Genetic Insights into ASD: Lasso Selected Genes Versus Ribosomal Markers

Hanbei Xiong[1]
[1]University of California, Los Angeles

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a disorder of the brain characterized by difficulties in emotional, verbal and non-verbal expression and social interaction. Autism is largely linked to genetic factors, so identifying important mutations in the genes involved is critical in helping researchers better understand the disorder and designing effective genetic screening tools for parents to avoid passing on the genetic change to their children (Chaste 2012). Given this background, understanding the biological insights behind gene expression data and selecting important genes related to ASD can be a pivotal approach.

Prior research indicates that the copy number of active ribosomal genes among individuals with ASD significantly deviates from that observed in the general healthy population (Porokhovnik 2015). Moreover, an investigation into Autism-related genes, employing unpaired t-tests and fold change analyses, unveiled a statistically noteworthy representation of ribosomal genes (Kuwano, 2011). Therefore, assessing the expression of ribosomal genes in mothers of children with autism is worth exploring. In the field of gene selection methods, the Least Absolute Shrinkage and Selection Operator (LASSO) plays a huge role in cancer classification based on gene expression data (Zhang, 2011). Based on these prior studies, our methods and investigations have implications in autism research.

In this study, a retrospective study design is carried out to identify important genetic changes between 21 healthy mothers having children with ASD and 21 healthy mothers having healthy children using DNA microarray gene expression profiling in peripheral blood. The objective of this investigation is to pinpoint noteworthy genetic variations in mothers linked to having offspring with ASD using Lasso feature selection. Additionally, we aim to assess the effectiveness of identifying mothers with ASD offspring by comparing the predictive capabilities of the Lasso selected genes against some ribosomal markers identified by Kuwano.

II. METHODS & MATERIALS

A. Data Set

The data used for this study is obtained from the Gene Expression Omnibus (GEO) (https://www.ncbi.nlm.nih.gov/geo/). The original dataset includes 84 observations which can be broken into four subgroups: 21 young adults with autism spectrum disorder (ASD), 21 age- and gender-matched healthy subjects (control), 21 healthy mothers having children with ASD (asdMO), and 21 healthy mothers having healthy children (ctrlMO). The ASD group and asdMO group are drawn from different pools of volunteers or specialized clinics, and they are not biologically related to each other. Other groups were formed using healthy volunteers matched for age and sex, drawn from students or faculty at the Faculty of Medicine, University of Tokushima, NCNP, and local communities. These individuals were enrolled as control subjects for both the ASD and asdMO groups. The subjects underwent thorough screenings to confirm the absence of any serious physical or mental conditions, including ASD, in their past and present. Children of the ctrlMO group were interviewed to ensure no ASD diagnoses and maintained good health, while all control subjects refrained from medication for at least three months before recruitment.

B. Study Population

This study focuses on comparing the 21 observations from the asdMO group and 21 observations from the ctrlMO group. Each of 19194 features represents normalized differential gene expression associated with each observation.

Participant demographics could be seen in Appendix Table 1. The distribution of age between asdMO and ctrlMO are similar with only slight differences on ranges. Gender is completely matched. We split the 42 observations into a training set and test set. Training set includes 30 observations with 15 coming from the asdMO group and 15 coming from the ctrlMO group. The test set contains the remaining observations. Therefore, we

can conclude that our training data and testing data are distributed evenly.

C. Statistical Methods

Lasso feature selection is conducted to select 9 important genes and use Lasso regression to fit a linear model with selected genes. We fit a logistic regression model with 9 significant ribosomal genes and compare the AUC score with the result given by Lasso regression. Two sample t tests with unequal variance assumption are used to evaluate the significance of all selected genes between asdMO group and ctrlMO group. All experiments are conducted using software R 4.3.2.

III. RESULTS

The experiment selected 9 significant genes using Lasso feature selection and fit in a Lasso regression. Another 9 ribosomal genes were randomly selected from Appendix Table 2 and fit into a logistic regression. The two sets of genes have no overlaps. The distributions of Lasso selected gene expression values within each group are displayed in Figure 1. The distributions of selected ribosomal gene expression values within each group are displayed in Figure 2, It is evident that the asdMO(1) group exhibits consistently lower gene expression values compared to the ctrlMO(0) group for these ribosomal genes.
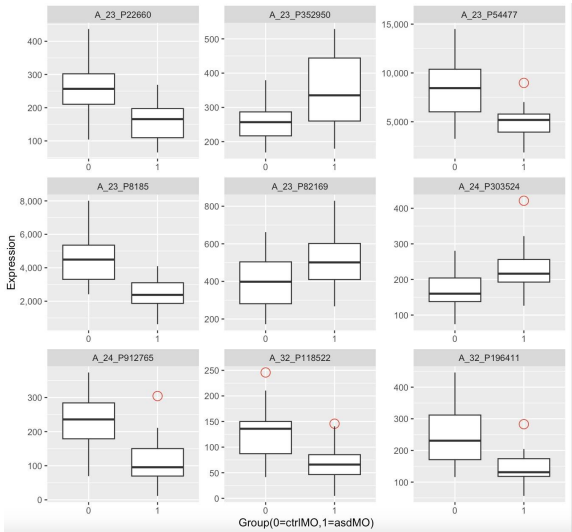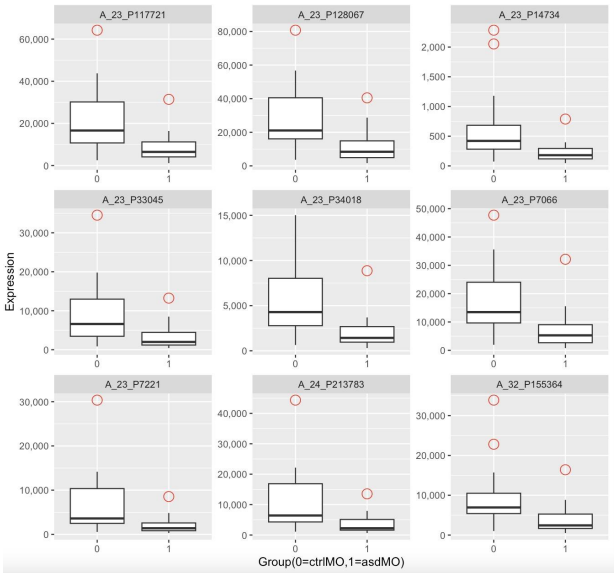

Figure 2: Expression Distribution
of 9 Ribosomal Genes

We validate the significance of Lasso selected genes and ribosomal genes using two sample t tests with unequal variance assumption. Information of selected genes and P values are displayed in Table 1. All differences of gene expression values between two groups are statistically significant to conclude the two groups are different in selected gene expression. Three unlabeled genes are identified by Lasso. Among these three unknown gene, asdMO group displays consistently lower expression values than the ctrlMO group as shown in Figure 1. The locations of these three genes in the human genome are presented in the description section of Table 1 part (a).

Table 1: Selected Genes Description

Figure 1: Expression Distribution of 9 Lasso
Selected Genes

(a) Lasso Selected Genes

Label	Gene Symbol	Description	p-value
A_23_P22660	CYSLTR1	cysteinyl leukotriene receptor 1	1.09×10^{-4}
A_23_P352950	PNMA5	paraneoplastic antigen like 5	1.63×10^{-3}
A_23_P54477	NOP10	ribonucleoprotein homolog	5.73×10^{-5}
A_23_P8185	DYNLT1	dynein, light chain, Tctex-type 1	5.25×10^{-6}
A_23_P82169	SOX4	SRY (sex determining region Y)-box 4	2.83×10^{-2}
A_24_P303524	MICALL2	MICAL-like 2	2.68×10^{-3}
A_24_P912765	Unknown	Location: chr2:144084039-144084098	1.72×10^{-5}
A_32_P118522	Unknown	Location: chr11:010877147-010877206	1.50×10^{-4}
A_32_P196411	Unknown	Location: chr19:052646336-052646277	2.33×10^{-4}

(b) Ribosomal Genes

Label	Gene Symbol	Description	p-value
A_23_P117721	RPS17	ribosomal protein S17	1.72×10^{-5}
A_23_P128067	RPL41	ribosomal protein L41	2.44×10^{-3}
A_23_P14734	RPS27L	ribosomal protein S27-like	6.22×10^{-3}
A_23_P33045	RPL26	ribosomal protein L26	6.52×10^{-3}
A_23_P34018	RPL39	ribosomal protein L39	1.31×10^{-3}
A_23_P7066	RPL9	ribosomal protein L9	2.07×10^{-3}
A_23_P7221	RPL34	ribosomal protein L34	7.83×10^{-3}
A_24_P213783	RPL31	ribosomal protein L31	6.13×10^{-3}
A_32_P155364	RPL7	ribosomal protein L7	8.16×10^{-3}

Figure 3 shows the ROC curves of two models. The logistic regression model with ribosomal genes has AUC score 0.444 and Lasso regression model has AUC score 0.944 which indicates that the Lasso regression model shows much better predictive power than logistic regression model with only ribosomal genes.
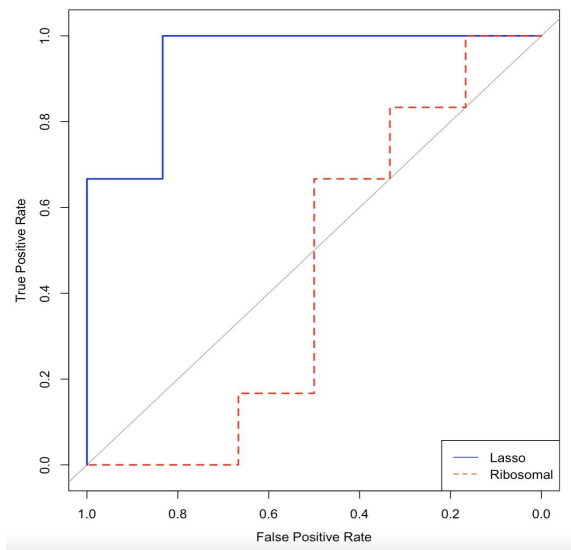

Figure 3: ROC Curve for Comparison

IV. CONCLUSIONS

In conclusion, our study successfully identified nine key genes associated with mothers of children with ASD using Lasso feature selection. Six of these genes were previously labeled, and three were unknown. Comparing our findings with ribosomal markers, we observed that Lasso-selected genes outperformed in predicting mothers with ASD offspring. This highlights the potential clinical relevance of our identified genes as superior markers for identifying at-risk mothers.

V. DISCUSSION

The exploration of genetic factors associated with Autism Spectrum Disorder (ASD) holds paramount importance in unraveling the complexities of this neurodevelopmental condition. Our study delved into identifying significant genetic variations in mothers with children diagnosed with ASD, utilizing a retrospective study design and DNA microarray gene expression profiling.

We discovered 9 genes using Lasso feature selection and extracted 9 ribosomal genes presented in Kuwano's paper (Kuwano 2011). The predictability of Lasso selected genes is more effective in identifying mothers with ASD offspring than ribosomal genes in this study. Although other researchers had concluded that ASD patients tend to have lower ribosomal gene expression than healthy individuals and this pattern in mothers having ASD children had also been illustrated in our study, using ribosomal genes as the only biomarkers is not ideal in determining potential ASD conditions. Moreover, three of our Lasso selected genes are unlabeled. The asdMO group shows consistently lower gene expression values than ctrlMO group for these genes. It is worth investigating the functions of these genes and conducting further analysis to discover their biological relationship with ASD.

Some limitations of this study are that the sample size is too small for us to make general inference on the population. Enrolling more ASD patients and their relatives would greatly increment the credibility of ASD related study. In addition, the biological interpretation of Lasso selected genes can be carried out by Ingenuity Pathways Analysis (IPA) which can provide more biological understanding of the intercorrelation of genes and causality of every aspect (Krämer 2014). Since ASD is a heterogeneous disorder which can also relate to environmental factors, more factors should be taken into consideration for future research.

REFERENCES

1. Zhang, Songfeng et al.,. "An experimental comparison of gene selection by Lasso and Dantzig selector for cancer classification," Computers in Biology and Medicine, Volume 41, Issue 11, pp. 1033-1040, 2011.
2. Porokhovnik, Lev N et al. "Active ribosomal genes, translational homeostasis and oxidative stress in the pathogenesis of schizophrenia and autism." *Psychiatric genetics* volume 25, Issue 2, pp. 79-87, 2015.
3. Kuwano, Yuki et al. "Autism-associated gene expression in peripheral leucocytes commonly observed between subjects with autism and healthy women having autistic children." *PloS one* Volume 6, Issue 9, e24723, 2011.
4. Chaste, Pauline, and Marion Leboyer. "Autism risk factors: genes, environment, and gene-environment interactions." *Dialogues in clinical neuroscience* Volume 14, Issue 3, pp. 281-92, 2012.
5. Krämer, Andreas et al. "Causal analysis approaches in Ingenuity Pathway Analysis." *Bioinformatics (Oxford, England),* Volume 30, Issue 4, pp. 523-530, 2014.

APPENDIX

I.   Source Data File

Due to some issues in raw datasets format, slight manual modifications of format on txt files were mandatory. The modified files used in R are uploaded to Bruinlearn. The raw public datasets can be found in the following links:

(1) Sample Labels and Gene Expression:
https://ftp.ncbi.nlm.nih.gov/geo/series/GSE26nnn/GSE26415/matrix/

(2) Gene Labels:
https://ftp.ncbi.nlm.nih.gov/geo/series/GSE26nnn/GSE26415/miniml/

Modified File name in Bruin learn:

(1) Sample Labels and Gene Expression:
Sample.status.txt
austismData.txt

(2) Gene Labels:
GPL6480-tbl-1.txt

I.   R data set

The cleaned data set of gene expression, sample labels, and gene labels used for our study can be obtained by running our R code provided in Appendix. The cleaned datasets are uploaded into Bruinlearn.
ASD_expression_all.csv
training_set_Ribosomal.csv
testing_set_Ribosomal.csv
training_set_LASSO.csv
testing_set_LASSO.csv

II.   Additional Appendix Tables

Appendix Table 1: Participants Demographics

Group	Males	Females	Age (mean±SD)	Age Range
ASD	17	4	26.7 ± 5.5	18–38
ASD Control	17	4	27.0 ± 5.5	19–39
Mother with ASD Child	0	21	44.7 ± 6.7	33–58
Mother Control	0	21	44.7 ± 6.7	31–59

.

Appendix Table 2: Differentially expressed genes for the asdMO group compared with the ctrlMO

group from IPA analysis presented by Kuwano.

Accession No.	GeneSymbol	Description	fold change	corrected p-value
Up-regulated gene				
AK024445	C14ORF56	chromosome 14 open reading frame 56	2.85	1.16E-03
NM_020478	ANK1	ankyrin 1, erythrocytic	2.73	1.32E-03
NM_015431	TRIM58	tripartite motif-containing 58	2.66	3.81E-03
NM_138368	DKFZP761E198	DKFZp761E198 protein	2.54	2.40E-03
BC009106	LZTR2 (SEC16B)	SEC16 homolog B (S. cerevisiae)	2.33	4.54E-02
NM_000419	ITGA2B	integrin, alpha 2b (platelet glycoprotein IIb of IIb/IIIa complex, antigen CD41)	2.29	6.45E-03
NM_001266	CES1	carboxylesterase 1 (monocyte/macrophage serine esterase 1)	2.23	5.81E-03
NM_001020029	C4B	complement component 4B (Chido blood group)	2.21	2.46E-02
NM_001001957	OR2W3	olfactory receptor, family 2, subfamily W, member 3	2.15	1.50E-02
NM_000894	LHB	luteinizing hormone beta polypeptide	2.15	4.37E-02
NM_006121	KRT1	keratin 1	2.13	1.67E-02
NM_002501	NFIX	nuclear factor I/X (CCAAT-binding transcription factor)	2.09	3.31E-03
NM_001039476	NPRL3 (C16ORF35)	nitrogen permease regulator-like 3 (S. cerevisiae)	2.06	4.84E-03
NM_006798	UGT2A1	UDP glucuronosyltransferase 2 family, polypeptide A1	2.06	2.74E-02
NM_001024858	SPTB	spectrin, beta, erythrocytic	2.05	4.05E-03
NM_198149	SHISA4	shisa homolog 4 (Xenopus laevis)	2.02	1.20E-02
NM_007371	BRD3	bromodomain containing 3	2.02	1.11E-03
Down-regulated gene				
NM_002624	PFDN5	prefoldin subunit 5	−2.00	7.37E-03
NM_012459	TIMM8B	translocase of inner mitochondrial membrane 8 homolog B (yeast)	−2.01	8.93E-04
NM_014463	LSM3	LSM3 homolog, U6 small nuclear RNA associated (S. cerevisiae)	−2.01	1.59E-03
AB075859	ZNF525	zinc finger protein 525	−2.01	2.31E-03
NM_002489	NDUFA4	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 4, 9kDa	−2.02	2.77E-03
NM_001803	CD52	CD52 molecule	−2.03	2.58E-03
NM_001019	RPS15A	ribosomal protein S15a	−2.03	3.73E-03
NM_005034	POLR2K	polymerase (RNA) II (DNA directed) polypeptide K, 7.0kDa	−2.05	2.56E-03
NM_001021	RPS17	ribosomal protein S17	−2.06	4.70E-03
NM_021104	RPL41	ribosomal protein L41	−2.09	6.64E-03
NM_001781	CD69	CD69 molecule	−2.09	4.54E-03
NM_005340	HINT1	histidine triad nucleotide binding protein 1	−2.12	1.96E-03
NM_000971	RPL7	ribosomal protein L7	−2.12	4.29E-03
NM_007333	KLRC3	killer cell lectin-like receptor subfamily C, member 3	−2.12	6.86E-03
NM_016304	RSL24D1 (C15orf15)	ribosomal L24 domain containing 1	−2.12	5.56E-03
NM_203495	COMMD6	COMM domain containing 6	−2.15	5.13E-03
NM_015920	RPS27L	ribosomal protein S27-like	−2.17	2.01E-03
NM_019051	MRPL50	mitochondrial ribosomal protein L50	−2.18	1.16E-03
NM_001026	RPS24	ribosomal protein S24	−2.18	2.80E-03
NM_002370	MAGOH	mago-nashi homolog, proliferation-associated (Drosophila)	−2.19	5.86E-04
NM_001192	TNFRSF17	tumor necrosis factor receptor superfamily, member 17	−2.23	1.83E-02
NM_005213	CSTA	cystatin A (stefin A)	−2.23	1.70E-03
NM_002586	PBX2	pre-B-cell leukemia homeobox 2	−2.27	3.17E-03
NM_000987	RPL26	ribosomal protein L26	−2.27	6.54E-03
NM_003096	SNRPG	small nuclear ribonucleoprotein polypeptide G	−2.30	2.26E-03
NM_004049	BCL2A1	BCL2-related protein A1	−2.32	4.72E-03
NM_001000	RPL39	ribosomal protein L39	−2.34	2.94E-03
NM_006294	UQCRB	ubiquinol-cytochrome c reductase binding protein	−2.34	2.79E-03
NM_000985	RPL17	ribosomal protein L17	−2.37	1.61E-03
NM_016093	RPL26L1	ribosomal protein L26-like 1	−2.37	2.28E-03
NM_015235	CSTF2T	cleavage stimulation factor, 3' pre-RNA, subunit 2, 64kDa, tau variant	−2.41	1.80E-03
NM_000661	RPL9	ribosomal protein L9	−2.47	1.82E-03
NM_004280	EEF1E1	eukaryotic translation elongation factor 1 epsilon 1	−2.49	8.51E-04
NM_005127	CLEC2B	C-type lectin domain family 2, member B	−2.53	1.45E-03
NM_000993	RPL31	ribosomal protein L31	−2.57	3.59E-03
NM_006713	SUB1	SUB1 homolog (S. cerevisiae)	−2.59	2.15E-03
NM_001011	RPS7	ribosomal protein S7	−2.59	1.70E-03
NM_033625	RPL34	ribosomal protein L34	−2.64	3.73E-03
BC049823	RPL22L1	ribosomal protein L22-like 1	−2.84	1.16E-03
NM_001006	RPS3A	ribosomal protein S3A	−3.08	1.64E-03

*corrected p-value was calculated by unpaired t test with Benjamini-Hochberg correction for multiple comparisons at the 0.05 FDR.
doi:10.1371/journal.pone.0024723.t003

III.   R Code

Code will be submitted in file called: finalproject.R

#Figure 1
df_long_lasso <- gather(df, key = "Gene", value = "Value", all_of(gene_lasso))

```
# Creating a boxplot with facets for each 'y' variable
ggplot(df_long_lasso, aes(x = as.factor(status), y = Value)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 1, outlier.size = 4) +
  labs(title = "Panel for Selected Genes",
       x = "Group(0=ctrlMO,1=asdMO)",
       y = "Expression") +
  facet_wrap(~ Gene, scales = "free", nrow=3,ncol = 3) +
  scale_x_discrete(labels = c("0", "1")) +  # Replace with appropriate labels for 'status'
  scale_y_continuous(labels = scales::comma)
```

#Figure 2
df_long_ribosomal <- gather(df, key = "Gene", value = "Value", all_of(gene_ribosomal))

```
ggplot(df_long_ribosomal, aes(x = as.factor(status), y = Value)) +
  geom_boxplot(outlier.colour = "red", outlier.shape = 1, outlier.size = 4) +
  labs(title = "Panel for Selected Genes",
       x = "Group(0=ctrlMO,1=asdMO)",
       y = "Expression") +
  facet_wrap(~ Gene, scales = "free", nrow=3,ncol = 3) +
  scale_x_discrete(labels = c("0", "1")) +  # Replace with appropriate labels for 'status'
  scale_y_continuous(labels = scales::comma)
```

#Figure 3
```
plot(auc_score, col = "blue", type = "l", lty = 1, main = "ROC Curve Comparison", xlab = "False Positive
Rate", ylab = "True Positive Rate", ylim = c(0, 1))
lines(auc_score_2, col = "red", type = "l", lty = 2)
legend("bottomright", legend = c("Lasso", "Ribosomal"), col = c("blue", "red"), lty = 1:2)
```

#Table 1
```
# Merge df_gene_selected_Lasso with result_table
merged_df_lasso <- merge(gene_selected_Lasso, result_table, by.x = "ID", by.y = "feature", all.x = TRUE)

# Merge df_gene_selected_IPA with result_table_2
merged_df_ribosomal <- merge(unique_df_ribosomal, result_table_2, by.x = "ID", by.y = "feature", all.x =
TRUE)

#Tabel 1 display
merged_df_lasso
merged_df_ribosomal
```