Analysis Pipeline of Prostate Cancer Data

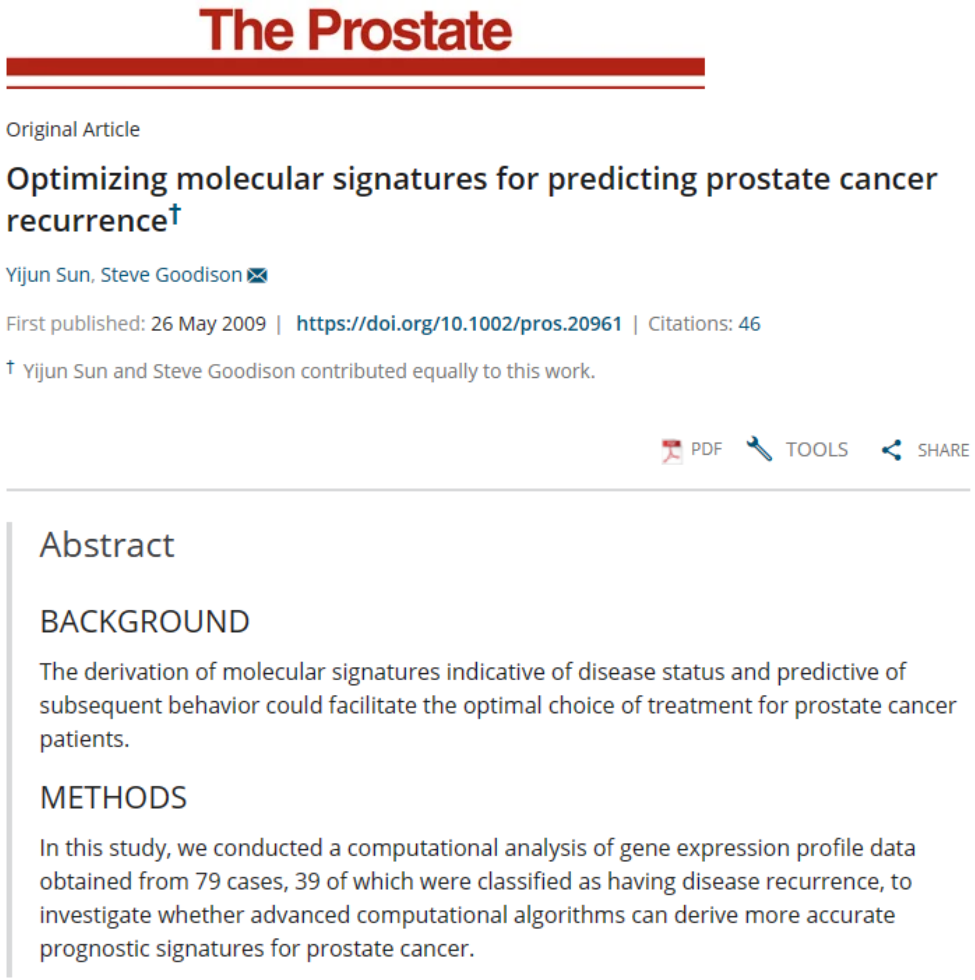
Gene Expression Data Analysis and Visualization
410.671.81
Summer 2021
Final Project
By: Zach Young

Introduction

- For this project I will be performing an analysis pipeline on a gene expression dataset from prostate tumor samples
- The goal of this project will be to clean the data, find and filter out significant genes and these will then be used to classify samples into two groups
- The pipeline for this project will follow this outline:
  - Normalize Data
  - Outlier Detection
    - Clustering dendrogram, CV vs. Mean plot and Average Correlation
  - Gene Filtering
    - Filtered low expressed genes using mean values
  - Feature Selection
    - Non-parametric and permutation methods were examined
  - Dimensionality Reduction using PCA
  - Classification with Linear Discriminant Analysis

Background

- The data for this analysis comes from a 2009 study that was concerned with using gene expression profiles to better predict prostate cancer prognosis in order to facilitate more optimal treatment choice for cancer patients[5]
- In the study, the researchers were able to successfully develop a computational algorithm based off specific genetic signatures that outperformed previous clinical standards in terms of predicting cancer recurrence[5]
- The data consists of a total of 79 tumor samples (40 non-recurrent and 39 recurrent) and was generated using the Affymetrix Human Genome U133A Array platform



The Prostate

Original Article

Optimizing molecular signatures for predicting prostate cancer recurrence†

Yijun Sun, Steve Goodison ✉

First published: 26 May 2009 | https://doi.org/10.1002/pros.20961 | Citations: 46

† Yijun Sun and Steve Goodison contributed equally to this work.

Abstract

BACKGROUND

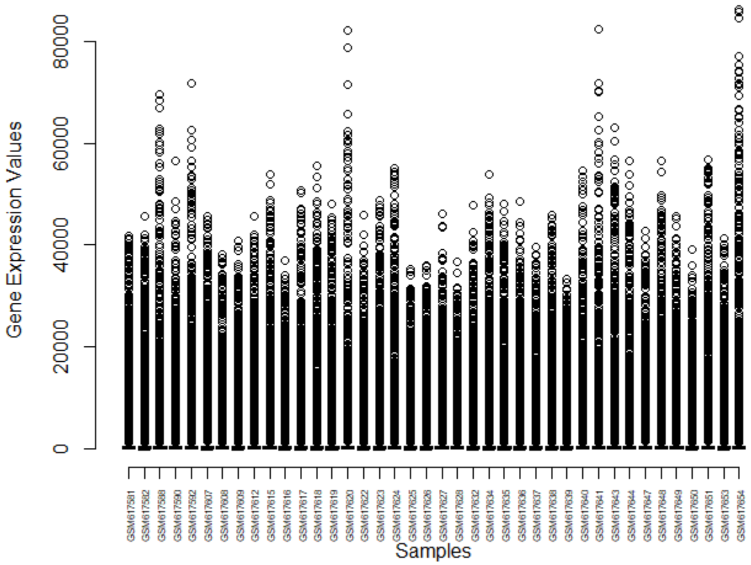
The derivation of molecular signatures indicative of disease status and predictive of subsequent behavior could facilitate the optimal choice of treatment for prostate cancer patients.

METHODS

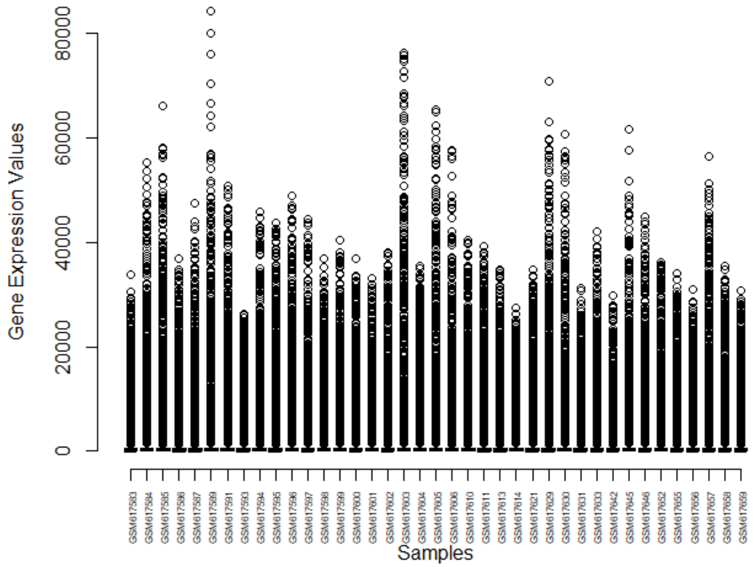
In this study, we conducted a computational analysis of gene expression profile data obtained from 79 cases, 39 of which were classified as having disease recurrence, to investigate whether advanced computational algorithms can derive more accurate prognostic signatures for prostate cancer.

Normalizing Data

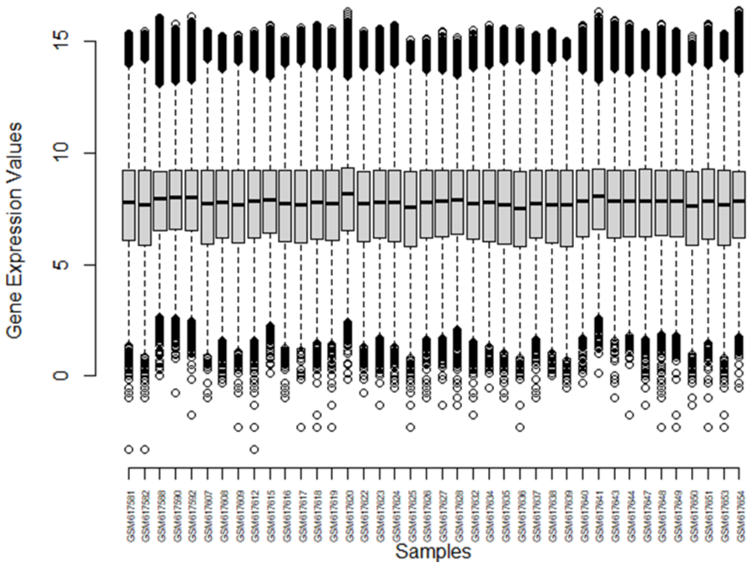
- The first step was to normalize the data
- The boxplots of the non-normalized data (top) from both groups clearly shows that the data is skewed and needs to be normalized
- After data was normalized using a log2 transformation, the boxplots on the bottom show very normal looking distributions



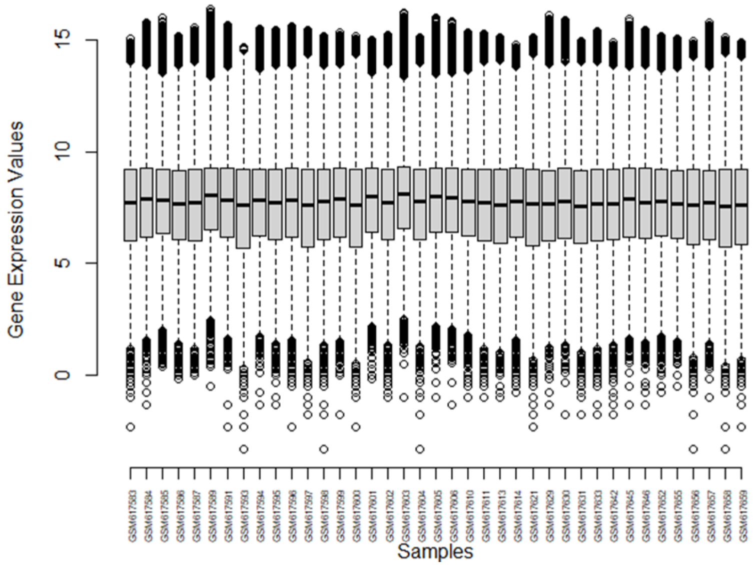
Boxplot of Non-Normalized Data
Non-Recurrent Samples



Boxplot of Non-Normalized Data
Recurrent Samples



Boxplot of Normalized Data
Non-Recurrent Samples

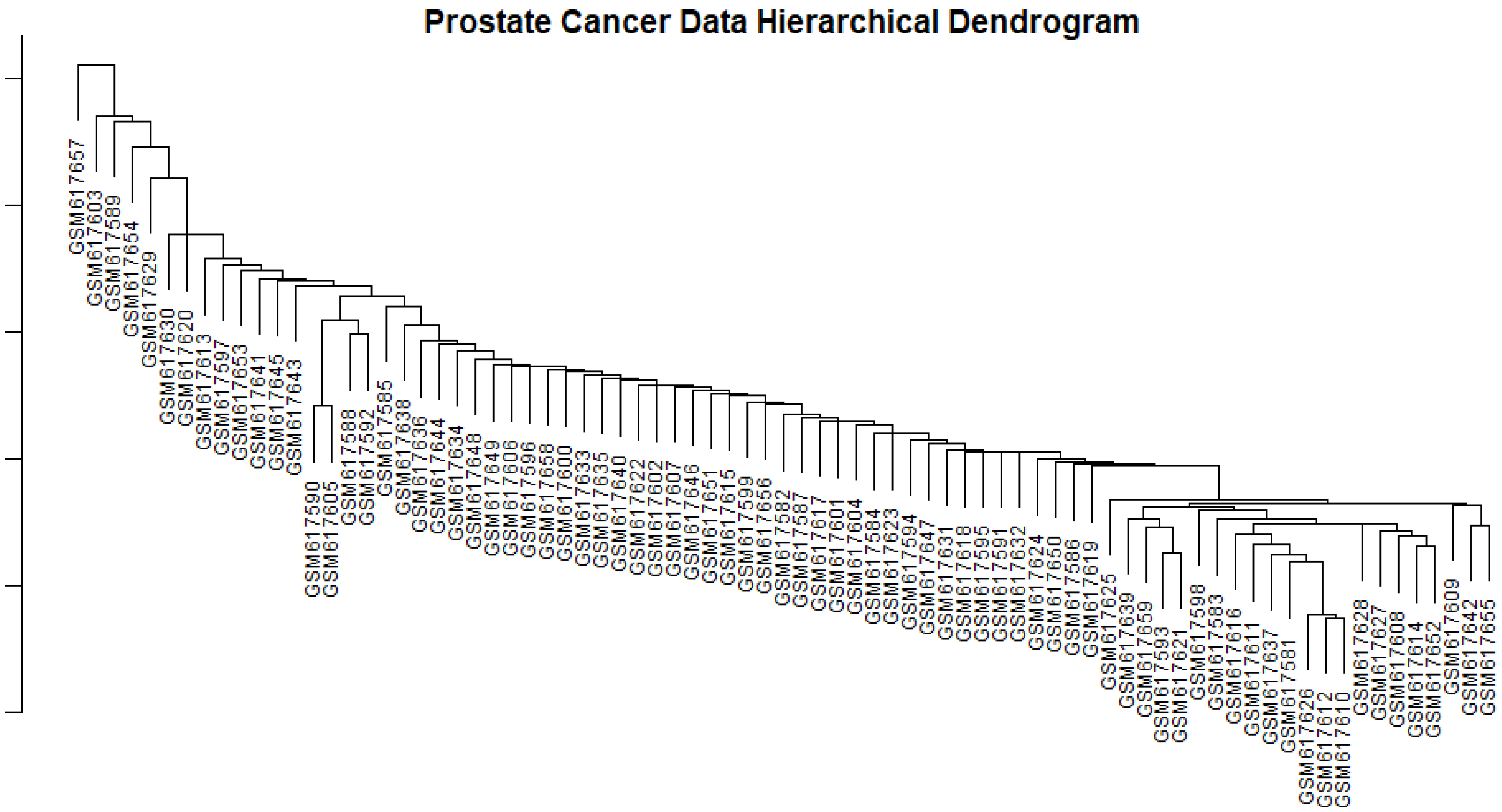


Boxplot of Normalized Data
Recurrent Samples

Outlier Detection – Clustering Dendrogram

- Next the data was examined for potential outliers
- The first method to detect outliers was a clustering dendrogram
    - Euclidean distance
    - Single linkage method
- Potential Outliers include those at the top of the tree: GSM617657, GSM617603, GSM617589, GSM617654 and GSM617629

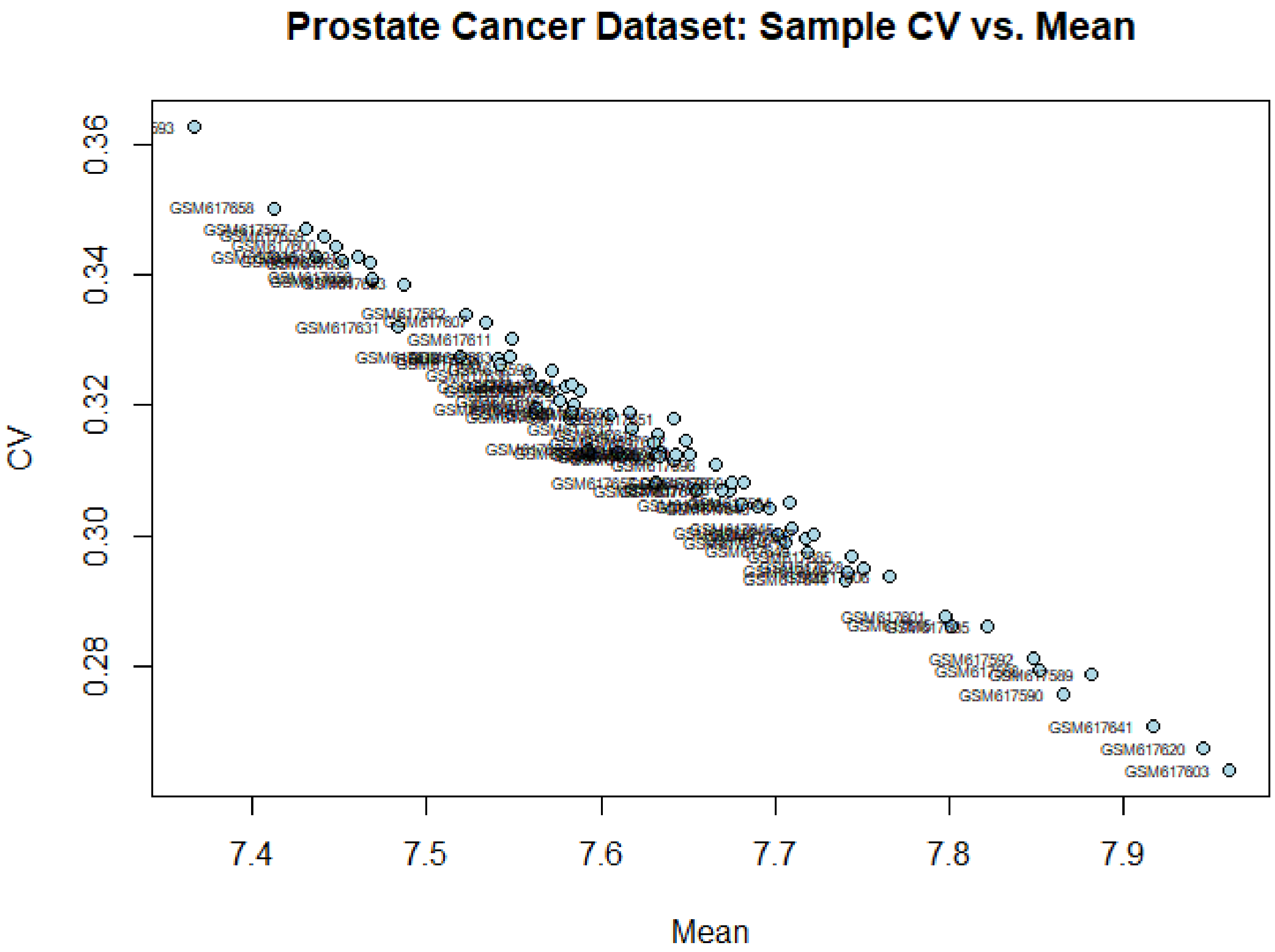
.



Prostate Cancer Data Hierarchical Dendrogram

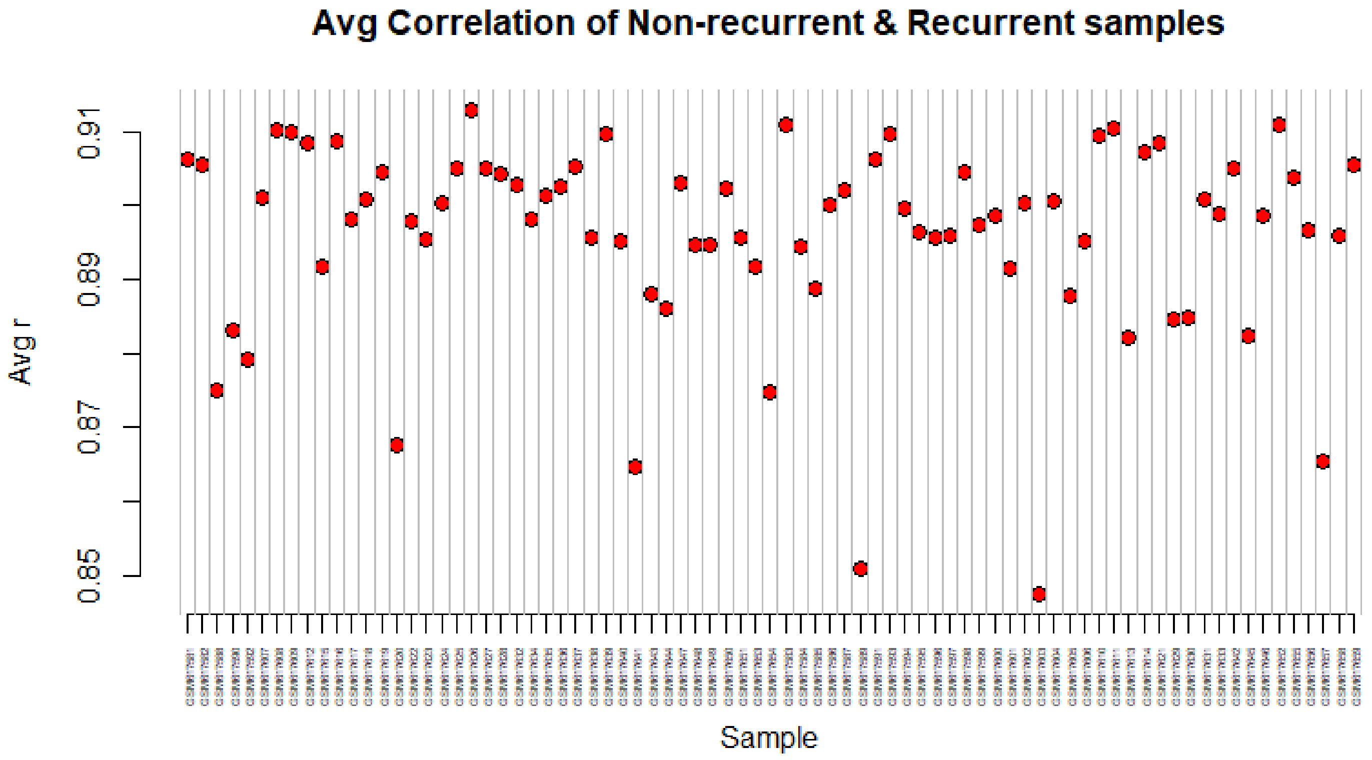
Outlier Detection – CV vs. Mean Plot

- Next a CV vs Mean plot was generated
- Potential Outliers from this test included samples: GSM617593, GSM617603, GSM617620, GSM617841, GSM617589

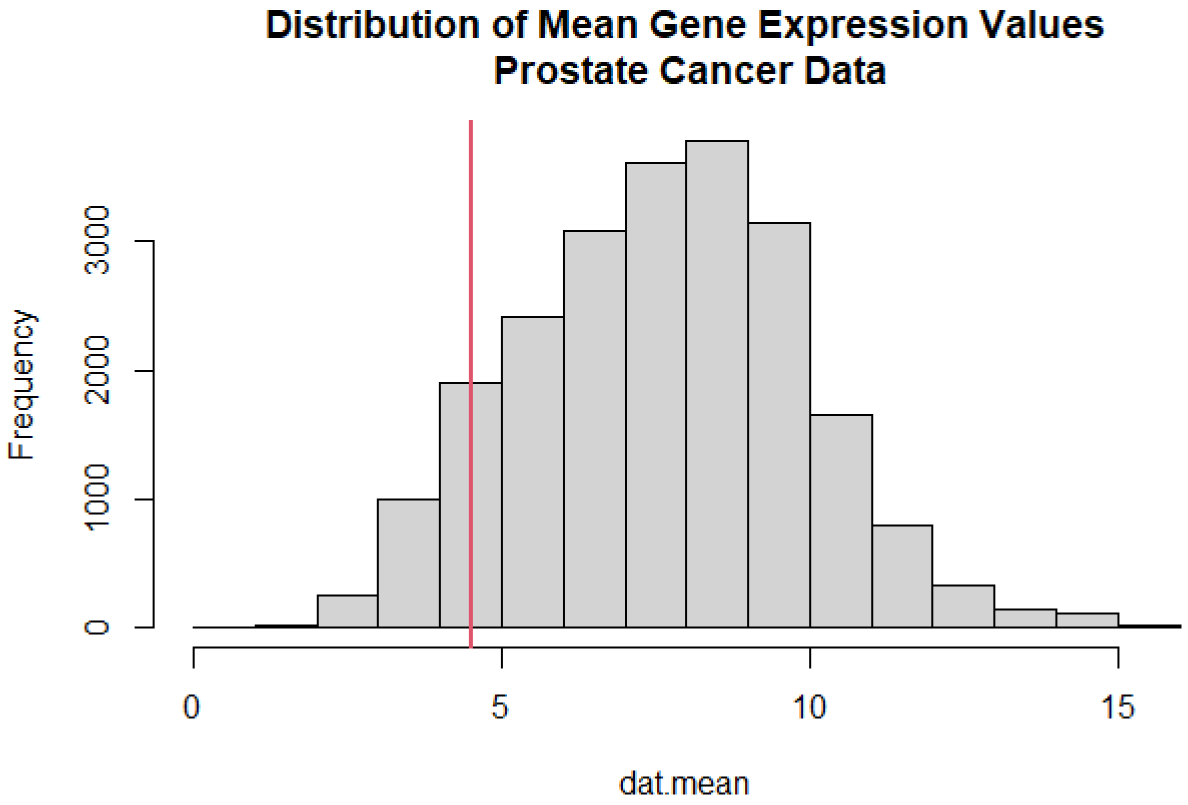


Outlier Detection – Average Correlation

- Finally, an average correlation plot was generated
- Two samples appeared as outliers at an Avg r < 0.86
  - GSM617589 & GSM617603
- Both of these samples appeared as potential outliers in all tests and were removed for further analysis
  - Both were recurrent samples



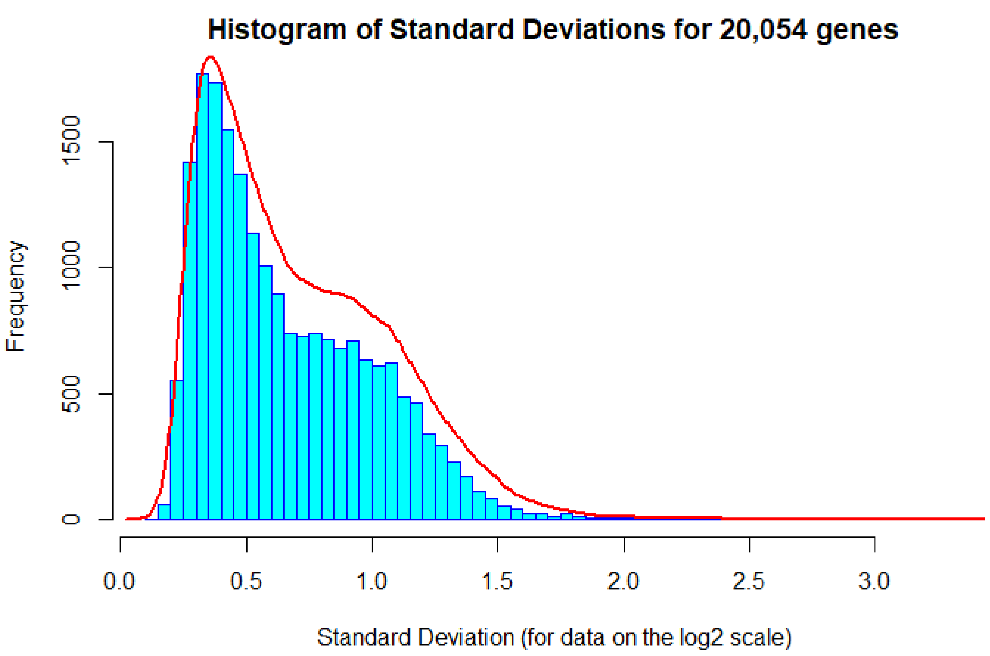
Gene Filtering

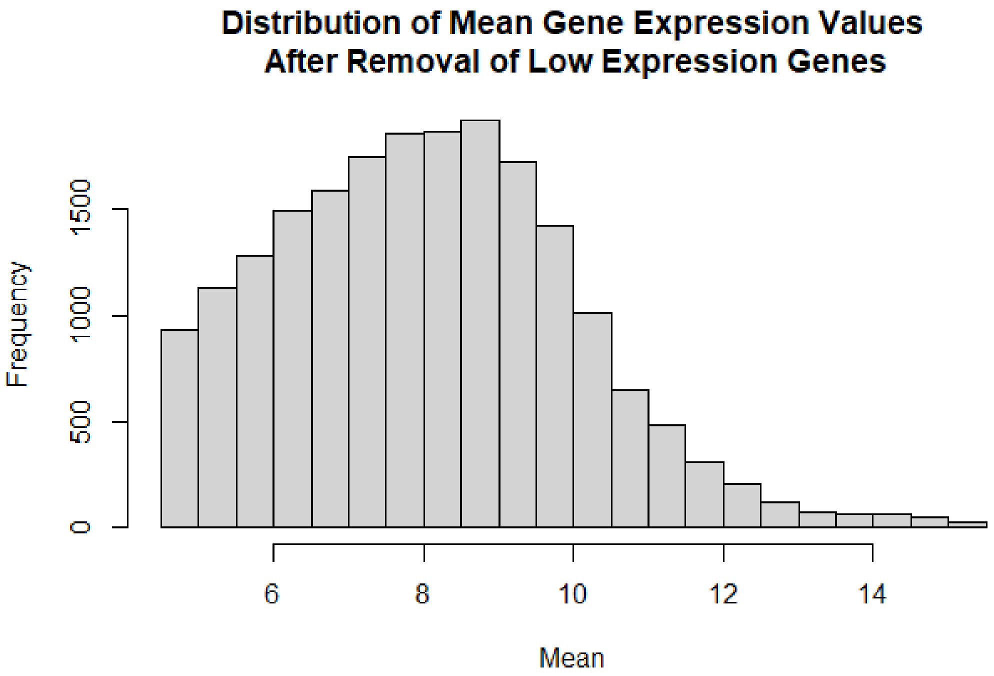


- Filtered out low expression genes using mean values
- Red line indicates bottom 10% of low expressed genes
- All genes below this threshold were removed prior to further analysis

Feature Selection

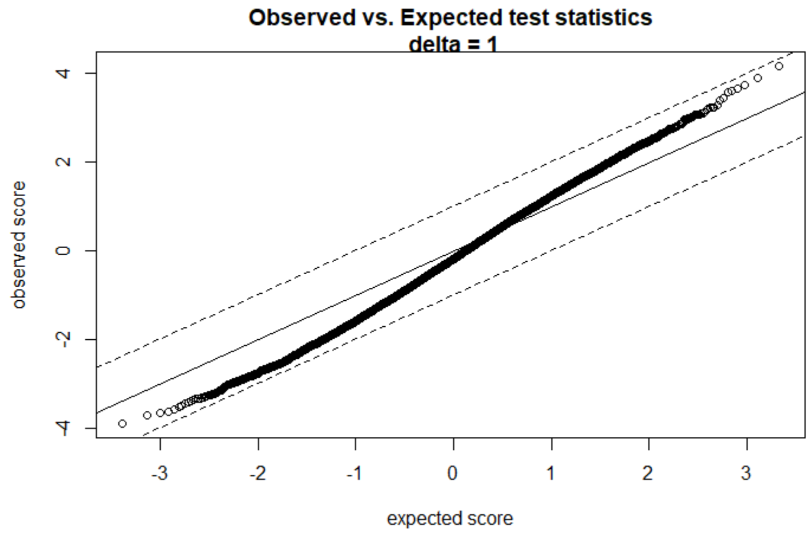
- To determine which statistical test to use for selecting features, the first step was to examine the distributions of the new, cleaned data
- The distributions of the standard deviations and of the means suggest that a non-parametric or permutation test would be most appropriate

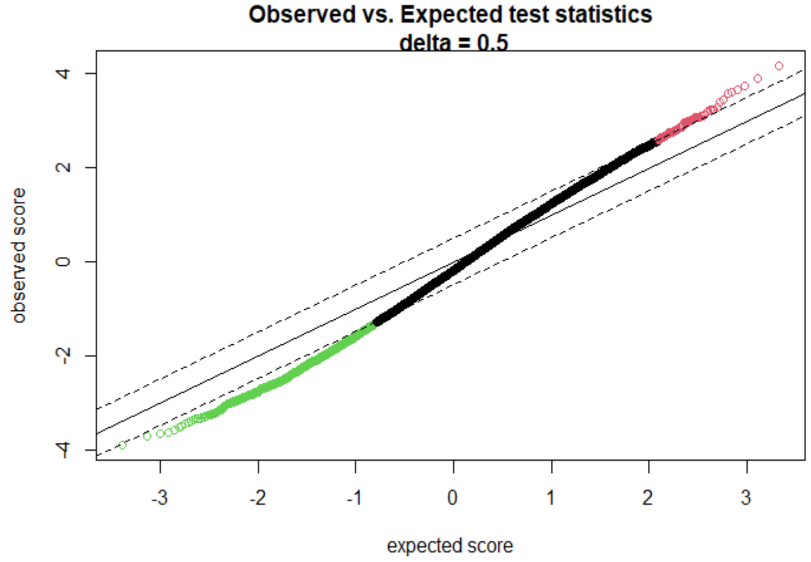




Feature Selection - SAM

- The first method of feature selection was to try a permutation method
- The SAM method was calculated with 100 permutations
- At a conservative delta of 1, no genes were found to be differentially expressed
- At delta of 0.5, over 3,000 genes were differentially expressed
- However, given this is a pretty low threshold value it also makes sense to examine the non-parametric methods



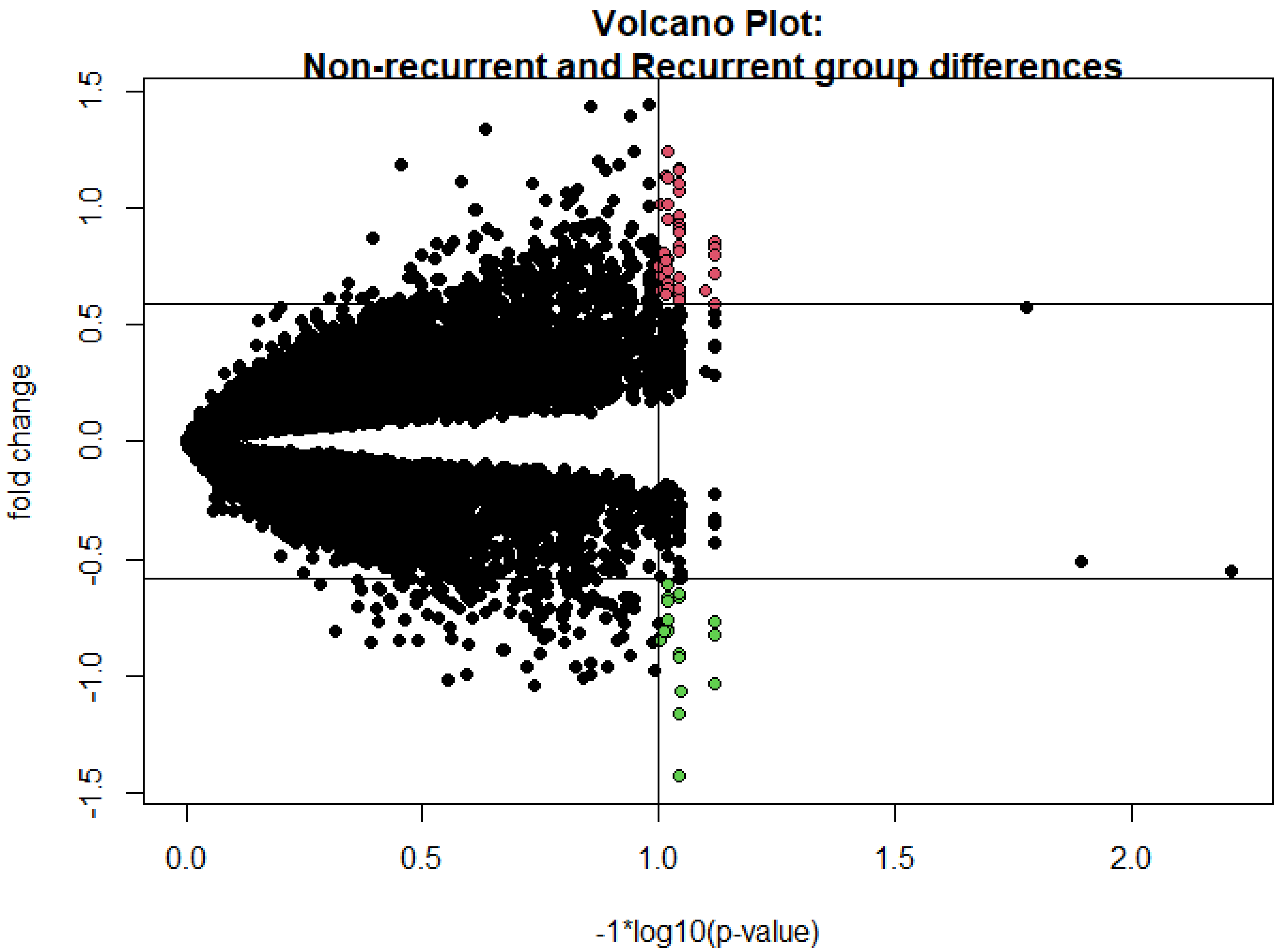


Feature Selection – Non-parametric tests

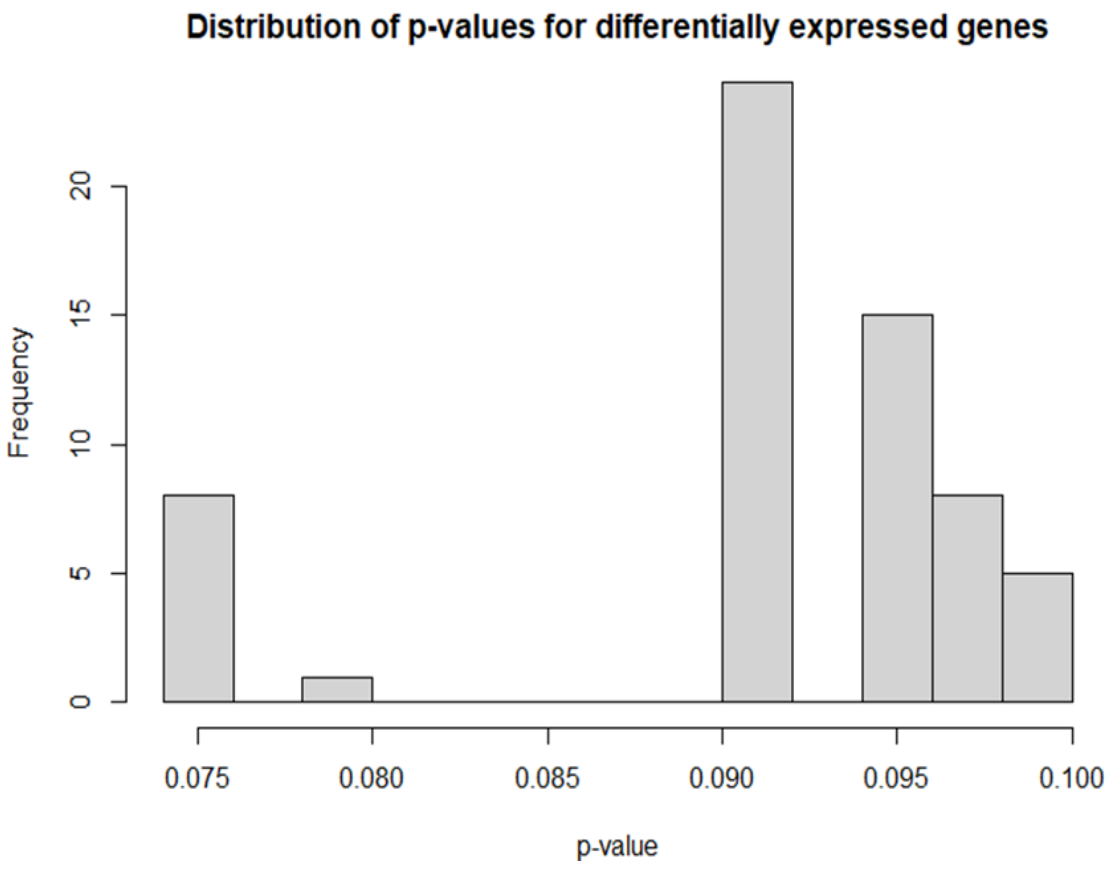
- Empirical Bayes and the Wilcoxon-Mann-Whitney test were also used to look for differentially expressed genes
- Empirical Bayes
  - The number of genes with p-value < 0.05 was 3,235
  - After adjusting for multiplicity using Benjamini Hochberg correction the number of genes with p-value < 0.05 was 3
  - Number of genes with a p-value < 0.1 after adjustment was 182
- Wilcoxon-Mann-Whitney
  - The number of genes with p-value < 0.05 was 3,245
  - After adjusting for multiplicity using Benjamini Hochberg correction the number of genes with p-value < 0.05 was 3
  - Number of genes with a p-value < 0.1 after adjustment was 211
  - Given that slightly more genes were found to be significant with this method versus the empirical bayes method, these adjusted p-values were used in the rest of the analysis

Fold Changes

- Fold changes were also calculated for all genes
- Similar to the previous tests, only a small number of genes had significant fold changes
  - Only 28 genes had a linear fold change > 1 in the positive direction and only 7 had a linear fold change > 1 in the negative direction
  - At a more conservative fold change of log2(1.5), 349 genes were found in the positive direction and 151 were found in the negative direction
- The volcano plot shows differentially expressed genes using the criteria of adjusted p-value < 0.1 and fold changes > log2(1.5) and < -log2(1.5)
  - Red points represent differentially expressed genes in the positive direction
  - Green points represent differentially expressed genes in the negative direction
- These differentially expressed genes were ultimately used for the final analysis



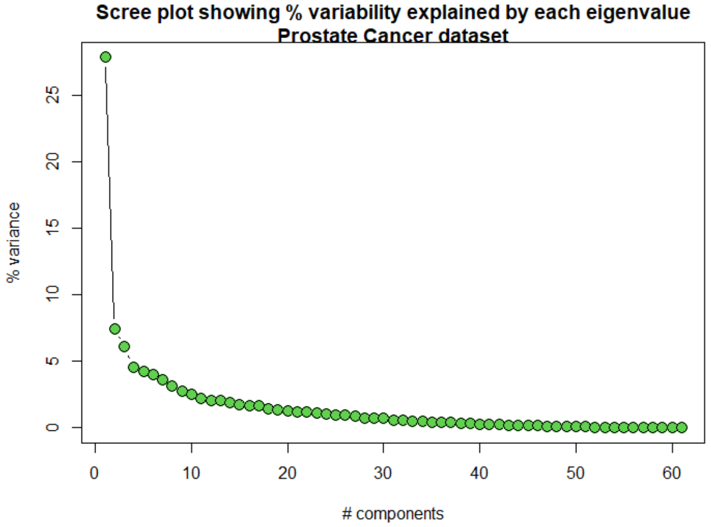
Differentially Expressed Genes

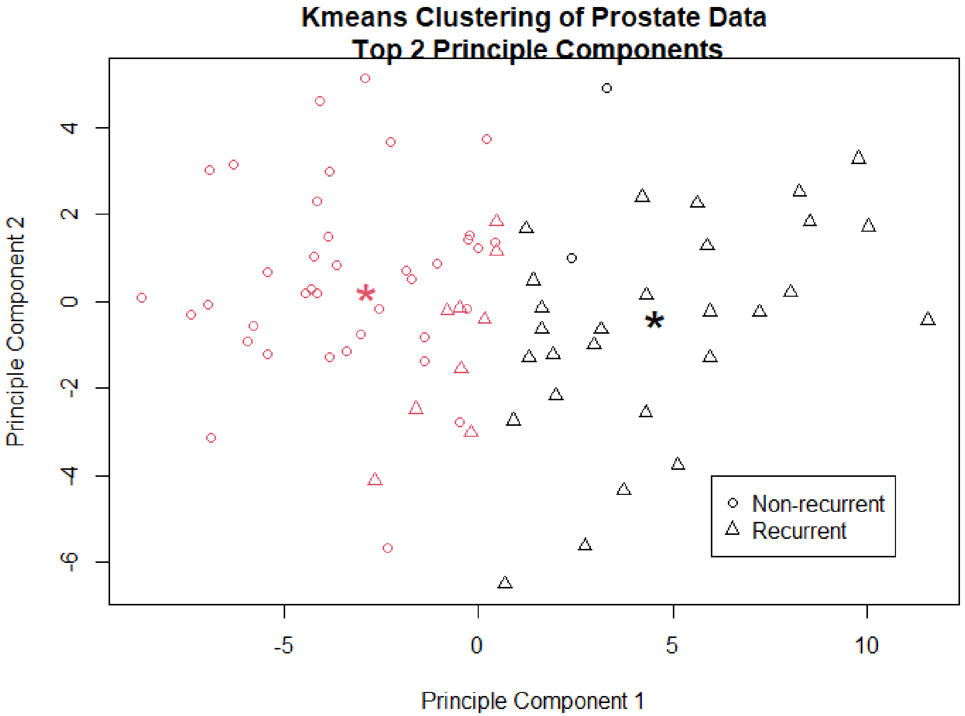


- A total of 61 genes were selected using p-values under 0.1 and fold changes greater than and less than log2(1.5)
  - 43 were found to be more highly expressed in the non-recurrent samples (positive direction)
  - 18 were found to be more highly expressed in the recurrent samples (negative direction)
- These are very conservative estimates on significance but in this case these thresholds were necessary in order to find an adequate number of differentially expressed genes to be used for further analysis
- The lack of significance between non-recurrent and recurrent tumor samples could indicate that both these tumor types have fairly similar molecular profiles

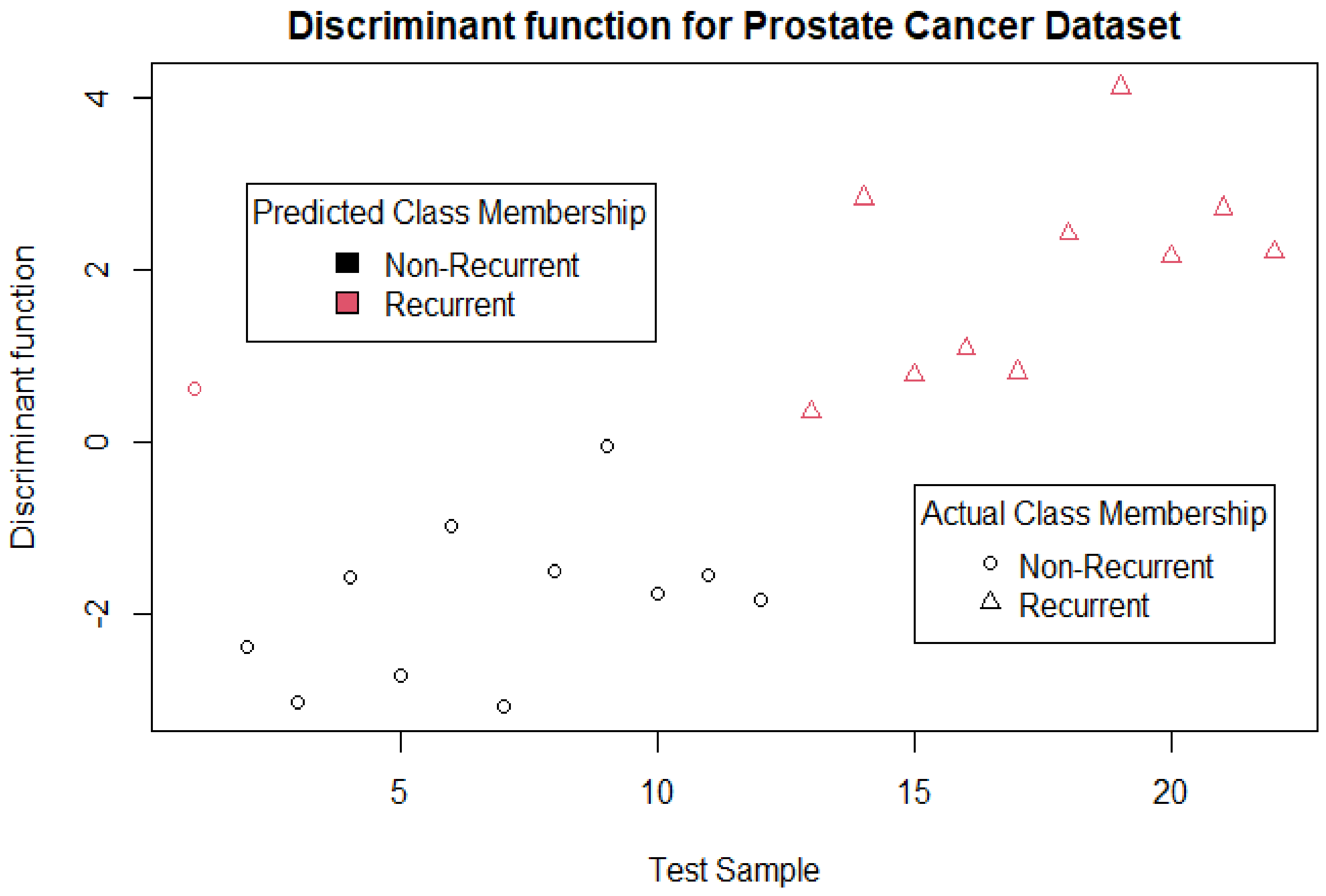
Principle Component Analysis

- To visualize the data in two dimensional space, principle component analysis was performed on the 61 DEGs and the top 2 principle components were plotted against one another following K-means clustering of these components
- The scree plot shows that roughly 35% of the variability within the data is captured by these first 2 components
- The figure at the bottom shows that PCA has done a fairly good job at separating the two groups from one another





Classification



- The final step was to perform a method of classification using the principal components
- The data was split into a training set consisting of 55 samples (28 non-recurrent and 27 recurrent) and 22 samples in the test set (12 non-recurrent & 10 recurrent)
- Linear discriminant analysis was used for the classification method
- The model was able to accurately predict 21 out of 22 of the test samples
    - However, this high performance may indicate overfitting

Top 5 Discriminant Genes – Negative Direction

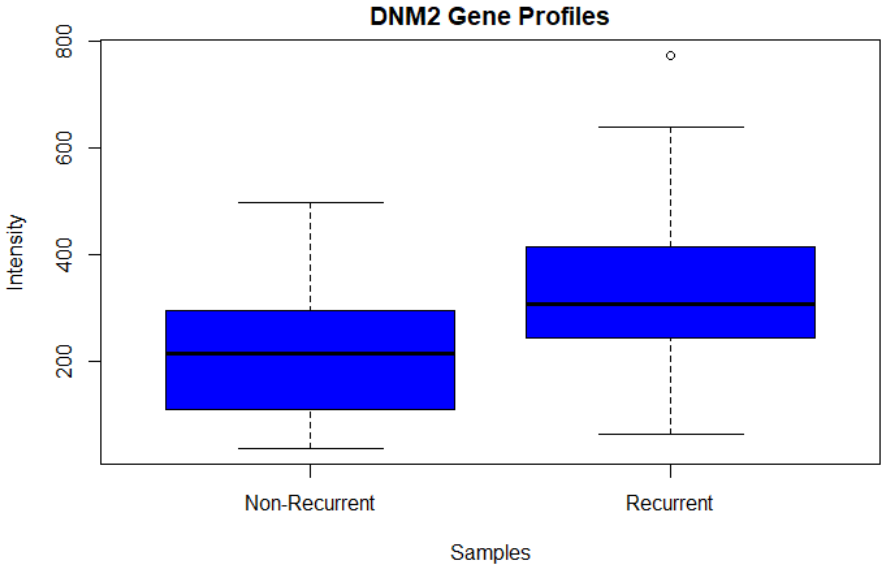
- Genes more highly expressed in the recurrent samples
- ICAM1 - Intracellular adhesion molecule 1(*215485_s_at*)
    - KEGG Pathways: NF-kappa B signaling pathway, cell adhesion molecules (CAMs), natural killer cell mediated cytotoxicity, TNF signaling pathway, leukocyte trans endothelial migration
    - GO Terms: regulation of leukocyte mediated cytotoxicity, T cell antigen processing and presentation, positive regulation of NF-kappa B transcription factor activity
- LRRC41 – Leucine rich repeat containing 41 (*215765_s_at*)
    - KEGG Pathways: none listed
    - GO Terms: protein ubiquitination, protein homodimerization activity
- PNMA2 - Paraneoplastic Ma antigen 2 (*209597_s_at*)
    - KEGG Pathways: none listed
    - GO Terms: Positive regulation of apoptotic process
- PPBP – Pro-platelet basic protein (*214146_s_at*)
    - KEGG Pathways: Cytokine-cytokine receptor interaction, chemokine signaling pathway
    - GO Terms: Leukocyte migration involved in inflammatory response, positive regulation of leukocyte chemotaxis, G-protein coupled receptor signaling pathway, positive regulation of cell division and proliferation
- DNM2 - Dynamin 2 (*216024_at*)
    - KEGG Pathways: Endocytosis, synaptic vesicle cycle, endocrine and other factor-regulated calcium reabsorption
    - GO Terms: G2/M transition of mitotic cell cycle, regulation of transcription, signal transduction, spermatogenesis, negative regulation of transforming growth factor beta receptor signaling pathway

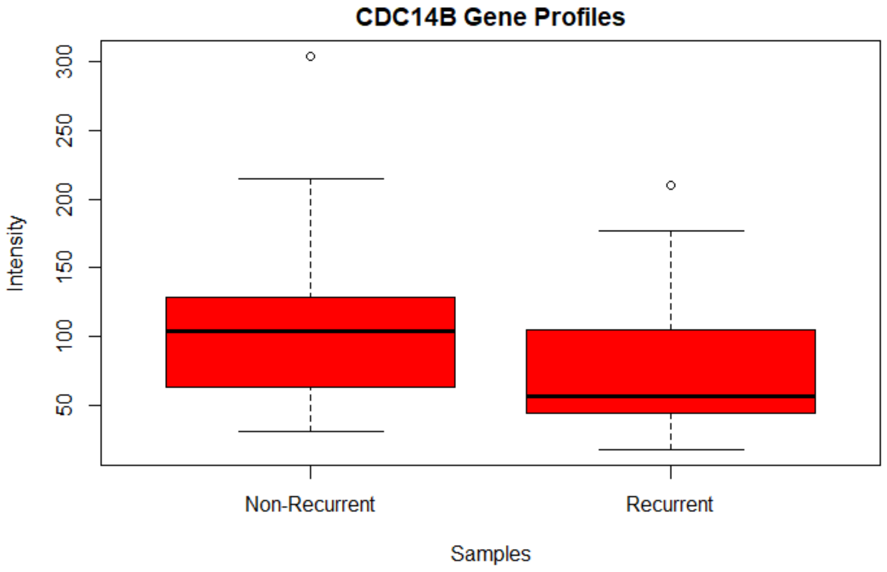
Top 5 Discriminant Genes – Positive Direction

- These genes were more highly expressed in the non-recurrent samples
- TTLL5 – tubulin tyrosine ligase like 5 (*208099_x_at*)
  - KEGG Pathway: none listed
  - GO Terms: ATP binding, transcription, cellular protein modification process, sperm axoneme assembly, fertilization, protein polyglutamylation, sperm motility
- FGFR2 – fibroblast growth factor receptor 2 (*208225_at*)
  - KEGG Pathway: MAPK signaling pathway, Ras signaling pathway, Rap1 signaling pathway, P13K-Ak1 signaling pathway, pathways in cancer, prostate cancer
  - GO Terms: Protein kinase activity, MAPK cascade, angiogenesis, apoptotic process, cell-cell signaling, positive regulation of cell proliferation, positive regulation of ERK1 and ERK2 cascade, positive regulation of cell division, cell fate commitment
- MTIF2 – mitochondrial translational initiation factor 2 (*203095_at*)
  - KEGG Pathway: none listed
  - GO Terms: regulation of translational initiation, ribosome disassembly, mitochondrial translational initiation, translation factor activity, GTPase activity
- MAGEA10 – MAGE family member A10 (*210295_at*)
  - KEGG Pathway: function not known but may play a role in embryonal development and tumor transformation or aspects of tumor progression
  - GO Terms: nucleus
- CDC14B – cell division cycle 14B (*211348_s_at*)
  - KEGG Pathway: Cell cycle
  - GO Terms: DNA repair, protein dephosphorylation, G2 DNA damage checkpoint, positive regulation of ubiquitin protein ligase activity

DNM2 and CDC14B

- Exploring the top positive and negative genes a little more
- DNM2 was shown in this analysis to be more highly expressed within the recurrent samples and has recently been shown to promote cancer cell growth, migration and invasion in a diverse set of cancers[6]
  - This has the potential to be used as a biomarker to predict prostate cancer recurrence
- CDC14B is believed to be involved in cell cycle control and has even been shown to interact with the tumor suppressor protein p53, perhaps playing a role in its regulation[1]
  - The lower levels within the recurrent samples may indicate less p53 activity which may be playing a part in tumor recurrence





Conclusions

- It was difficult to find many statistically significant differences between the two groups using common threshold values which may indicate similar molecular profiles; however, this may also mean that there are a few key genes that are primarily responsible for the difference between tumor recurrence and non-recurrence and these could possibly be used as effective biomarkers in the future for better prognosis
- Although conservative thresholds had to be used in terms of identifying differentially expressed genes, the principal component analysis and classification model have shown that these genes performed well in separating out the non-recurrent versus recurrent tumor types
- The fact that the classification model performed as well as it did may indicate that the model was overfitting to the data, finding new data or a using a leave one out approach would be warranted to better assess how well these variables can really predict tumor recurrence
- Among the genes identified in this analysis, DNM2 is an interesting target for further exploration given that it has recently been found to play an important role in cancer development and proliferation in a diverse set of tumors

References

1. CDC14B Gene. Gene Cards: The Human Gene Database. Retrieved August 14, 2021 from https://www.genecards.org/cgi-bin/carddisp.pl?gene=CDC14B
2. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. Nature Protoc. 2009;4(1):44-57.
3. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.
4. Sorin, D. (2012). Statistics and Data Analysis for Microarrays Using R and Bioconductor.
5. Sun, Y., & Goodison, S. (2009). Optimizing molecular signatures for predicting prostate cancer recurrence. The Prostate, 69(10), 1119–1127. https://doi.org/10.1002/pros.20961
6. Raja SA, Shah STA, Tariq A, et al. Caveolin-1 and dynamin-2 overexpression is associated with the progression of bladder cancer. Oncol Lett. 2019;18(1):219-226. doi:10.3892/ol.2019.10310