Biomarkers of ASD

If you want a subtitle put it here

Lucas Childs, Bahaar Ahuja, …

2025-11-05

Use this as a template. Keep the headers and remove all other text. In all, your report can be quite short. When it is complete, render and then push changes to your team repository.

Abstract

This report investigates blood biomarkers for autism spectrum disorder (ASD) with a focus on evaluating how methodological choices influence results. Using the dataset from Hewitson et al. (2021), we examine protein distributions and justify the use of log-transformations to stabilize variance and reduce skewness. We explore subject-level outliers and assess whether their frequency differs between ASD and typically developing (TD) groups. We then evaluate how modifications to protein selection procedures, such as varying the number of top predictive proteins, using training/test partitions, and applying fuzzy intersections, affect classification performance. Finally, we identify alternative protein panels that achieve comparable or improved classification accuracy and benchmark these results against the original in-class analysis. This work highlights the sensitivity of biomarker analyses to preprocessing and methodological decisions and provides guidance for more robust ASD protein panel selection.

Aim

The aim of this report is to explore the sensitivity of ASD biomarker analyses to key methodological choices. Specifically, we examine the effects of log-transforming protein levels, assess the presence of subject-level outliers and their group distribution, and investigate how modifications to protein selection procedures (e.g., training/test splits, number of top proteins, fuzzy intersections) impact classification performance. Finally, we identify alternative protein panels that achieve comparable or improved classification accuracy and benchmark these results against the original in-class analysis.

Dataset

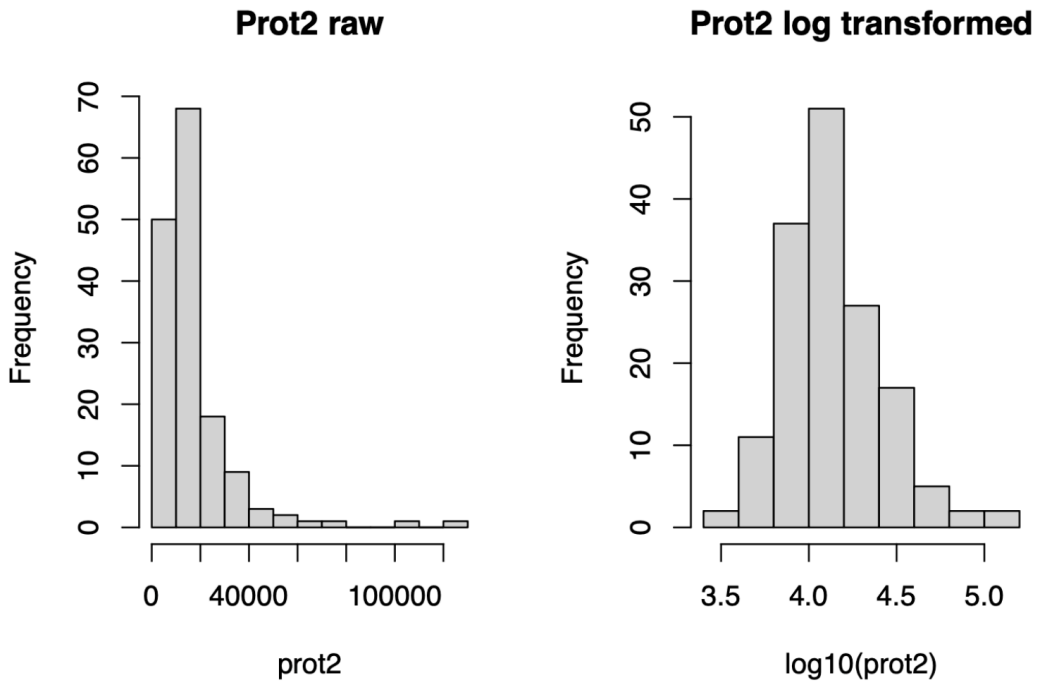
The data consist of blood serum samples from 76 boys with autism spectrum disorder (ASD) and 78 typically developing (TD) boys, aged 18 months to 8 years. Proteomic analysis was performed using SomaLogic's SOMAScan™ 1.3K platform, measuring levels of 1,317 proteins. Two additional variables, ADOS (ASD severity) and group (ASD vs. TD), were included.

Preprocessing involved removing missing values, applying a log transformation to stabilize variance and reduce skewness, centering and scaling each protein to standardize distributions, and trimming extreme outliers to prevent disproportionate influence. For exploratory analyses, we also considered an untrimmed dataset to examine subject-level outliers.

Summary of published analysis

Summarize the methodology of the paper in 1-3 paragraphs. You need not explain the methods in depth as we did in class; just indicate what methods were used and how they were combined. If possible, include a diagram that depicts the methodological design. (Quarto has support for GraphViz and Mermaid flowcharts.) Provide key results: the proteins selected for the classifier and the estimated accuracy.

Q1: A logarithmic transformation was done to the biomarker levels in order to reduce scale, variance, and skewedness of the data. Furthermore, the data was centered and scaled to standardize the distribution of each biomarker.



From the density plot, `prot2` looks skewed (strongly right-skewed). It contain large outliers, and the scale of `prot2`'s protein level is very high. But, after log transforming, the right-skewed `prot2` now appears more symmetric and of a much smaller and more readable scale.

Q4: We chose to find an alternate protein panel that achieves improved classification accuracy using a guided LASSO approach with differential penalties. Proteins identified through both multiple testing with t-tests and Random Forest importance (`proteins_sstar`) were assigned zero penalty to ensure these were included in the panel without significant shrinkage due to regularization. So, this prior validation was used since there was strong prior evidence that these proteins were relevant. The remaining proteins not selected by the two aforementioned methods received L1 regularization (LASSO regularization), allowing us to discover additional predictive biomarkers while controlling overfitting. We can think of this method from a Bayesian standpoint where the prior data (pre-selected panel of 5 proteins, `proteins_sstar`) informed our belief about protein importance. Overall, our method displayed increased classification accuracy, specifically highlighted by the metric: 'accuracy'. The metrics are depicted below:

metric	estimator	estimate
sensitivity	binary	0.750
specificity	binary	0.933
accuracy	binary	0.839
roc auc	binary	0.858

Compared to the benchmark results:

metric	estimator	estimate
sensitivity	binary	0.812
specificity	binary	0.733
accuracy	binary	0.774
roc auc	binary	0.883

The proteins selected for the classifier are the following:

- CD59
- 4-1BB
- Dtk
- Cadherin-5
- HAI-1
- Kallikrein 11
- PAI-1
- Growth hormone receptor
- IGFBP-4
- MRC2
- CRDL1
- IL-17 RD
- TPSG1
- MP2K2
- ENPP7
- MFGM
- PCSK7
- ITI heavy chain H4
- IgD
- DBNL
- DERM
- Elafin
- RELT
- PPID
- Semaphorin 3E
- CD27
- CNDP1

- IL-17 RC
- SRCN1
- Epo
- GDNF
- 14-3-3 protein zeta/delta
- a-Synuclein
- CSRP3
- MIG

Findings

Summarize your findings here. I've included some subheaders in a way that seems natural to me; you can structure this section however you like.

Impact of preprocessing and outliers

Task 1: After looking at the raw protein level distributions for a sample of 4 proteins in the dataset, it became clear that the scale of each protein level was very different across proteins. For example, one protein's mean was 17,164 while another's was 458. Furthermore, we found certain proteins had skewed distributions where one protein had a mean roughly 5,000 units larger than its median, indicating a right skew. Looking at each protein's density further revealed the differing scales of each protein level, large outliers, and skewed distributions apparent in 2 out of the 4 randomly selected proteins.

The log transformation of the protein levels helps to compress the scale of protein levels across the wide range of positive values that were encountered. Additionally, the logarithm helps with skewed data, helping to make data more symmetric and closer to a normal distribution (as confirmed with histograms and a QQ Plot).

Task 2:

Methodological variations

Task 3

Improved classifier

Task 4: Our alternate protein panel consisted of 35 proteins versus the 5 used in the `inclass-analysis`. We found that combining prior knowledge of protein importance with regularized classification lead to the best model accuracy. Overall our method of combining multiple t-tests, Random Forest importance, and the guided LASSO approach led to improved

results compared to the baseline displaying 8% increase in accuracy, 27% increase in specificity, $\approx 3\%$ decrease in AUC, and an 8% decrease in sensitivity. Even though AUC and sensitivity decreased slightly, the other metrics show an overall improvement in classification accuracy.

We also found that just running LASSO regularization on the whole set of proteins led to similar but slightly decreased classification accuracy compared to the baseline approach. In order to navigate around this, the incorporated prior knowledge of the 2-method statistically validated panel of proteins proved useful.