Global Gene Expression in Autism Spectrum Disorder

Background

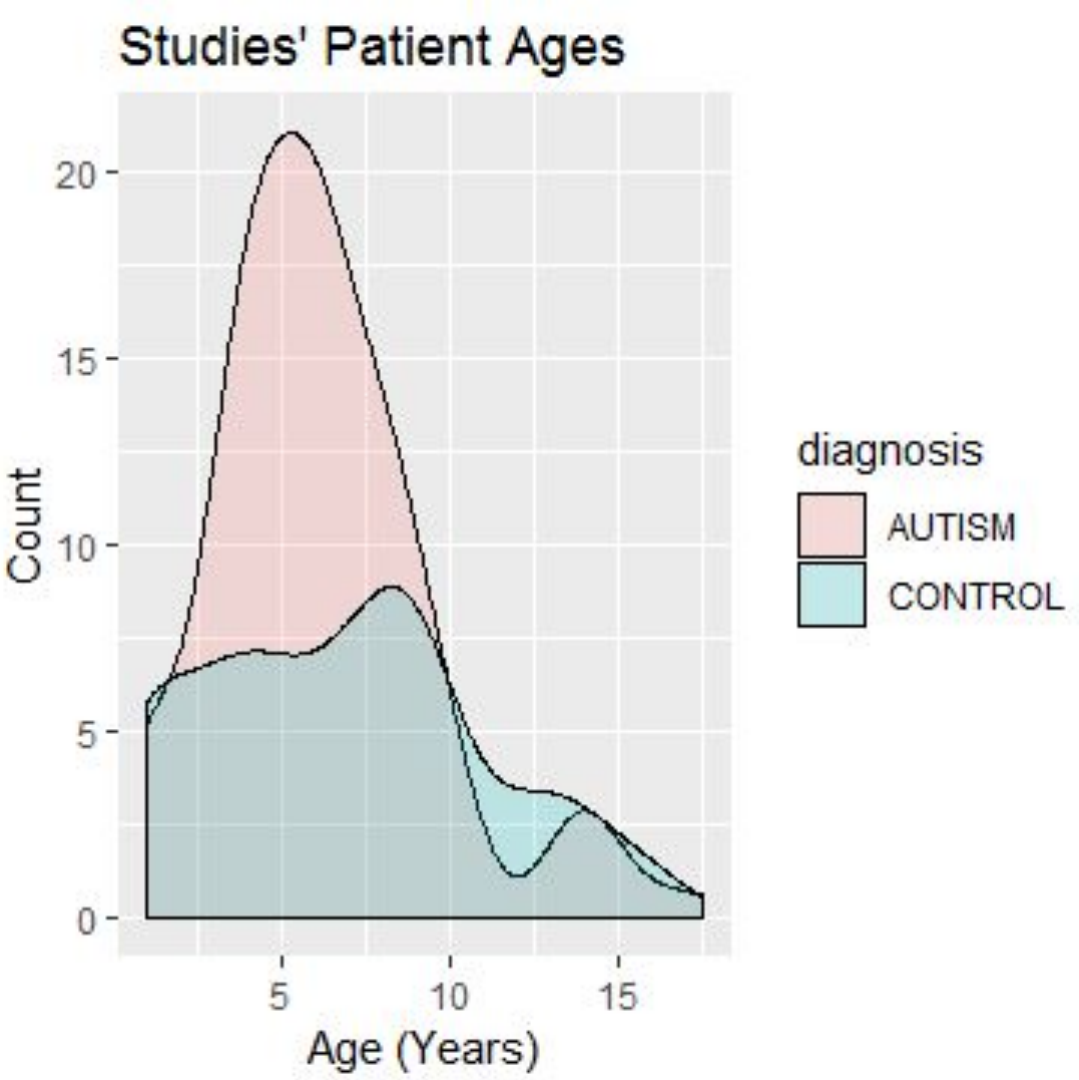
- Collection of rare variants distributed across many genes that confer the manifestation of ASD
- No secure molecular diagnostic tool for ASD and therapy targets

Question: What is the predictive power of a gene expression model combining two, independent datasets and can it successfully act as a molecular diagnostic tool for ASD in other gene expression datasets?

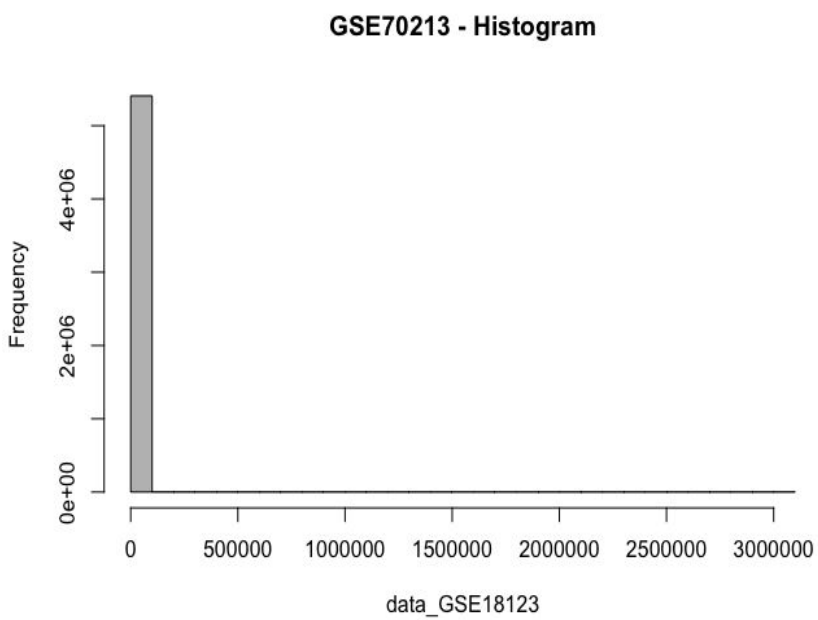
1. Evaluate differential gene expression with variables: diagnosis, age, and batch.
2. Combine datasets to validate and improve predictive model with gene expression signatures.

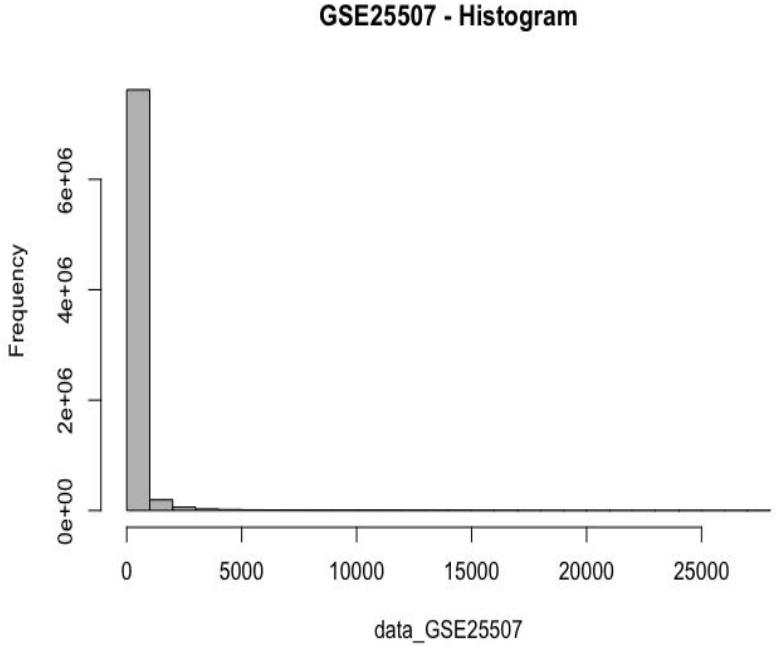
Datasets:

- Expression profiling by microarray (Affymetrix Human Genome U133 Plus 2.0 Array Platform)
- 54,613 genes, 245 samples
  1. Kong et al.(2012): entire peripheral blood
     Samples= 99
     (66 autism, 33 control)
  2. Alter et al. (2011): peripheral blood lymphocytes,
     Samples = 146
     (82 autism, 64 control)
- Ages 1-17.5 years
  (mean = 6.4 years)

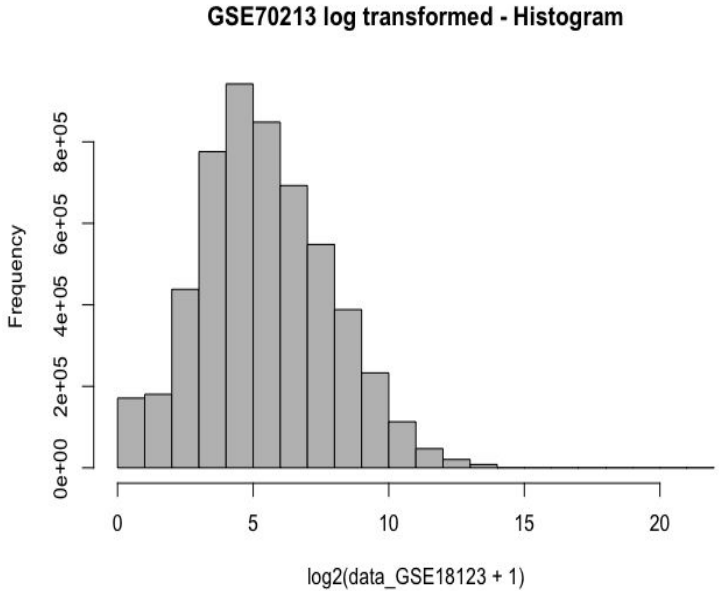


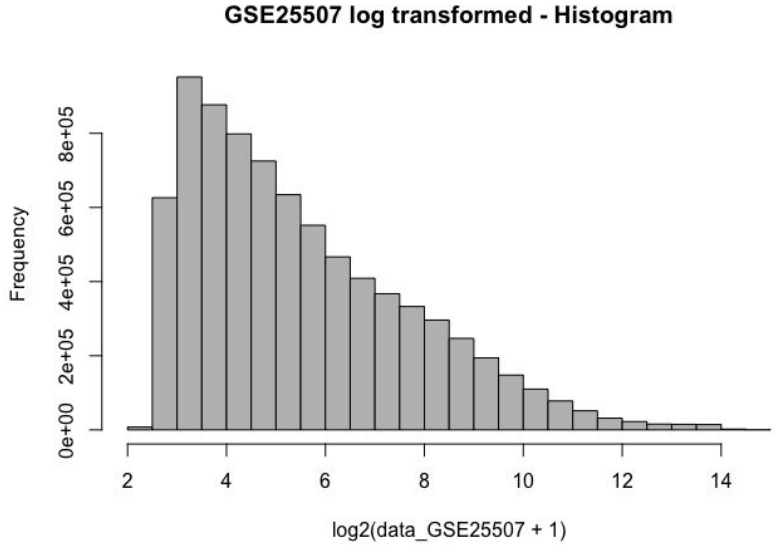
Description of data



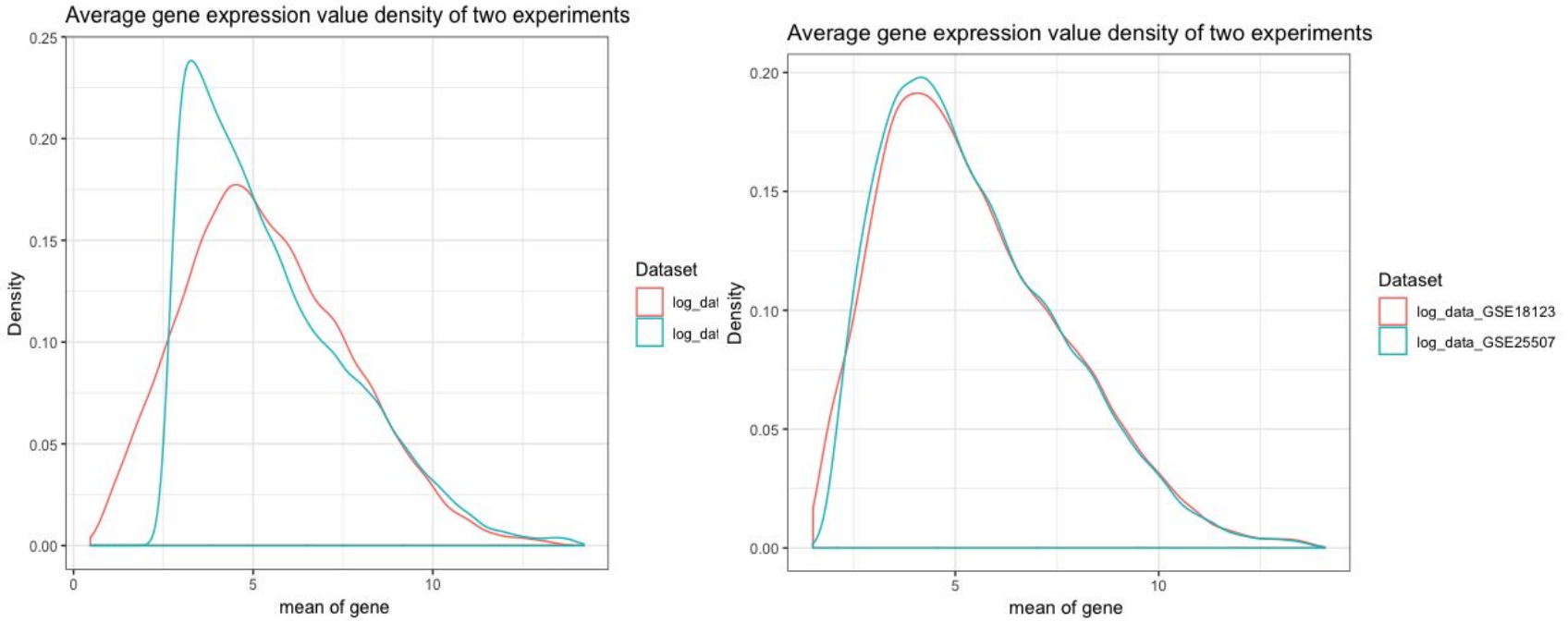


log2 transformation





Quantile normalization



Variables in Metadata

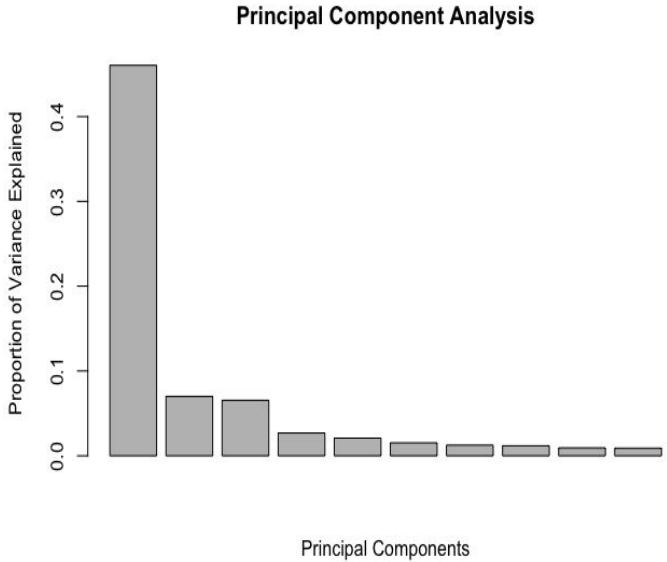
1. Diagnosis is a categorical variable with two levels(autism and control).

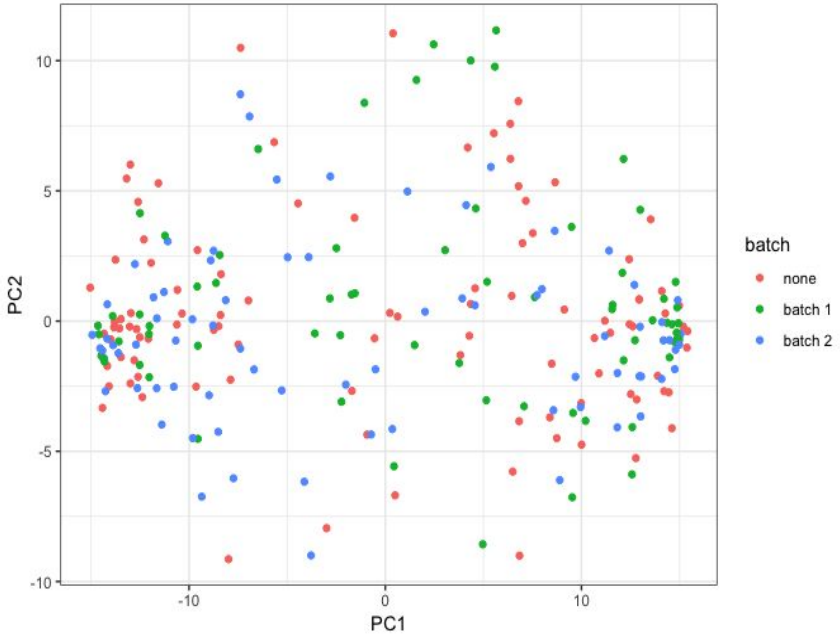
2. Batch is a categorical variable with three levels(batch 1, batch 2 and none)

3. Age is a continuous variable.
   - 15 missing values
   - Multiple imputations for missing values

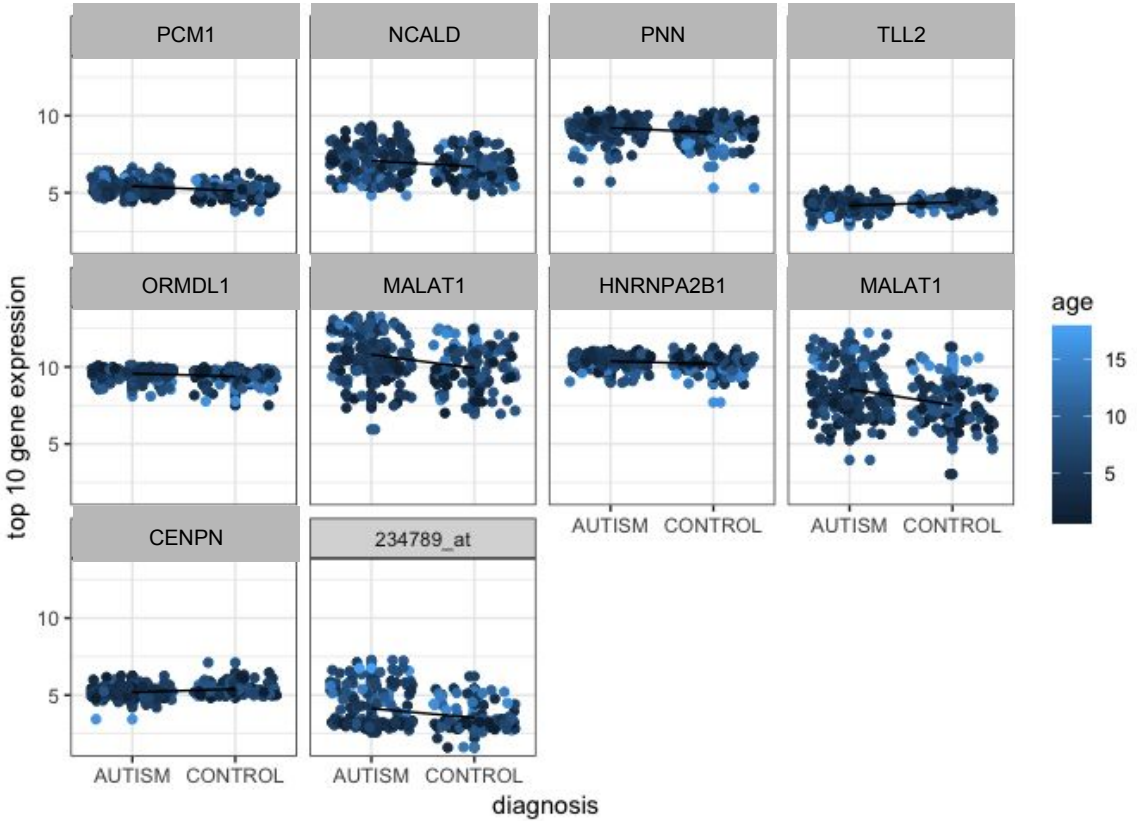
Principal Component Analysis

Goal: We want to use PCA to identify whether there is a batch effect of our combined data.





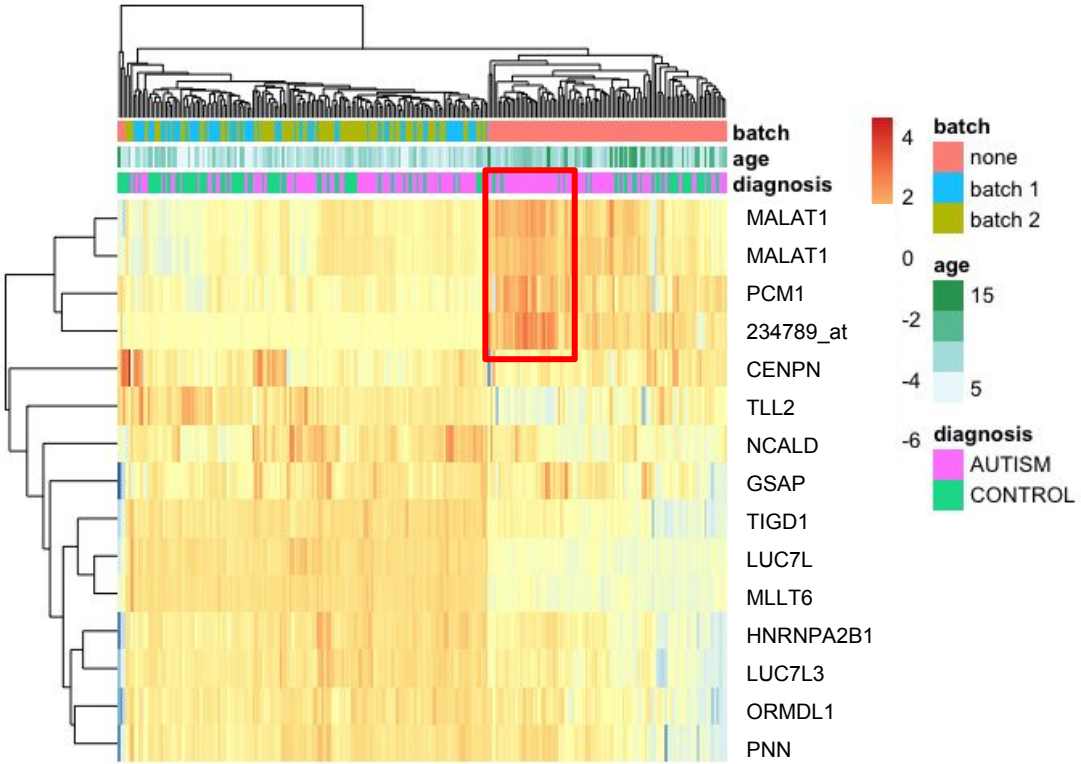
Limma analysis



Search for statistically relevant differentially expressed genes via a linear model fit

(probe 234789_at maps to an unknown gene)

Clustering heatmap of top genes



Cluster of patient samples with autism having relative increased expression of
MALAT1
PCM1
234789_at

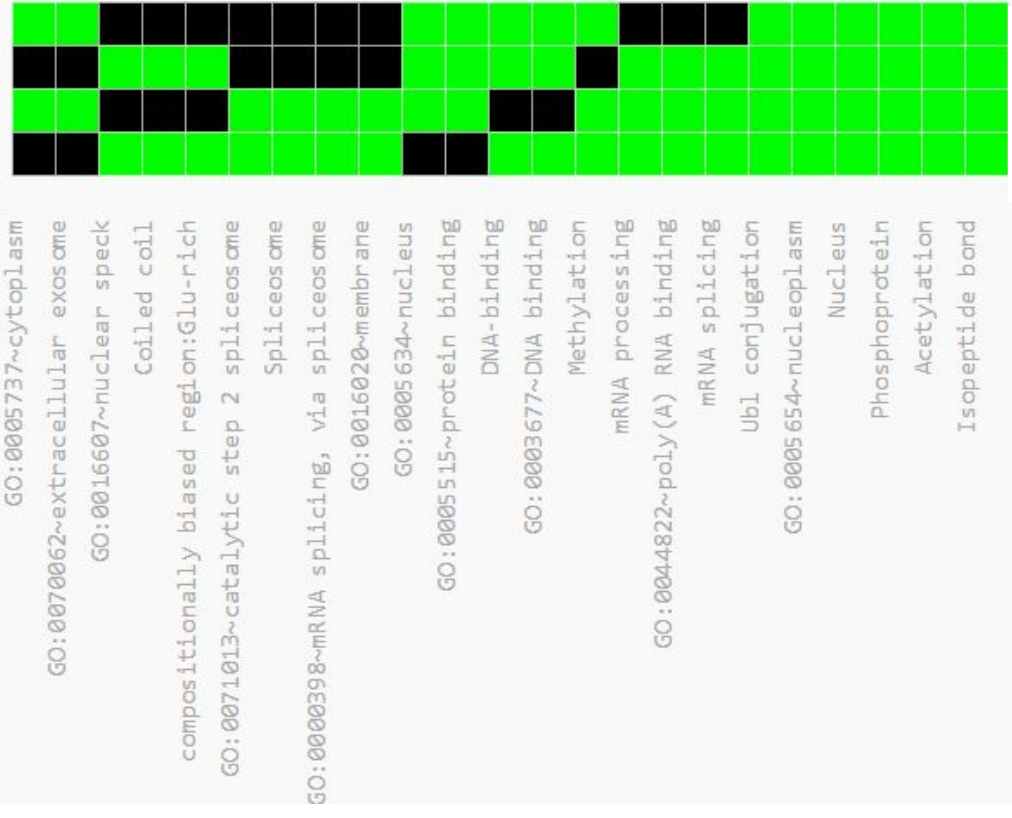
MALAT1
https://www.ncbi.nlm.nih.gov/pubmed/22960213

PCM1
https://www.ncbi.nlm.nih.gov/pubmed/26883496

Top 13 “statistically relevant” genes

PNN	pinin, desmosome associated protein
LUC7L	LUC7 like
LUC7L3	LUC7 like 3 pre-mRNA splicing factor
TIGD1	tigger transposable element derived 1
MALAT1	metastasis associated lung adenocarcinoma transcript 1
MLLT6	MLLT6, PHD finger containing
TLL2	tolloid like 2
HIST1H2BG	histone cluster 1 H2B family member g
PCM1	pericentriolar material 1
HNRNPA2B1	heterogeneous nuclear ribonucleoprotein A2/B1
ORMDL1	ORMDL sphingolipid biosynthesis regulator 1
GSAP	gamma-secretase activating protein
AC092718.4	(unknown gene transcript)

DAVID: (KEGG-pathway) gene function classification



histone cluster 1 H2B family member g(HIST1H2BG)
LUC7 like 3 pre-mRNA splicing factor (LUC7L3)
heterogeneous nuclear ribonucleoprotein A2/B1(HNRNPA2B1)
pinin, desmosome associated protein (PNN)

HIST1H2BG gene related to histone function/formation

LUC7L3, HNRNPA2B1,PNN genes associated with mRNA splicing functions in the nucleus

Machine Learning

Goal: build a binary classifier that predicts diagnosis based on the gene expression profile of a patient.

Motivation: LASSO or Elastic Net regularizer produce sparse solutions

Analysis: compare the genes selected by different models with the statistically relevant genes.

Results - accuracy

Linear SVM trained with Stochastic Gradient Descent: 75% classification accuracy using 39 genes

Logistic Regression: 84% classification accuracy using 50 genes

(Ran the training overnight, maximum 100 iterations for every model, store the model with the highest accuracy every 100 epochs. Run 4-fold cross-validation first to determine the regularization strength.)

Results - matching genes

Logistic reg. vs SGD	Logistic reg. vs Stat	SGD vs Stat	All
GATA2	AC092718.4	HISTH2BG	
CXCR3	PCM1		
STATH	ORMDL1		
MMP27			

Results

Matching genes seem uncorrelated to each other

Matching genes seem too unspecific:

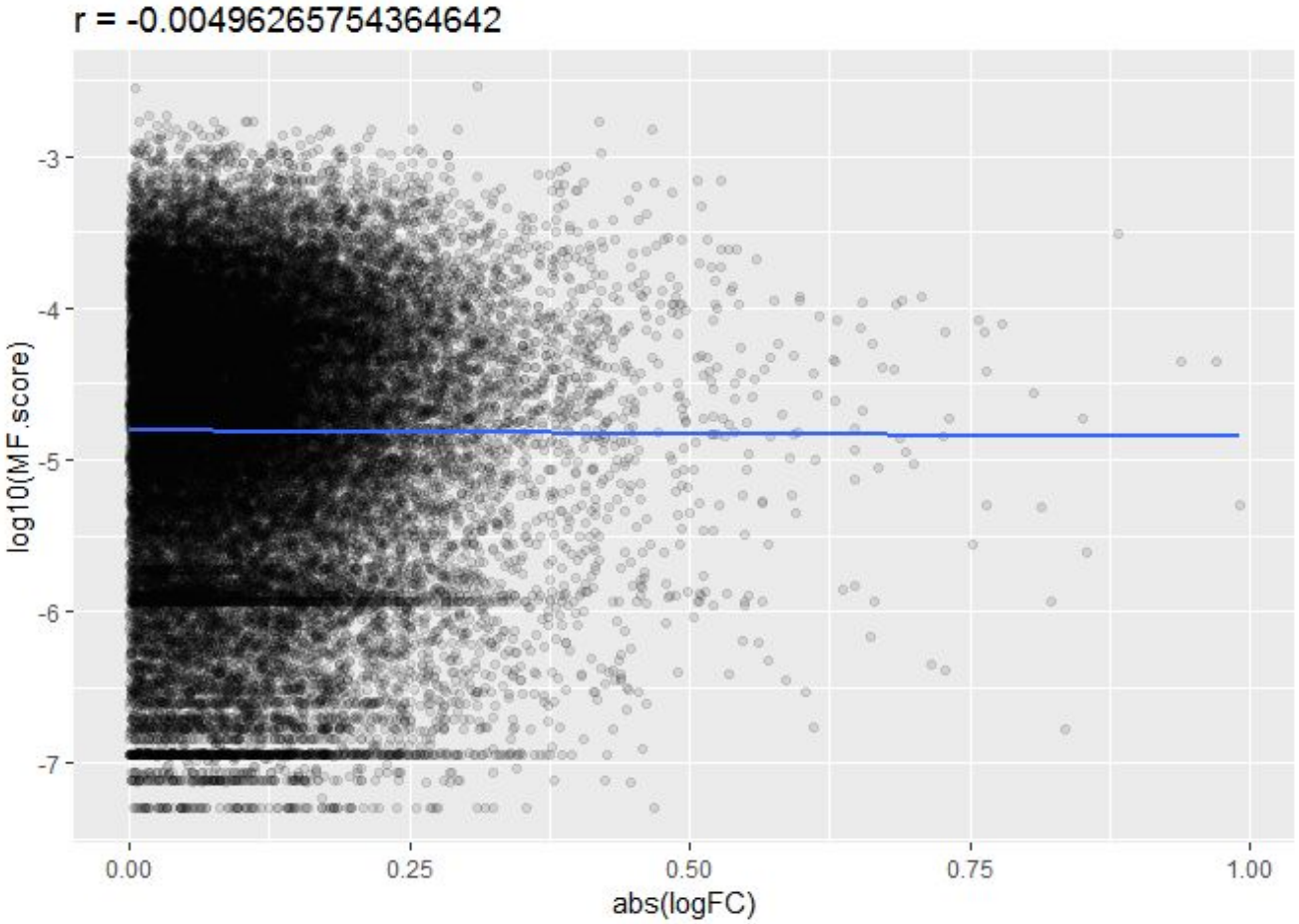
- HISTH2BG: histone-coding proteins
- GATA2: transcription factor whose mutation is associated with a wide range of diseases)

Models offer promising results and high classification accuracy - overfitting?

Multifunctional Bias

Spearman's correlation

- r = -0.00496…
- No/weak monotonic relationship between variables (MF scores & genes)
- Weak multifunctional bias



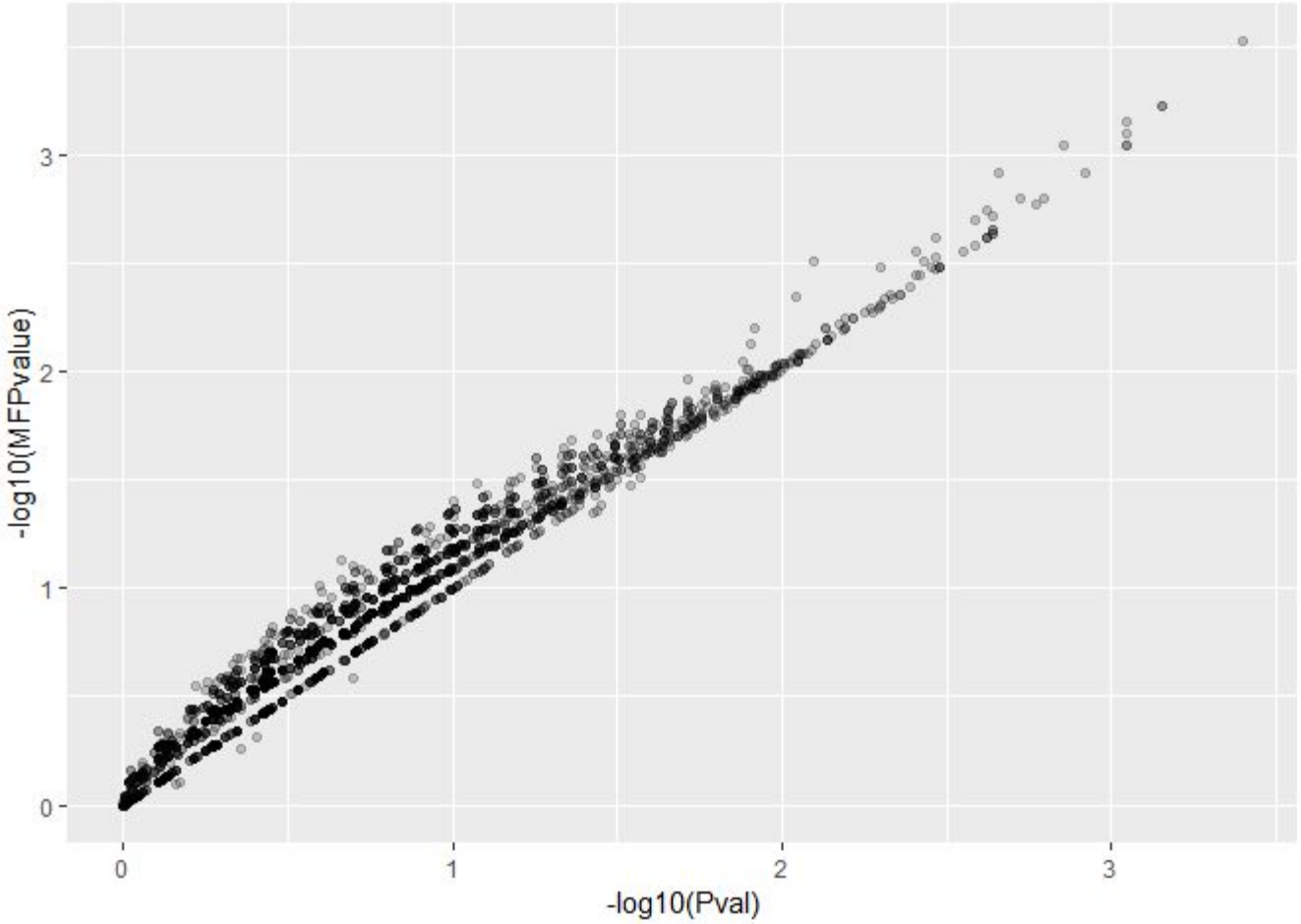
Geneset Enrichment Analysis

Precision-Recall Method

```
## # A tibble: 3,494 x 12
##    Name  ID    NumProbes NumGenes RawScore    Pval CorrectedPvalue MFPvalue
##    <chr> <chr>     <dbl>    <dbl>    <dbl>   <dbl>           <dbl>    <dbl>
##  1 erro~ GO:0~        20       20   0.0517 0.0004              1     0.000300
##  2 nucl~ GO:0~        36       36   0.0308 0.0007              1     0.000600
##  3 DNA ~ GO:0~       191      191   0.0194 0.0007              0.781 0.000600
##  4 regu~ GO:2~       102      102   0.0178 0.0009              0.502 0.0007
##  5 posi~ GO:2~        62       62   0.0225 0.0009              0.431 0.0008
##  6 thyr~ GO:0~        21       21   0.0458 0.0009              0.754 0.0009
##  7 telo~ GO:0~        22       22   0.0473 0.0009              0.603 0.0009
##  8 nucl~ GO:0~       106      106   0.0173 0.0014              0.521 0.0009
##  9 nucl~ GO:0~        24       24   0.0439 0.00120             0.502 0.00120
## 10 DNA-~ GO:0~       110      110   0.0172 0.0022              0.567 0.00120
## # ... with 3,484 more rows, and 4 more variables: CorrectedMFPvalue <dbl>,
## #   Multifunctionality <dbl>, `Same as` <chr>, GeneMembers <chr>
```

MF scores: 0.528-0.929

GO Terms Multifunctionality Adjustment



- Largest adjustment = 0.471
- No large losses to statistical significance of GO terms with multifunctionality adjustment

GO Terms

ID	Term
GO:2000573	positive regulation of DNA biosynthetic process
GO:0006296	nucleotide-excision repair, DNA incision, 5'-to lesion
GO:0006260	DNA replication
GO:0042276	error-prone translesion synthesis
GO:2000278	regulation of DNA biosynthetic process
GO:0030878	thyroid gland development
GO:0006297	nucleotide-excision repair, DNA gap filling
GO:0006261	DNA-dependent DNA replication
GO:0006289	nucleotide-excision repair
GO:0000723	telomere maintenance

- DNA regulation
- No ML genes in enriched GO terms

GATA2 : chromosome 3 → loss GATAD2B (gene family)*

CXCR3: chromosome x (chemokine)--> upregulated signaling in ASD (https://patentimages.storage.googleapis.com/88/a1/6a/8397de58196fa9/US20070048801A1.pdf)

STATH (statherin): chromosome 4 → peptide with reduced phosphate group (than control) (https://www.spectrumnews.org/news/search-for-autism-biomarkers-turns-to-saliva/)

MMP27: chromosome 11 → gain *

- * https://www.malacards.org/card/autism?limit[RelatedGenes]=158&limit[CnvdVariations]=2458

Question: What is the predictive power of a gene expression model combining two, independent datasets and can it successfully act as a molecular diagnostic tool for ASD in other gene expression datasets?

There is promise in applying machine learning to develop a molecular diagnostic tool to test for ASD via a gene expression profile.

Limitations:

- Incomplete data sets (unmatched gene IDS)
- Cell heterogeneity → couldn't control proportional difference in peripheral blood lymphocytes vs. all peripheral blood cells
  - Method used only in single-cell studies and epigenome-wide association studies.
- Relative low fold change of differentially expressed genes, questionable statistical significance