of transcription factors, reducing the expression of a gene.
- Recruit proteins involved in gene expression (Moore et al., 2013).

# Research Aim

This research aims to use DNA methylation levels at different sites in children with and without ASD to reveal genetic biomarkers and biological pathways involved in ASD. These can be used for the diagnosis, prediction, and treatment of ASD, leading to improved outcomes for affected individuals.

# Method


Data and Data Processing
GEO DATABASE:
27,578 DNA Methylation sites from children with ASD and unaffected siblings (GSE27044)
970 samples
Data Preprocessing and Processing:
KNNImputation for data imputation
Feature Selection
Select 2,500 features with LinearSVC
Select 2,000 features with MI, ANOVA, and Chi2
Select 640 features with rfeSVM, rfeLR, and rfeXGB
Computational Analysis and Evaluation
Classification with SVM, LR, and XGB
Evaluation of models using accuracy, precision, and recall using
10-fold cross validation
Post-analysis of highest-performing model:
Rank sites based on paired t-test + analyze genes of top 15 sites using DAVID and literature review + conduct GSEA using top 500 genes to determine genes that have been linked to ASD and reveal potential biomarkers and involved biological pathways

# Method - Dataset

The GSE27044 dataset (Barwick BG et al. 2015) was used for this analysis.

- 27,578 methylation sites from peripheral blood leukocytes
- 488 pediatric-age males with ASD, 482 of their unaffected brothers, and 158 of indeterminable class. This analysis considered only the 970 samples whose class is known.
- Different ethnicities but a white majority.

# Method - Preprocessing and Feature Selection

- 397 missing values (0.00148%) - k-Nearest Neighbor Imputer (kNNImputer) from Scikit-learn was used (Troyanskaya et al., 2001).

- Feature Selection Methods:
  - 2,500 features: **LinearSVC** (Feng et al., 2019)
  - 2,000 features: **ANOVA** (Sthle and Wold 1989), **Mutual information** (Kraskov et al., 2004), **Chi2** (H. Zhang et al., 2014).
  - 710 features: **Recursive Feature Elimination** (Guyon et al. 2002).

# Method - Classification and Evaluation

- Classification Methods:
  - Support Vector Machine (Cortes and Vapnik, 1995).
  - Logistic Regression (Montgomery et al., 2021).
  - XGBoost (Chen and Guestrin, 2016).
- Evaluation:
  - 10-fold cross validation (Kohavi, 1995) + accuracy, precision, and recall

# Results

| Classifier | Feature Selection Method | Accuracy | Precision | Recall |
|---|---|---|---|---|
| SVM | ANOVA | 99.79% | 99.80% | 99.80% |
| | MI | 99.18% | 99.39% | 98.98% |
| | Chi2 | 99.18% | 99.17% | 99.16% |
| LR | ANOVA | 90.93% | 92.17% | 92.21% |
| | MI | 92.37% | 92.78% | 92.21% |
| | Chi2 | 92.47% | 93.23% | 91.80% |
| XGB | ANOVA | 66.70% | 68.00% | 65.35% |
| | MI | 65.26% | 64.99% | 66.77% |
| | Chi2 | 67.11% | 67.38% | 67.48% |

***Summary of Performances***

| Existing Work | Accuracy of Existing Work | Accuracy Achieved |
|---|---|---|
| Gopinathan & Geetha (2023) | 99.00% (500 CpG sites) | 99.79% (640 CpG sites) |
| Feng et al. (2019) | 99.70% (678 CpG sites) | |

***Comparison with Existing Work***

# Method - Analysis

- Paired t-test to rank features (Hsu & Lachenbruch, 2014; Kim, 2015)- method of comparing paired data to determine the significance of features (to reject the null hypothesis). Using a paired t-test would account for the differences between sibling pairs. An adjusted p-value (Benjamini and Hochberg, 1995) was used to determine significance.

- Gene Set Enrichment Analysis - compares selected genes to existing gene sets to find overlaps (Subramanian et al., 2005)

# Results - Some Key Genes

| Site | Location | Gene | Function |
|---|---|---|---|
| cg12010995 | Intronic | CYP7A1 | CYP7A1 codes for an enzyme involved in bile acid synthesis and has been linked to impaired liver regeneration (L. Zhang et al., 2009). While gastrointestinal abnormalities have been associated with ASD (Nadeem et al., 2021), this gene has no clear link in existing literature. |
| cg23477967 | Exonic | TFPI | TFPI plays a role in blood coagulation. However, apart from its primary function, it has been linked to roles in innate immunity, microbial defense, inflammation (Massberg et al., 2010). It has also been linked to venous thrombosis (Schmidt et al., 1999). |
| cg21660392 | Intronic | ABCA8 | While the exact function is unknown, the gene is highly expressed in the heart, liver, and muscle. It may play a role in transporting substances across the blood-brain barrier (Albrecht and Viturro 2007). |

# Results - Some Key Genes

| Site | Location | Gene | Function |
|---|---|---|---|
| cg11161873 | Exonic | LSMEM1 | The exact function of the protein is unknown. However, it has been linked to Parkinson's disease and Alzheimer's disease, suggesting neurological function (Whittle et al. 2024; Bergen et al. 2015). |
| cg04223956 | Intronic | NEK7 | NEK7 has been linked to gastric cancer progression (Li et al., 2021) and is involved in protein phosphorylation and regulation of mitotic cell cycle. |
| cg13059335 | Intronic | ADAMDEC1 | It is involved in proteolysis, immune response, and negative regulation of cell adhesion. It has been found that it is expressed exclusively in the gastrointestinal tract and has been linked to inflammatory diseases and cancers (Kumagai et al., 2020). |

# Results - Key Pathways

| Gene Set | No. of Overlaps | % of Selected Genes | p-value (↑) |
|---|---|---|---|
| Genes differentially expressed in ovarian cancer patients | 62 | 9.69% | $4.79\ e^{-15}$ |
| Abnormal immune system physiology | 60 | 9.38% | $9.17\ e^{-14}$ |
| Increased inflammatory response | 42 | 6.56% | $6.87\ e^{-10}$ |
| Abnormal respiratory system physiology | 46 | 7.19% | $1.81\ e^{-9}$ |
| Abnormal reproductive system morphology | 40 | 6.25% | $2.95\ e^{-9}$ |
| Abnormal intestine morphology | 33 | 5.16% | $3.21\ e^{-9}$ |
| Abnormal digestive system morphology | 38 | 5.94% | $4.00\ e^{-9}$ |
| Abnormal lung morphology | 40 | 6.25% | $5.17\ e^{-9}$ |

# Conclusions

- The genes CYP7A1, NEK7, TFPI, LSMEM1, ABCA8, and ADAMDEC1 could play a key role in ASD, despite not having been seen previously. These genes correspond to the CpG sites cg12010995, cg04223956, cg23477967, cg11161873, cg21660392, and cg13059335.
- Also of note, genes related to the reproductive and respiratory systems were found to have significant overlap with the selected genes, despite minimal existing research on this link. Additionally, the immune and gastrointestinal systems were revealed to be significantly involved in ASD.

# Implications

- Genes hold promise as biomarkers for the diagnosis and prediction of ASD.
- Reveals previously unrecognized links, offering insights into potential symptomatology and therapeutic avenues for ASD and related disorders.
- Reinforces existing evidence of associations between ASD and the gastrointestinal and immune systems, elucidating the physiological effects. However, there is still minimal existing investigation about all these pathways in ASD. Understanding these complex interactions could significantly advance diagnostic approaches.
- Machine learning approach could be made more suited for diverse populations.
- Future research should investigate the roles of KERA, CSN3, C21orf94, ZNF619, MTTP, and PKIA in ASD.

# References


- Albrecht, C., & Viturro, E. (2007). The ABCA subfamily—gene and protein structures, functions and associated hereditary diseases. Pflügers Archiv - European Journal of Physiology, 453(5), 581–589.
- Arnett, A. B., Rhoads, C. L., Hoekzema, K., Turner, T. N., Gerdts, J., Wallace, A. S., Bedrosian-Sermone, S., Eichler, E. E., & Bernier, R. A. (2018). The autism spectrum phenotype in ADNP syndrome. Autism Research: Official Journal of the International Society for Autism Research, 11(9), 1300–1310.
- Barwick BG, Alisch RS, Chopra P, Warren ST (2015). Gene Expression Omnibus GEO Omnibus GSE27044, (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27044). [DATASET]
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B, Statistical Methodology, 57(1), 289–300.
- Bergen, A. A., Kaing, S., ten Brink, J. B., Netherlands Brain Bank, Gorgels, T. G., & Janssen, S. F. (2015). Gene expression and functional annotation of human choroid plexus epithelium failure in Alzheimer's disease. BMC Genomics, 16, 956.
- Charman, T., & Baird, G. (2002). Practitioner review: Diagnosis of autism spectrum disorder in 2- and 3-year-old children. Journal of Child Psychology and Psychiatry, and Allied Disciplines, 43(3), 289–305.
- Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273–297.
- Depino, A. M. (2013). Peripheral and central inflammation in autism spectrum disorders. Molecular and Cellular Neurosciences, 53, 69–76.
- Feng, X., Hao, X., Xin, R., Gao, X., Liu, M., Li, F., Wang, Y., Shi, R., Zhao, S., & Zhou, F. (2019). Detecting Methylomic Biomarkers of Pediatric Autism in the Peripheral Blood Leukocytes. Interdisciplinary Sciences, Computational Life Sciences, 11(2), 237–246.
- Gopinathan, A., & Geetha, P. (2023). Autism Gene Subset Selection from Microarray data – A Wrapper Approach: AUTISM GENE SUBSET FROM MICROARRAY DATA – A WRAPPER APPROACH. Journal of Scientific & Industrial Research (JSIR), 82(08), 841–850.
- Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. Machine Learning, 46(1), 389–422.

# References


- Hsu, H., & Lachenbruch, P. A. (2014). PairedtTest. In Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat05929
- Kim, T. K. (2015). T test as a parametric statistic. Korean Journal of Anesthesiology, 68(6), 540–546.
- Klin, A. (2018). Biomarkers in Autism Spectrum Disorder: Challenges, Advances, and the Need for Biomarkers of Relevance to Public Health. Focus , 16(2), 135–142.
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. Proceedings of the 14th International Joint Conference on Artificial Intelligence, 2, 1137-1143.
- Kraskov, A., Stögbauer, H., & Grassberger, P. (2004). Estimating mutual information. Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics, 69(6 Pt 2), 066138.
- Kumagai, T., Fan, S., & Smith, A. M. (2020). ADAMDEC1 and Its Role in Inflammatory Disease and Cancer. Metalloproteinases in Medicine, 7, 15–28.
- Li, Y.-K., Zhu, X.-R., Zhan, Y., Yuan, W.-Z., & Jin, W.-L. (2021). NEK7 promotes gastric cancer progression as a cell proliferation regulator. Cancer Cell International, 21(1), 438.
- Massberg, S., Grahl, L., von Bruehl, M.-L., Manukyan, D., Pfeiler, S., Goosmann, C., Brinkmann, V., Lorenz, M., Bidzhekov, K., Khandagale, A. B., Konrad, I., Kennerknecht, E., Reges, K., Holdenrieder, S., Braun, S., Reinhardt, C., Spannagl, M., Preissner, K. T., & Engelmann, B. (2010). Reciprocal coupling of coagulation and innate immunity via neutrophil serine proteases. Nature Medicine, 16(8), 887–896.
- Matson, J. L., Wilkins, J., & Macken, J. (2008). The Relationship of Challenging Behaviors to Severity and Symptoms of Autism Spectrum Disorders. Journal of Mental Health Research in Intellectual Disabilities, 2(1), 29–44.
- McCarty, P., & Frye, R. E. (2020). Early Detection and Diagnosis of Autism Spectrum Disorder: Why Is It So Difficult? Seminars in Pediatric Neurology, 35, 100831.
- Montgomery, D. C., Peck, E. A., & Geoffrey Vining, G. (2021). Introduction to Linear Regression Analysis. John Wiley & Sons.
- Moore, L. D., Le, T., & Fan, G. (2013). DNA methylation and its basic function. Neuropsychopharmacology: Official Publication of the American College of Neuropsychopharmacology, 38(1), 23–38.

# References


- Nadeem, M. S., Murtaza, B. N., Al-Ghamdi, M. A., Ali, A., Zamzami, M. A., Khan, J. A., Ahmad, A., Rehman, M. U., & Kazmi, I. (2021). Autism - A Comprehensive Array of Prominent Signs and Symptoms. Current Pharmaceutical Design, 27(11), 1418–1433.
- Sanders, S. J., Campbell, A. J., Cottrell, J. R., Moller, R. S., Wagner, F. F., Auldridge, A. L., Bernier, R. A., Catterall, W. A., Chung, W. K., Empfield, J. R., George, A. L., Jr, Hipp, J. F., Khwaja, O., Kiskinis, E., Lal, D., Malhotra, D., Millichap, J. J., Otis, T. S., Petrou, S., … Bender, K. J. (2018). Progress in Understanding and Treating SCN2A-Mediated Disorders. Trends in Neurosciences, 41(7), 442–456.
- Schmidt, M., Götting, C., Schwenz, B., Lange, S., Müller-Berghaus, G., Brinkmann, T., Prohaska, W., & Kleesiek, K. (1999). The 536C→T transition in the human tissue factor pathway inhibitor (TFPI) gene is statistically associated with a higher risk for venous thrombosis. Thrombosis and Haemostasis, 82(07), 1–5.
- Sthle, L., & Wold, S. (1989). Analysis of variance (ANOVA). Chemometrics and Intelligent Laboratory Systems, 6(4), 259–272.
- Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., 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, 102(43), 15545–15550.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., & Altman, R. B. (2001). Missing value estimation methods for DNA microarrays. Bioinformatics , 17(6), 520–525.
- Vogt, D., Cho, K. K. A., Lee, A. T., Sohal, V. S., & Rubenstein, J. L. R. (2015). The parvalbumin/somatostatin ratio is increased in Pten mutant mice and by human PTEN ASD alleles. Cell Reports, 11(6), 944–956.
- Whittle, B. J., Izuogu, O. G., Lowes, H., Deen, D., Pyle, A., Coxhead, J., Lawson, R. A., Yarnall, A. J., Jackson, M. S., Santibanez-Koref, M., & Hudson, G. (2024). Early-stage idiopathic Parkinson’s disease is associated with reduced circular RNA expression. NPJ Parkinson’s Disease, 10(1), 25.

# References


- Yu, Y., Zhang, B., Ji, P., Zuo, Z., Huang, Y., Wang, N., Liu, C., Liu, S.-J., & Zhao, F. (2022). Changes to gut amino acid transporters and microbiome associated with increased E/I ratio in Chd8+/- mouse model of ASD-like behavior. Nature Communications, 13(1), 1151.
- Zeidan, J., Fombonne, E., Scorah, J., Ibrahim, A., Durkin, M. S., Saxena, S., Yusuf, A., Shih, A., & Elsabbagh, M. (2022). Global prevalence of autism: A systematic review update. Autism Research: Official Journal of the International Society for Autism Research, 15(5), 778–790.
- Zhang, H., Li, L., Luo, C., Sun, C., Chen, Y., Dai, Z., & Yuan, Z. (2014). Informative gene selection and direct classification of tumor based on Chi-square test of pairwise gene interactions. BioMed Research International, 2014, 589290.
- Zhang, L., Huang, X., Meng, Z., Dong, B., Shiah, S., Moore, D. D., & Huang, W. (2009). Significance and mechanism of CYP7a1 gene regulation during the acute phase of liver regeneration. Molecular Endocrinology , 23(2), 137–145.