HIGH SCHOOL EDITION
Journal of Student Research
Volume 13 Issue 1 (2024)

Statistical Analysis of Methylation Reveals Epigenetic Biomarkers for High- Functioning Autism Spectrum Disorder

Aarav Sharma[1] and Hayan Lee[#]

[1]Archbishop Mitty High School, USA
[#]Advisor

ABSTRACT

Autism spectrum disorder (ASD) is a neurological disorder that affects social behavior and learning ability. Although studies have identified certain regions of the genome (22q13.33) linked to the increased risk of ASD, there is no clear etiology, which is why psychologists theorize that ASD is a complex interaction between genetic and environmental factors. The current study was designed to identify genes with a positive correlation toward the diagnosis of high-function autism spectrum disorder (HF-ASD). Samples were obtained from National Center for Biotechnology Information (NCBI) containing 450K CpG methylation from 69 subjects diagnosed as HF-ASD or typically developing (TD). We applied Principal Component Analysis (PCA) and further utilized two sample Kolmogorov-Smirnov test (KS-test) to obtain the p-values for each patient. False Discovery Rate (FDR Correction) is computed by Benjamini-Hochberg Procedure. Based on the output of the FDR method, we identified the top ten genes that have CpGs with statistically differentially methylated. Such potential epigenetic biomarker includes WBP11, LDHB, DHX29, NUB1, C2orf67, SNAI2, MTOR, CLCN3, SYNJ2BP, and TTK. Two sample KS test and FDR Correction code is available at: https://www.kaggle.com/code/aaravsharma123/gse109905-fdr- pvalues/notebook.

Dataset Information

The DNA methylation data set was acquired from Kimura et al. [1]. The data files were downloaded from the NCBI Gene Expression Omnibus (GEO) under Accession Number GSE109905. This dataset extracted peripheral blood samples using the QIAamp DNA blood midi kit. It consists of 69 patients, of which 38 patients were high functioning ASD (HF-ASD) and the other 31 patients were Typically Developing (TD). Each patient was either male or female under the age of 18. DNA methylation profiles were assayed using the Illumina 450,000 methylation array. Additionally, there are 410,000 beta-methylation values for all 69 subjects [2].

Methods

Principal component analysis (PCA) [3] performs dimension reduction and projects data points into 2D space. We utilized PCA on DNA methylation data of 69 HF-ASD and TD Samples (Fig.1). The separation between the two groups is shown but not so clear, reflecting the difficulties and complexity of the neurodevelopmental disorder.

Further, we used Kolmogorov-Smirnov (two sample KS- test) [4]. We hypothesized that all data came from the same group and performed two sample KS-test, which suggests how likely data points of two groups are in the same distribution. Two-sample KS-test was applied per chromosome gene id (CG ID) and the p-

value was computed per CG ID. We immediately filtered those p- values to determine which CG ID rejects the null hypothesis, i.e., p-values less than or equal to 0.05.

Out of 410,000 CpGs, we were able to identify 71 CG IDs with p-values less than or equal to 0.05.

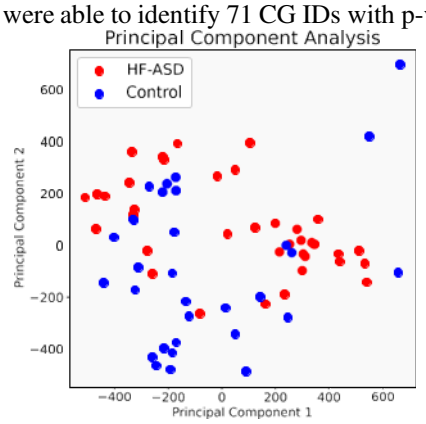

Figure 1. Principal Component Analysis plot comparing High Function ASD patients with Typically Developing patients.

The p-values are further adjusted by Benjamini-Hochberg (BH) procedure to compute the False Detection Rate (FDR Correction) [5]. FDR correction is a statistical approach that validates whether certain features that appear to be significant are truly null. We utilized Python's statsmodels library with default parameter settings. This returned outputs of 'reject' and corrected p-values. The reject variable consistently provided a single binary value of true or false, where true means a reject of the null hypothesis. We then sorted the FDR values to select the top 10 chromosome genes. We were unable to map all of the CG IDs to their correlating genes as some of them were not present in the data file provided by the National Center for Biotechnology Information (NCBI) [6]. Due to this reason, we needed to select the top 20 FDR values, and hence, the top 20 CG IDs. As shown in Table 1 and Table 2 below, we were able to identify their respective genes and functions, chromosomes, and chromosomal coordinates.

Table 1. Chromosome Genes Sorted Based On Their Respective False Detection Rate Values

CG id	Chr	Position	FDR	Gene
cg05515957	12	14956570	0.000044	WBP11; C12orf60
cg08148261	12	21810960	0.000073	LDHB
cg05225431	5	54604207	0.000073	SKIV2L2; DHX29
cg08134068	7	151039078	0.000094	NUB1
cg04052427	2	211035360	0.000142	C2orf67
cg00116554	8	49833833	0.000142	SNAI2
cg07029998	1	11322191	0.000142	MTOR
cg00069314	4	170542039	0.000142	CLCN3
cg05151393	14	70883883	0.000142	SYNJ2BP
cg08044663	6	80713582	0.000142	TTK

Table 2. Chromosome Genes Description

Gene	Description
WBP11;	Activates pre-mRNA splicing and may inhibit protein phosphatase I activity

C12orf60	
LDHB	Predominantly expressed in aerobic tissues such as cardiac muscle
SKIV2L2; DHX29	ATP-binding RNA helicase involved in translation initiation
NUB1	Predominantly shown in the Nervous system and acts as a protein coding gene
C2orf67	A potential marker reflecting the activation status of EGF signaling pathway
SNAI2	Role in the formation of tissues during embryonic development
MTOR	Expressed in numerous tissues, with highest levels in testis and Nervous System
CLCN3	Expressed primarily in tissues derived from neuroectoderm in the hippocampus
SYNJ2BP	Predicted to enable type II activin receptor binding activity
TTK	Evidence present in proliferating cell lines, Nervous System, and lymph nodes

Results

We were able to identify the top 10 CG IDs and their corresponding genes by combining our original dataset with the one provided by NCBI [6]. As shown in Table 1, the following genes were found to have a significant correlation with the diagnosis: WBP11, LDHB, DHX29, NUB1, C2orf67, SNAI2, MTOR, CLCN3, SYNJ2BP, and TTK.

Out of those ten genes, the four most outstanding CG IDs with matching gene names, WBP11, LDHB, DHX29, and NUB1, were found to be engaged in a variety of systems and tissues in the body, not just in the brain level like we hypothesized. For instance, LDHB, which is expressed in the cardiac muscle, was previously linked to be Maternal- Autoantibody Related (MAR) Autism [7], but through our study, its impact on autism is significantly greater.

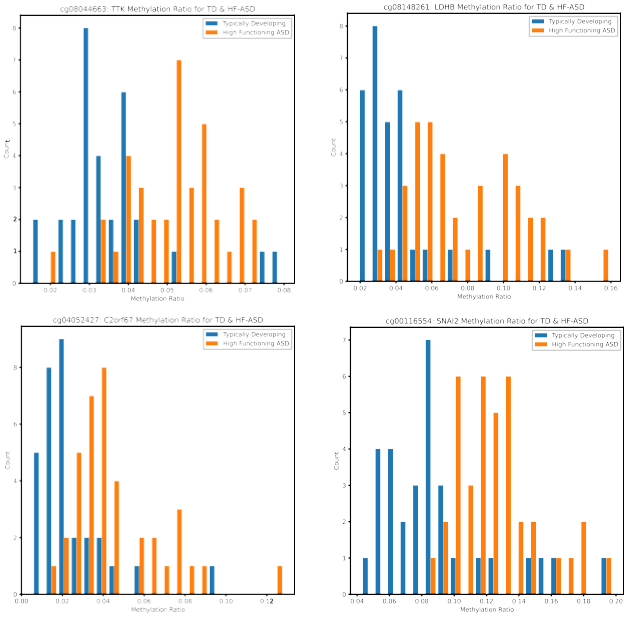

Figure 2. Histograms illustrating the most promising genes between the high functioning autism spectrum disorder (HF-ASD) subjects and typically developing (TD) subjects.

Discussion

Autism Spectrum Disorder (ASD) has dramatically increased over the past several decades. Despite certain regions of the genome being linked to autism [8], there is no clear etiology of ASD. Due to this reason, psychologists postulate that environmental factors, along with genetic factors, may play a key role in ASD [9]. This study introduces a novel approach towards the identification of genes with the most significant positive correlation towards the diagnosis of HF-ASD and TD.

Our study utilizes a public dataset provided by Kimura et al. [2]. We first applied PCA for the purpose of visualizing our dataset and comparing HF-ASD patients with TD patients.

KS-test was then applied onto the data to compute p- values. Further, we utilized FDR correction to confirm our p- values obtained from KS-test. Finally, we mapped the CG IDs associated with their p-values with the genes found on another dataset downloaded from NCBI [6]. Table 1, Table 2, and Fig. 2 summarize our findings: cg08044663 (FDR:0.000142, TTK), cg08148261 (FDR:0.000073, LDHB), cg04052427 (FDR:0.000142, C2orf67), and cg00116554 (FDR:0.000142, SNAI2). Our code can be found at: https://www.kaggle.com/code/aaravsharma123/gse109905- fdr-pvalues/notebook.

Although ten genes were identified to have a significant effect on the diagnosis of HF-ASD, a limitation of the current study is the small group size. Our dataset consists of 69 subjects, 38 of which are HF-ASD and the rest are TD. Although 69 subjects are a good representation of the total number of HF-ASD and TD patients, an expanded group size would be more beneficial to ensure the accuracy of KS-test and FDR correction.

Another limitation of this study stems from single omics data. This dataset solely contains methylation data. In the future, we hope to acquire other types of omics data, including mutation, copy number alterations, or gene- expression datasets to validate our findings. Additionally, we would like to conduct a three-way analysis among HF-ASD, ASD, and TD patients and note the particular differences among the three groups.

Acknowledgments

We would like to thank Michael Snyder at the Snyder Lab in Stanford University. Also, we would like to thank Kimura et al. for publishing his dataset.

References

[1] Kimura, Ryo, et al. “An Epigenetic Biomarker for Adult High- Functioning Autism Spectrum Disorder.” Nature News, Nature Publishing Group, 20 Sept.2019, https://www.nature.com/articles/s41598-019-50250-9#data- availability.
[2] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE109905.
[3] “PCA.” Pca Visualization in Python, https://plotly.com/python/pca-visualization/.
[4] “Scipy.stats.ks_2samp#.” Scipy.stats.ks_2samp - SciPy v1.9.1 Manual, https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ks_2samp.html.
[5] “Statsmodels.stats.multitest.fdrcorrection” Statsmodels, https://www.statsmodels.org/dev/generated/statsmodels.stats.mult itest.fdrcorrection.html.
[6] https://www.kaggle.com/datasets/hayan2021lee/hm450coords
[7] Marks, Katya, et al. “Maternal-Autoantibody-Related (MAR) Autism: Identifying Neuronal Antigens and Approaching Prospects for Intervention.” Journal of Clinical Medicine, MDPI, 7Aug.2020,

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7465310/.

[8] Zhu, Yihui, et al. “Placental Methylome Reveals a 22q13.33 Brain Regulatory Gene Locus Associated with Autism - Genome Biology.” BioMed Central, BioMed Central, 16 Feb. 2022, https://genomebiology.biomedcentral.com/articles/10.1186/s1305 9-022-02613-1.

[9] Tordjman, Sylvie, et al. “Gene × Environment Interactions in Autism Spectrum Disorders: Role of Epigenetic Mechanisms.” Frontiers, Frontiers, 1 Jan. 1AD, https://www.frontiersin.org/articles/10.3389/fpsyt.2014.00053/ful l.