Logistic and Lasso regression models in predicting clinical outcomes with significant analysis

Presented by Chongshu Chen
BST550

Department of Biostatistics and Computational Biology
University of Rochester

Dec 15th, 2015

Table of Contents

1. Introduction
   - Objectives
   - Gene Expression Omnibus (GEO)
2. Statistical methods
   - Logistic Regression Model
   - Lasso Regression Model
   - DESeq2 Negative Binomial Distribution
3. Summary

Background

- The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes
- Genes vary in size from a few hundred DNA bases to more than 2 million bases in humans
- Large scale RNA-seq experiments are become increasingly routine

The Path to Genomic Medicine

Human Genome Project (2003)

Sequence More Genomes

Interpret Genome Data

Identify Functions

Realization of Genomic Medicine (20XX?)

Scientific Objective

- Study the genes that are significantly differential expressed in RNA seq data that can be used to predict clinical outcomes
- Construct a set of biomarkers that will predict breast receptor status using gene expression measurement
- Extend the approach to other clinical outcomes such as age and other biological factors.

Statistical Objective

- Explore the data mining techniques, especially classification methods to be applied to Genomic Data
- Use multiple testing procedure with FDR control and apply GLM models methods to fit Genomic Data
- Apply the logistic regression and Lasso models to predict the clinical response variables, e.g breast cancer receptor status
- Implement the cross-validation procedure/nested cross validation procedure to access the performance of predictive models

Database Link: Gene Expression Omnibus GSE62944)

NCBI

GEO
Gene Expression Omnibus

HOME | SEARCH | SITE MAP | GEO Publications | FAQ | MIAME | Email GEO

NCBI > GEO > **Accession Display**

Not logged in | Login

Scope: Self | Format: HTML | Amount: Quick | GEO accession: GSE62944 | GO

Series GSE62944 — Query DataSets for GSE62944

Status	Public on Jan 27, 2015
Title	Alternatively processed and compiled RNA-Sequencing and clinical data for thousands of samples from The Cancer Genome Atlas
Organism	Homo sapiens
Experiment type	Expression profiling by high throughput sequencing
Summary	We reprocessed RNA-Seq data for 9264 tumor samples and 741 normal samples across 24 cancer types from The Cancer Genome Atlas with "Rsubread". Rsubread is an open source R package that has shown high concordance with other existing methods of alignment and summarization, but is simple to use and takes significantly less time to process data. Additionally, we provide clinical variables publicly available as of May 20, 2015 for the tumor samples where the TCGA ids are matched.
Overall design	Using Rsubread version 1.14.2 R package, we aligned the fastq files downloaded from Cghub. First, the reads were aligned with align() function to the UCSC hg19 reference genome. Second, featureCounts() function was used to summarize the gene level expression values as integer number. Third, these

Description of GSE62944 Series

- RNA-Seq data for 9264 tumor samples and 741 normal samples across 24 cancer types from The Cancer Genome Atlas with "Rsubread".
- Note that Rsubread is an open source R package that use and takes significantly less time to process data.
- 548 clinical variables for each sample are provided in the TCGA Clinica Variables samples via txt file

GSE62944 Data

- All 9264 tumor samples have been combined to create the processed matrix files for tumor samples
- All 741 normal samples have been combined to create the processed matrix files for normal samples
- The CancerType Samples.txt and TCGA24 Normal CancerType Samples.txt files list each sample tumor type for tumor samples and normal samples respectively
- 548 clinical variables for each sample are provided in the Clinical Variables 9264 Samples.txt
- Raw data mRNA sequence can be downloaded from CGHub (https://cghub.ucsc.edu/) with an access key and processed with pipeline available from github link

A classification problem in data mining

An objective: identifying biomarkers (genes) that are significant in predicting breast cancer estrogen receptor status

A technical issue: $p > n$

Y	genes	Models
Receptor Status	??	??

Proposed statistical methods

- A Logistic regression model with significant analysis of RNA-seq experiments
- A Lasso model: a regression shrinkage and selection method
- Implement CV/nested CV procedures to access the "TRUE" performance of models (Generalization)

Logistic regression model

Model Bernoulli outcome and select the most m differential gene expression levels as predictors. We will have the linear predictor term

$$\eta = \beta_0 + \beta_1 x_1 + ... + \beta_5 x_5 \tag{1}$$

Suppose Y = 1 if receptor status = positive,

$$\phi(t) = \frac{e^{\eta}}{1+e^{\eta}} = \frac{1}{e^{-\eta}+1}$$

According to the logistic regression model, we assume that the probability of recurrence give gene expression level x is

$$Pr(Y = 1 \mid x) = \phi(\eta).$$

	<=.005	<=.05	<=.5
1	0.60	0.70	0.88

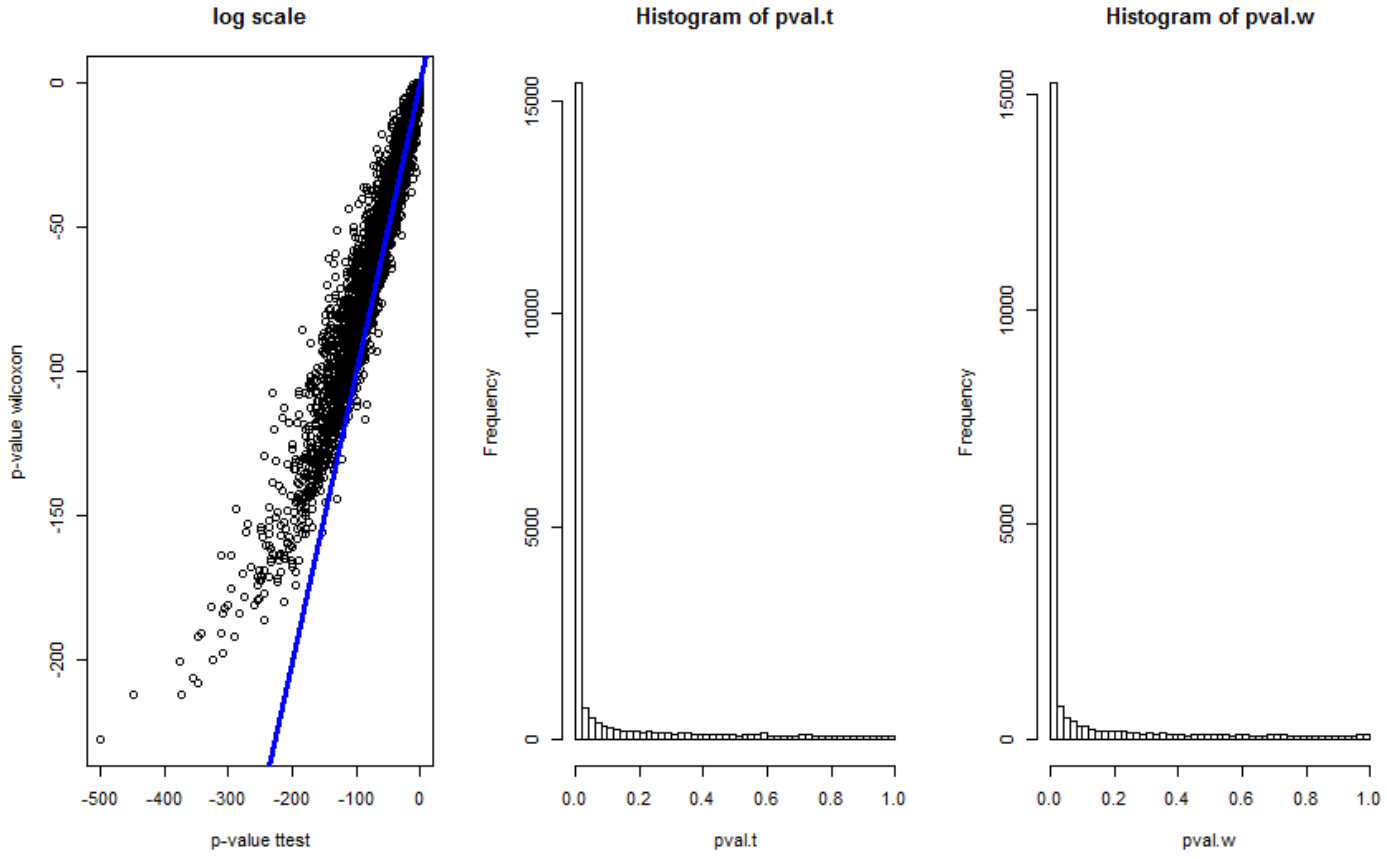
Table: t test

	<=.005	<=.05	<=.5
1	0.60	0.69	0.88

Table: wilcoxcon test

Tables show the proportion of genes has adjusted p-value less than given thresholds under two sample t tests and two sample Wilcoxon rank sum tests.

Results from T test and Wilcoxon test



Wilcoxon and t tests results are varied slightly in the histogram.

Cross validation for logistic regression models

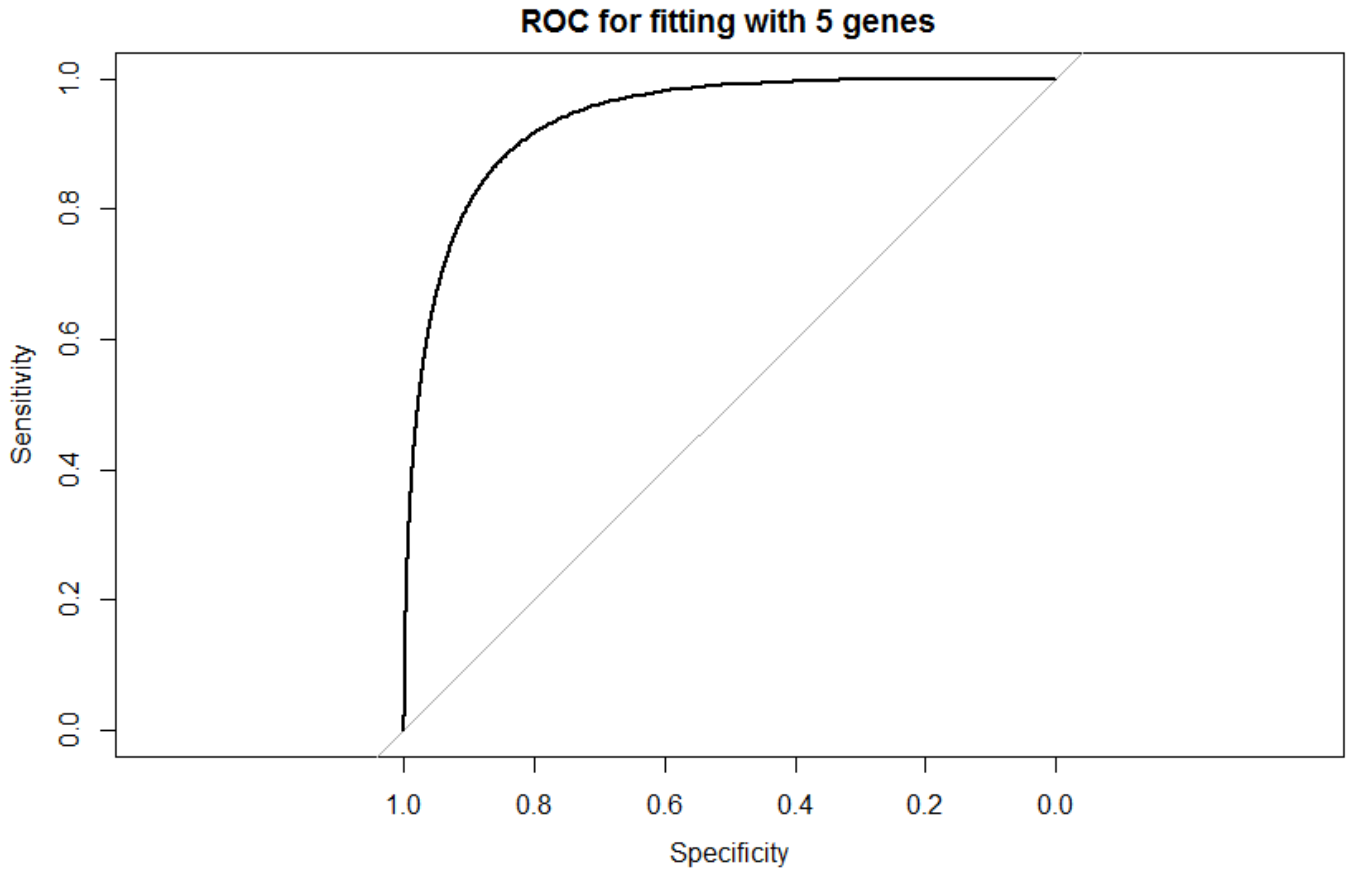
- Random assign rows with sampling ID and select training data and test data
- Select the m genes according to wilcoxon rank sum test with smallest p-values from the training data only
- Evaluate the risk score, the linear predictor term η for the test data
- Estimate the predictive outcomes under the model for all test set and repeat above steps
- Use the AUC (area under curve) to evaluate the model prediction errors

Logistic Model with 5 Covariates

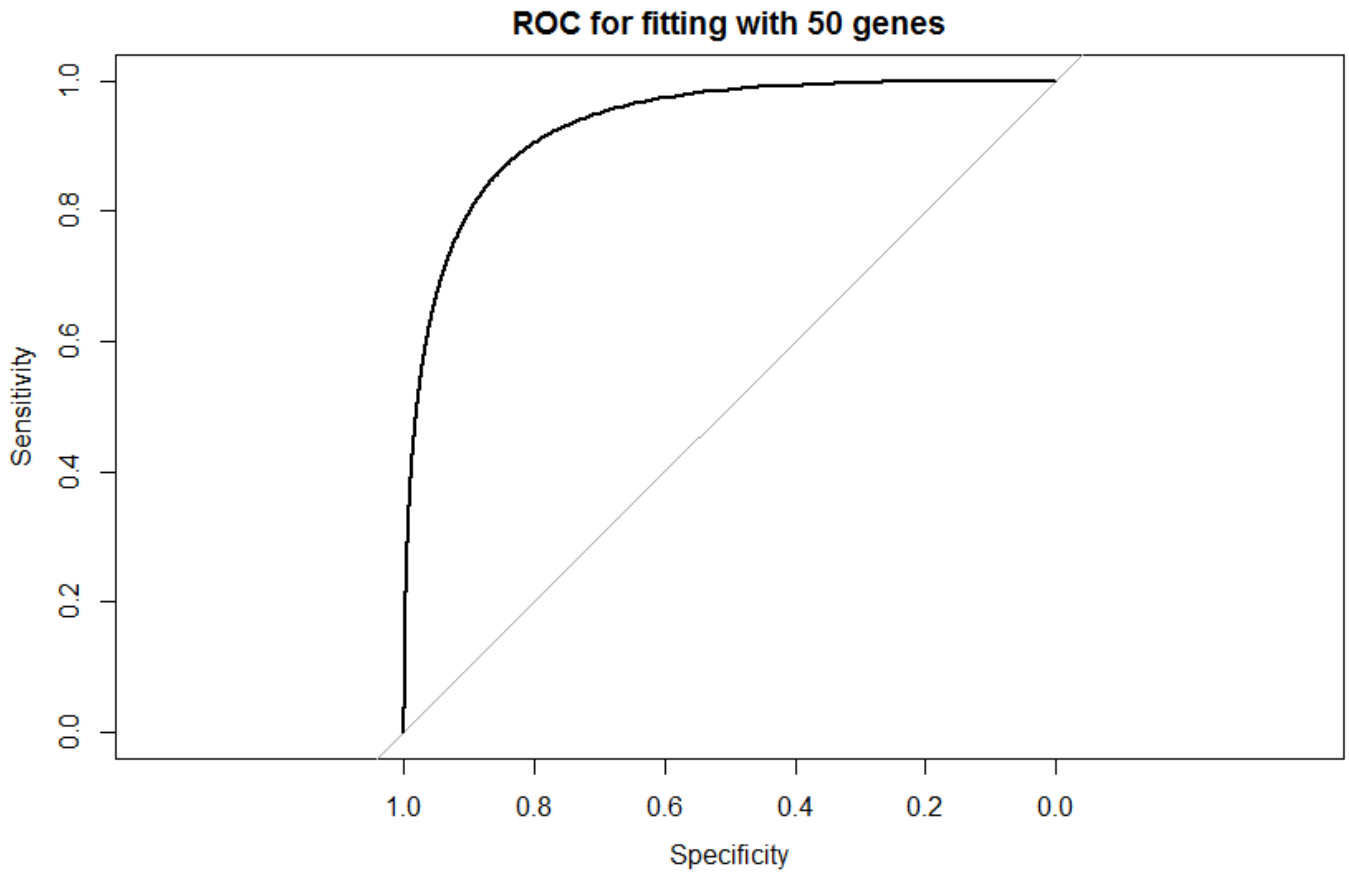
```
summary(fitm1)

##
## Call:
## glm(formula = status ~ ., family = binomial(link = logit), data = fitteddata)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.2367   0.0864   0.1645   0.2511   3.2020
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -9.13911    1.08458  -8.426  < 2e-16 ***
## ESR1         0.48883    0.09672   5.054 4.32e-07 ***
## GATA3        0.13486    0.11201   1.204 0.228579
## AGR3         0.19344    0.05462   3.542 0.000398 ***
## GPR77        0.33324    0.12212   2.729 0.006358 **
## C6orf97     -0.15680    0.13442  -1.166 0.243432
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1086.33  on 1014  degrees of freedom
## Residual deviance:  378.07  on 1009  degrees of freedom
```

Logistic Models with 5 Genes VS. 50 Genes

AUC statistic is used to evaluate model performance



Measuring Predictability with ROC



A brief explanation of Lasso model

Give a set of input measurements $x_1, x_2, ..., x_k$ and an outcome measurement y, the lasso fits a linear model

$$Y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k \tag{2}$$

subject to minimize $\sum(y_i - \beta^T x_i)^2 + \lambda \sum \mid \beta_j \mid$

Nested Cross Validation at Optimal λs

Implementation: ncv.lasso (gex, k1, k2, m)

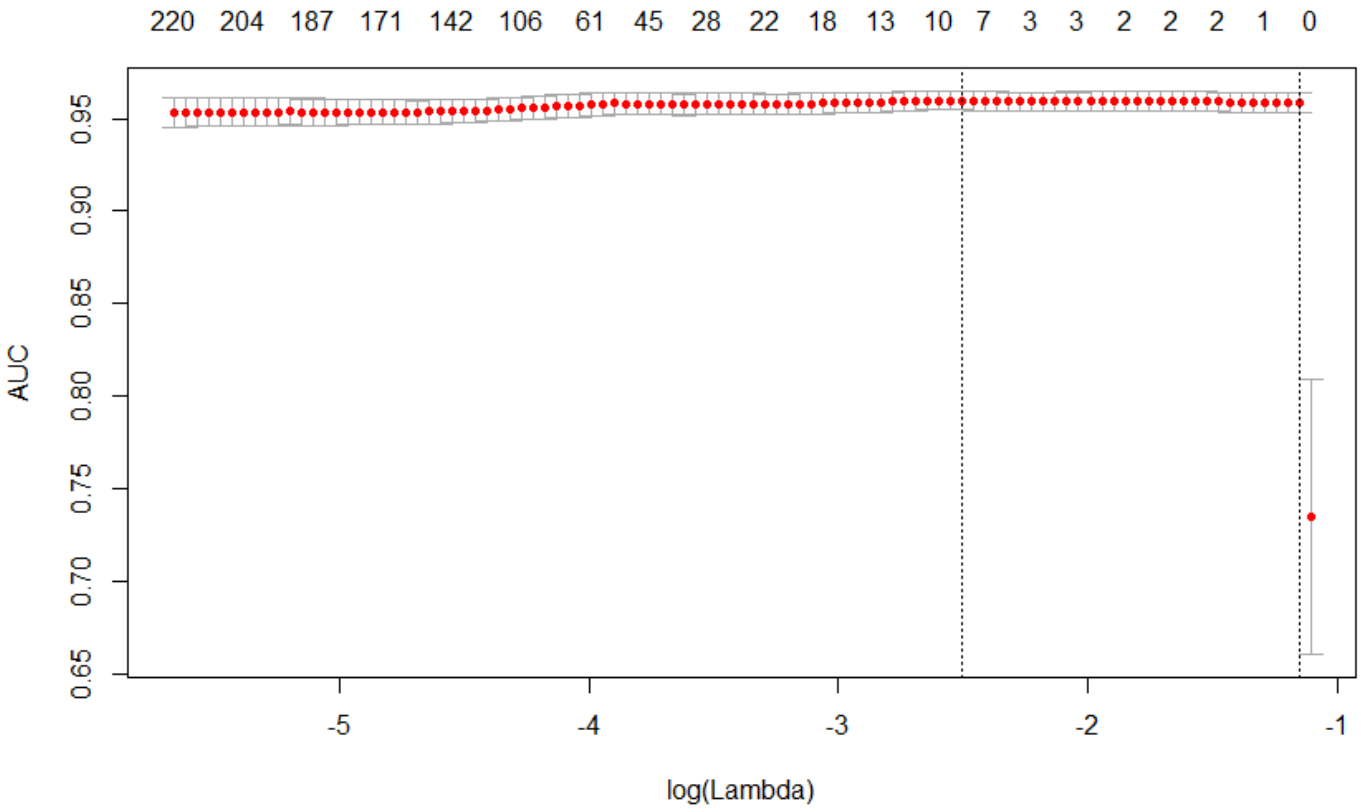
- Extend the cross validation procedure with nested cross-validation approach to evaluate robustness of the Lasso model in predicting receptor status
- The nested cross valuation method not only selects the optimal lambda, but also evaluates the accuracy of prediction
- Include the multiple hypothesis testing procedure for pre-selection in function

Nested Cross Validation Method

- Implement a cross validation procedure in an inner loop to choose the optimal λ value
- Make an outer loop function to fit the LASSO model at given optimal λ and validate the error rate with cross validation procedure.
- Set up a strong penalty for fitting Lasso models
- Expect the AUC statistic would be lower than previous cross validation procedures

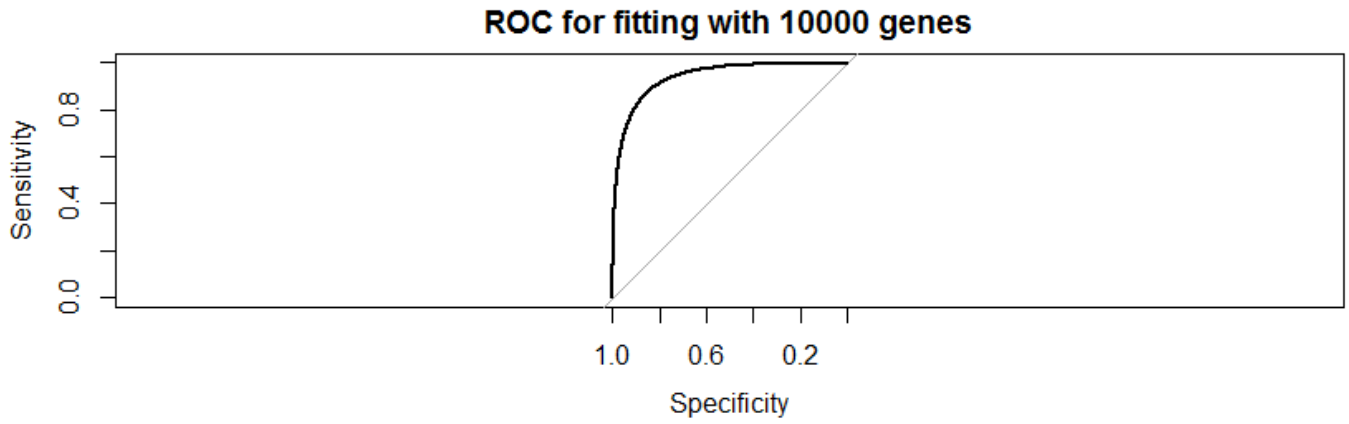
LASSO Model with 9 Covariates

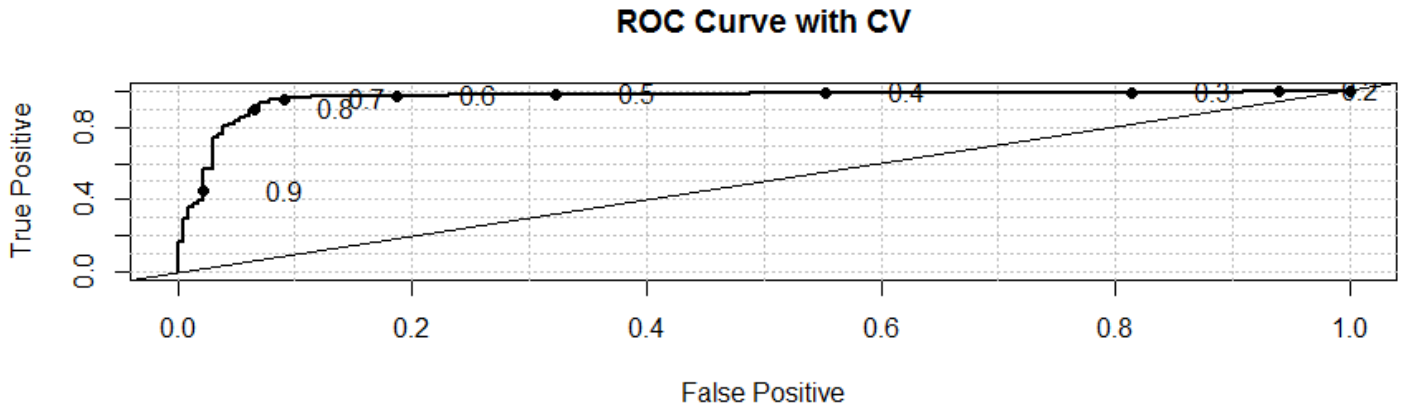
X1.2.SBSRNA4 A2MP1 AGRN B3GNT6 CA13 DEK DNASE1 ESR2 GPR78



Lasso with Nested CV with Gene Selection

ncv.lasso <- function (gex, k1, k2, m)





Lasso with Nested CV with Gene Selection

ncv.lasso <- function (gex, k1=5, k2=10, m=10000)

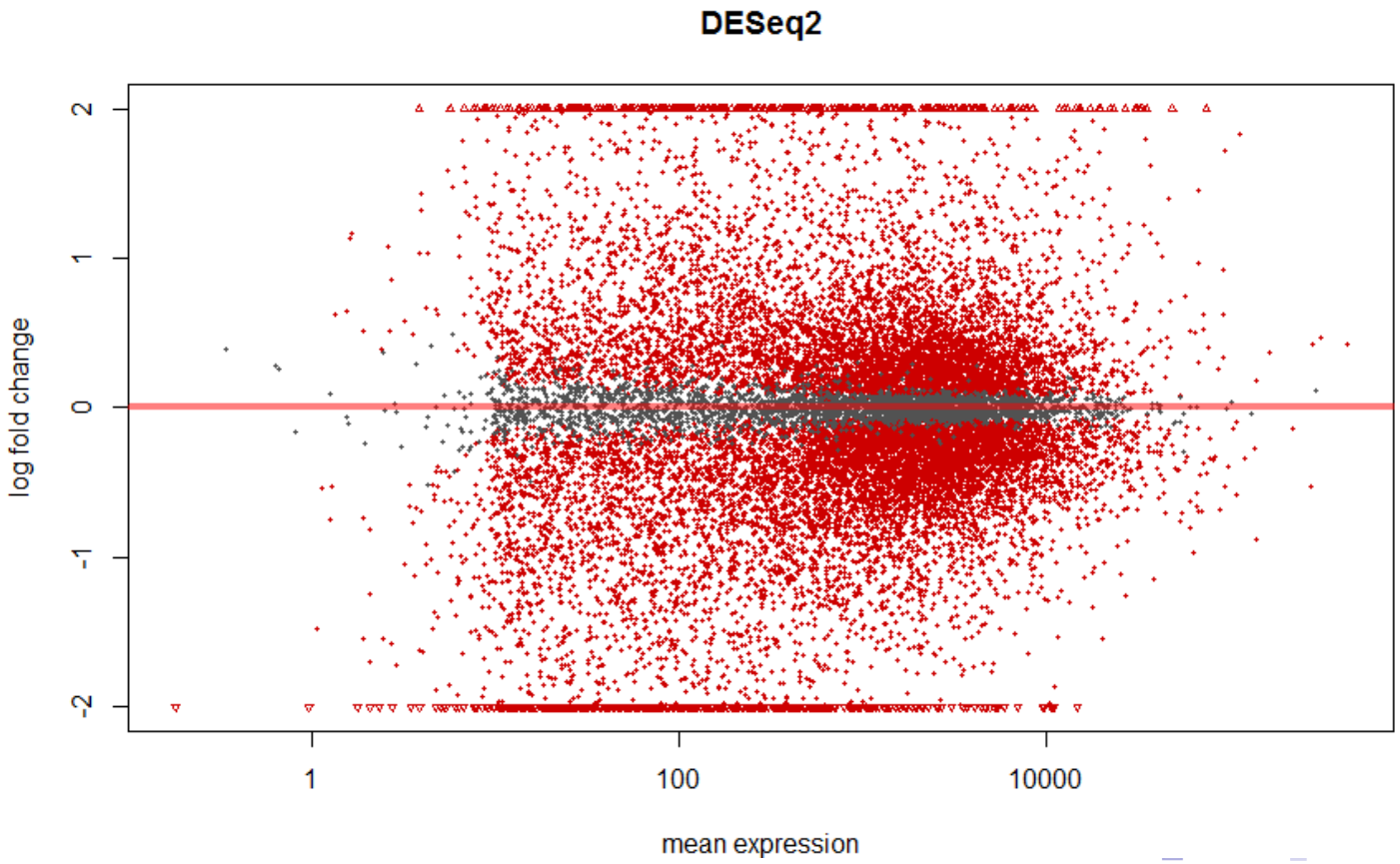
```
 the coeficient of beta not equal 0
X1.2.SBSRNA4         A1BG     A1BG.AS1        A2ML1
  0.85374442   0.02638241   0.00000000   0.00000000

 the coeficient of beta not equal 0
X1.2.SBSRNA4         A1BG     A1BG.AS1        A2LD1
  0.86411004   0.02560542   0.00000000   0.00000000

 the coeficient of beta not equal 0
X1.2.SBSRNA4         A1BG         A1CF        A2ML1      AADACL3
  0.90510576   0.02551822   0.00000000   0.00000000   0.00000000

 the coeficient of beta not equal 0
X1.2.SBSRNA4         A1BG         A1CF
  0.89199197   0.02598418   0.00000000

 the coeficient of beta not equal 0
X1.2.SBSRNA4         A1BG         A1CF        A2MP1         AA06         AAAS        AADAT
  0.91884530   0.02616571   0.00000000   0.00000000   0.00000000   0.00000000   0.00000000
       ABCC3
```

Deseq2 Results

```
> res
log2 fold change (MAP): typeInd Positive vs negative
Wald test p-value: typeInd Positive vs negative
DataFrame with 18167 rows and 6 columns
           baseMean log2FoldChange      lfcSE          stat        pvalue          padj
          <numeric>      <numeric>  <numeric>     <numeric>     <numeric>     <numeric>
ESR1     30883.2794       4.131422 0.12072807      34.22089 1.182693e-256 2.148599e-252
C6orf97   1519.4681       3.086806 0.09404742      32.82181 2.876892e-236 2.613224e-232
CPB1     49227.3899       8.144686 0.24970357      32.61742 2.322550e-233 1.406459e-229
GPR77      418.8725       2.834369 0.08860434      31.98905 1.548358e-224 7.032253e-221
COL9A1     117.3629      -5.136620 0.16329543     -31.45599 3.475798e-217 1.262897e-213
...             ...            ...        ...           ...           ...           ...
RING1    2606.87546  -5.057235e-05 0.04179816 -0.0012099180     0.9990346     0.9992546
CC2D1A   3147.60048  -5.584035e-06 0.04475611 -0.0001247659     0.9999005     0.9999093
HIST1H1E   30.89528   2.849029e-05 0.10325681  0.0002759169     0.9997799     0.9999093
NENF     5099.19726   7.164160e-06 0.06303084  0.0001136612     0.9999093     0.9999093
RSPRY1   1433.47734  -8.667644e-06 0.05087448 -0.0001703731     0.9998641     0.9999093
```

MA plot: an application for visual representation

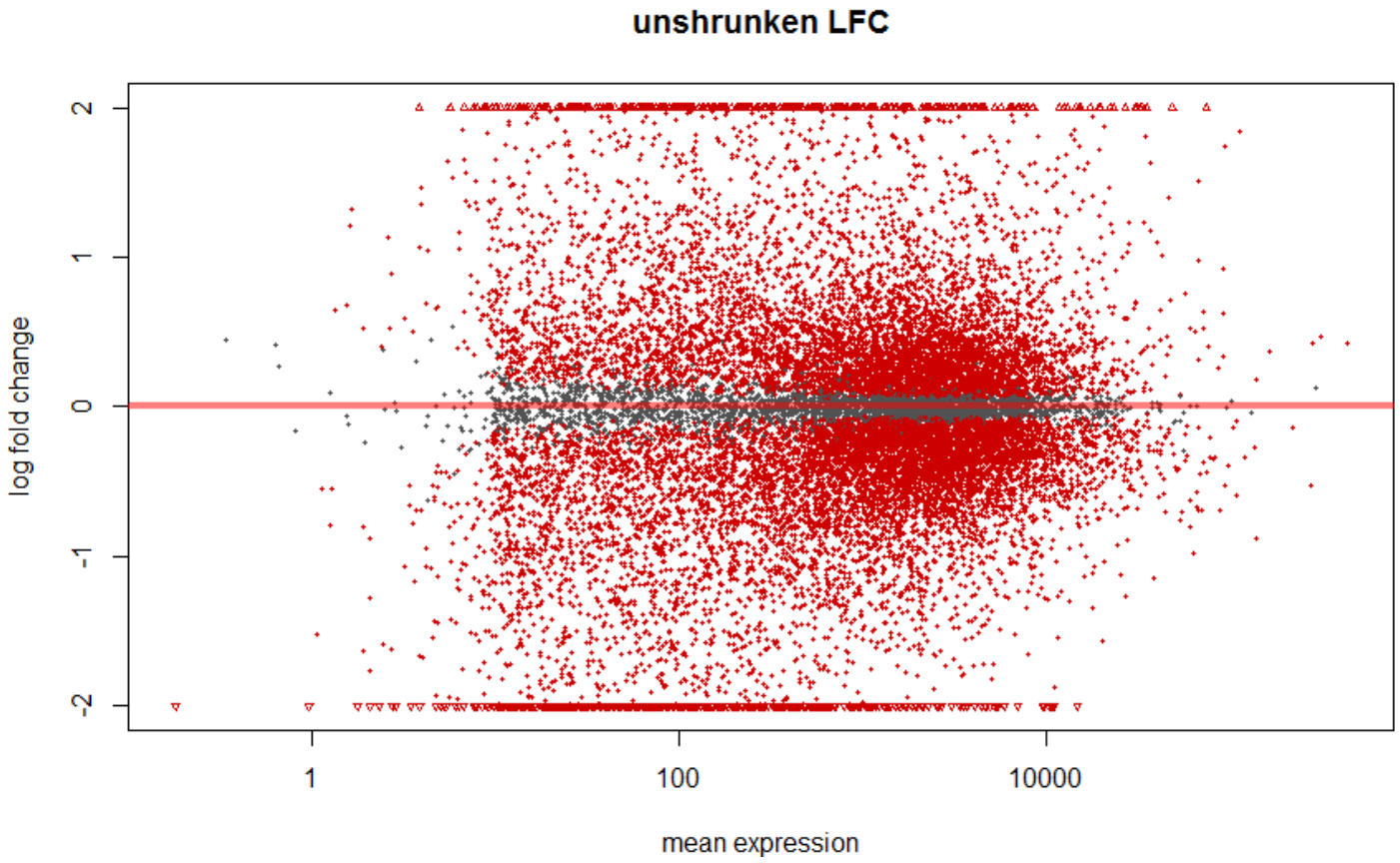
Two channel DNA gene expression data has been transformed onto the M (log ratios) and A (mean average) scale



MLE estimates

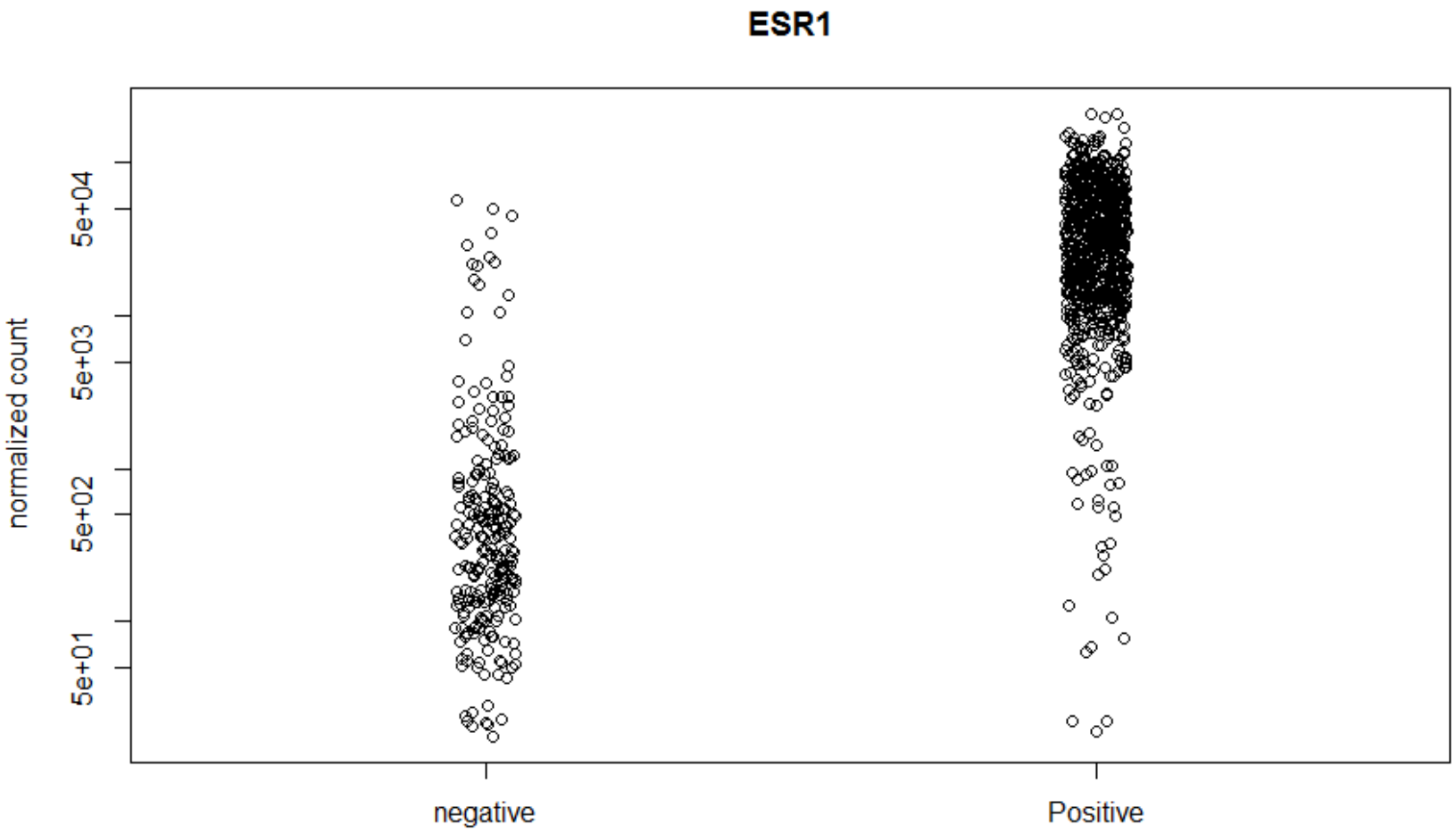
```
log2 fold change (MAP): typeInd Positive vs negative
Wald test p-value: typeInd Positive vs negative
DataFrame with 18167 rows and 7 columns
           baseMean log2FoldChange        lfcMLE      lfcSE          stat        pvalue
          <numeric>      <numeric>     <numeric>  <numeric>     <numeric>     <numeric>
ESR1     30883.2794       4.131422      4.174239 0.12072807      34.22089 1.182693e-256 2.14
C6orf97   1519.4681       3.086806      3.105975 0.09404742      32.82181 2.876892e-236 2.61
CPB1     49227.3899       8.144686      8.562797 0.24970357      32.61742 2.322550e-233 1.40
GPR77      418.8725       2.834369      2.849965 0.08860434      31.98905 1.548358e-224 7.03
COL9A1     117.3629      -5.136620     -5.233245 0.16329543     -31.45599 3.475798e-217 1.26
...             ...            ...           ...        ...           ...           ...
RING1    2606.87546  -5.057235e-05 -5.062315e-05 0.04179816 -0.0012099180     0.9990346
CC2D1A   3147.60048  -5.584035e-06 -5.580724e-06 0.04475611 -0.0001247659     0.9999005
HIST1H1E   30.89528   2.849029e-05  2.925217e-05 0.10325681  0.0002759169     0.9997799
NENF     5099.19726   7.164160e-06  7.175659e-06 0.06303084  0.0001136612     0.9999093
RSPRY1   1433.47734  -8.667644e-06 -8.664454e-06 0.05087448 -0.0001703731     0.9998641
```

MA plot with MLE estimates



Plot Count function

A plot of counts for ESR1 with min adjust p-value



Results and Comments

- Investigate on using logistic predictive models with significant analysis, Lasso penalized regression, Deseq2 method
- Logit models shows that ESR1 is the most significant gene for receptor status
- Deseq2 method also model ESR1 that it has the smallest p-value on receptor status
- Both Lasso models and logit models shows very high AUC values and great prediction performance

Summary and Future Work

- Identify the genes that are significantly predicting the breast cancer estrogen receptor status with three modeling techniques
- Evaluate the performance the logit and Lasso models with cross-validation procedures
- Implement the nested cross-validation procedure to evaluate the lasso model with gene pre selection procedure
- Extend these methods to typical types of comparisons and sampling schemes in RNA-seq data for clinical outcomes
- Apply these methods to other biological factors or clinical outcomes in the TCGA data set

References

[1] Rahman M, Jackson LK, Johnson WE, Li DY et al. (2015). Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics* 31(22):3666-72. PMID: 26209429'

[2] NCBI database http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE62944

[3] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.