Classification for Renal Cell Carcinomas
Final project presentation

Alex Anderson, K. Jarrod Millman, and Lara Troszak

Public Health 240F: Statistical Genomics II

Thursday, May 8, 2014

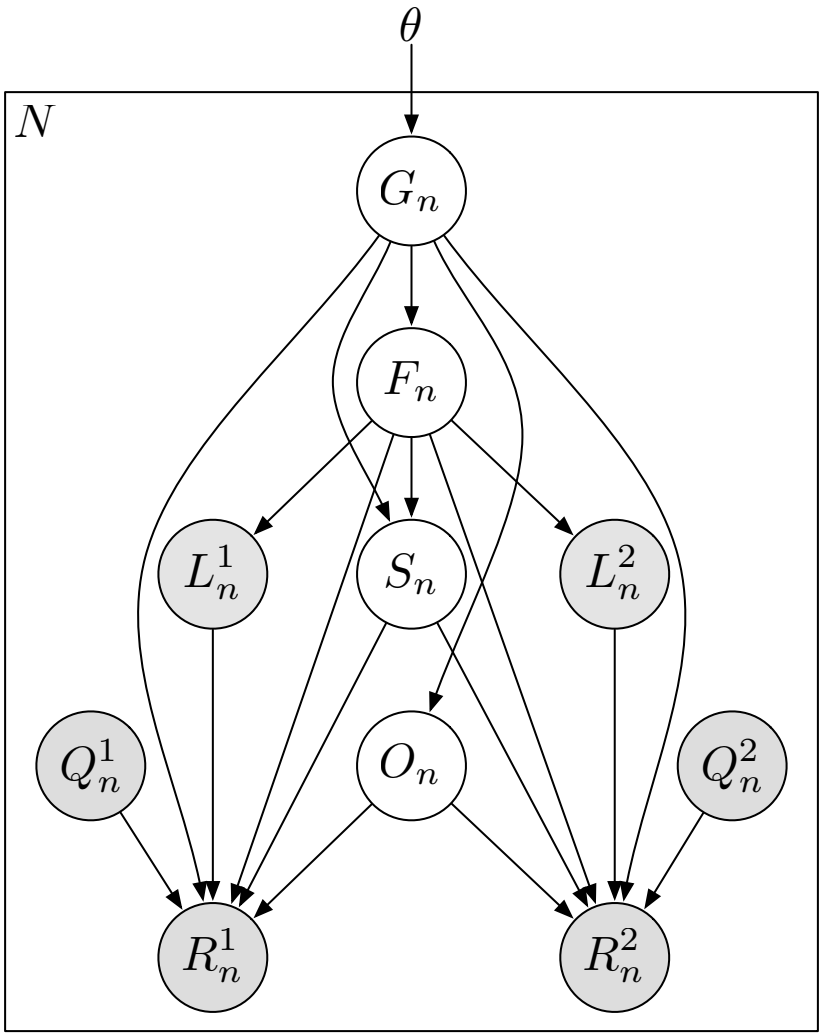
Goal

Classification of

- *Kidney renal papillary cell carcinoma (KIRP)*
- *Kidney renal clear cell carcinoma (KIRC)*

based on TCGA RNAseq abundance estimates

Data: Expected read counts with RSEM



Data: training and validation sets

- 72 samples
  - 35 KIRC / 37 KIRP
  - 47 training / 25 test
  - 24 KIRC / 23 KIRP (training)

Reproducibility

All code, figures, and text for this project are available on Github:
`https://github.com/jarrodmillman/ph240f`

Method

EDA

- Log-transformed
- Filtering
- Normalization

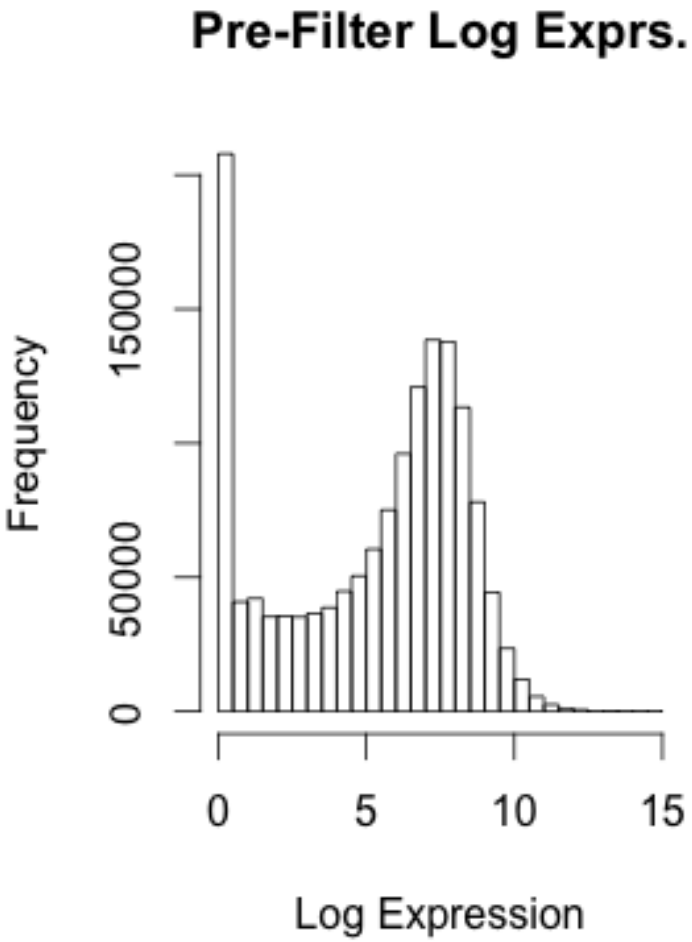
Feature Selection

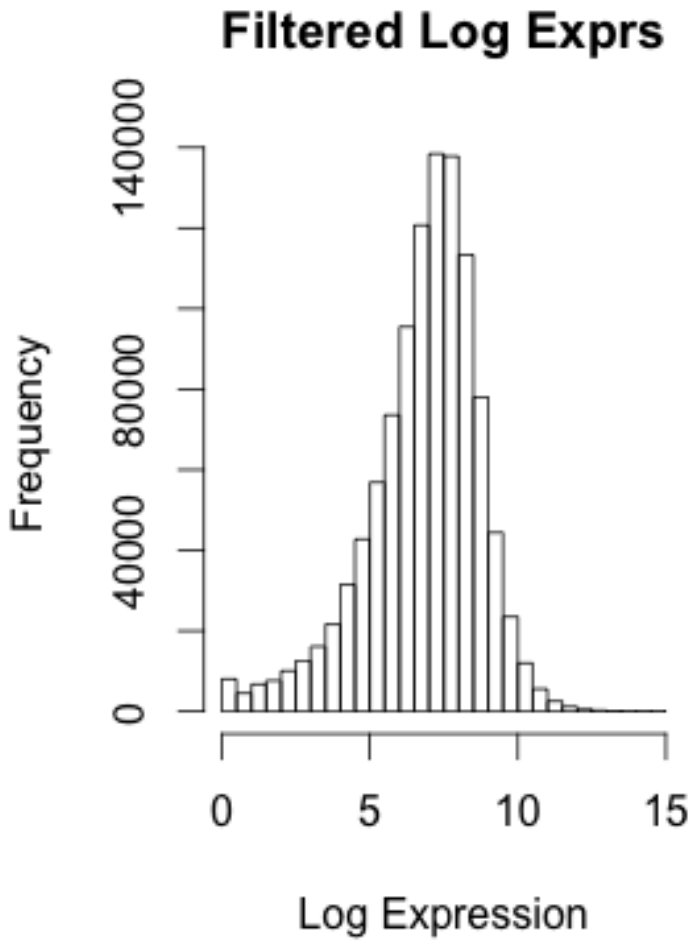
- Variance
- T-score

Classification methods

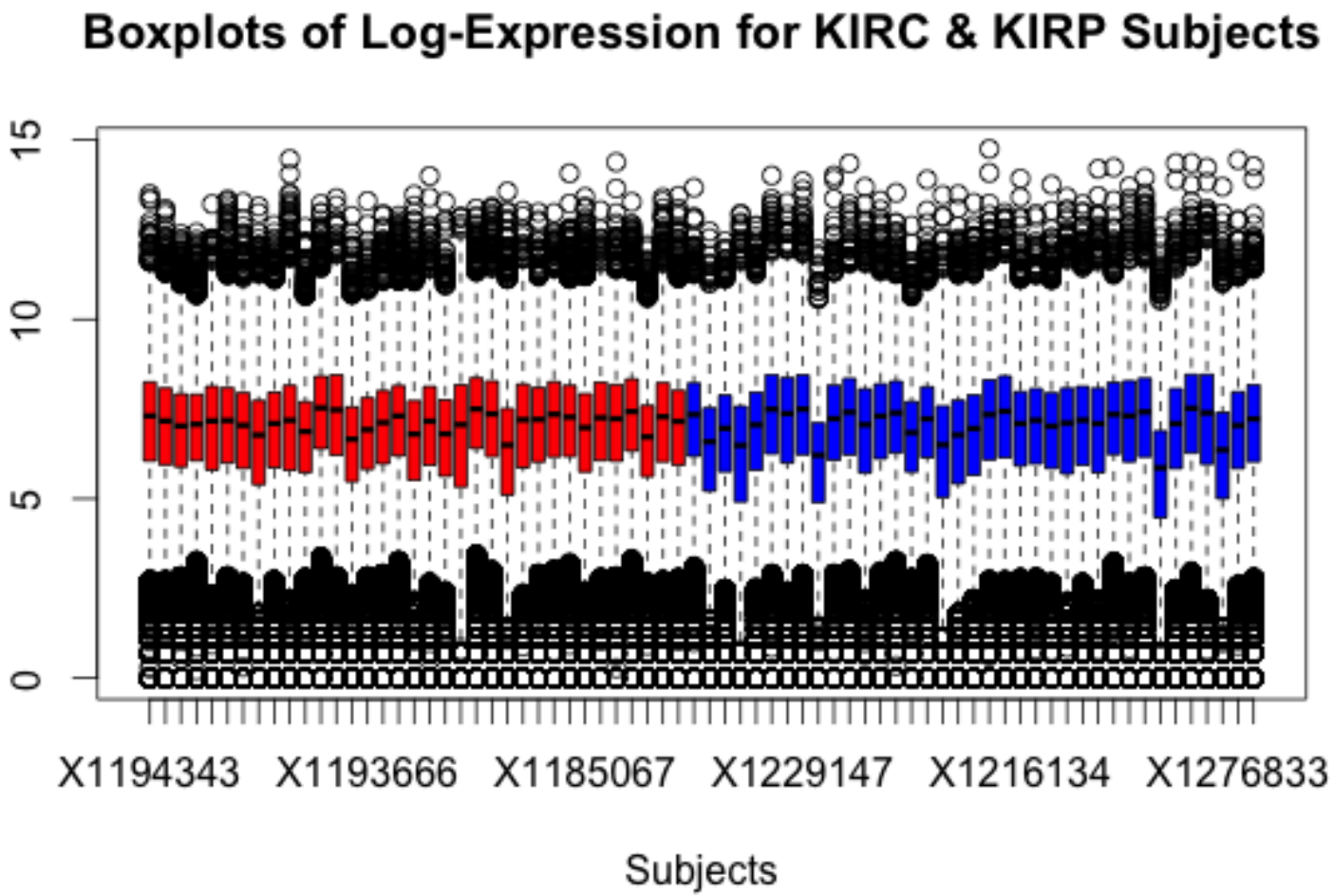
- LDA with features selected
- SVM with features selected
- Random forest

EDA: Filtering

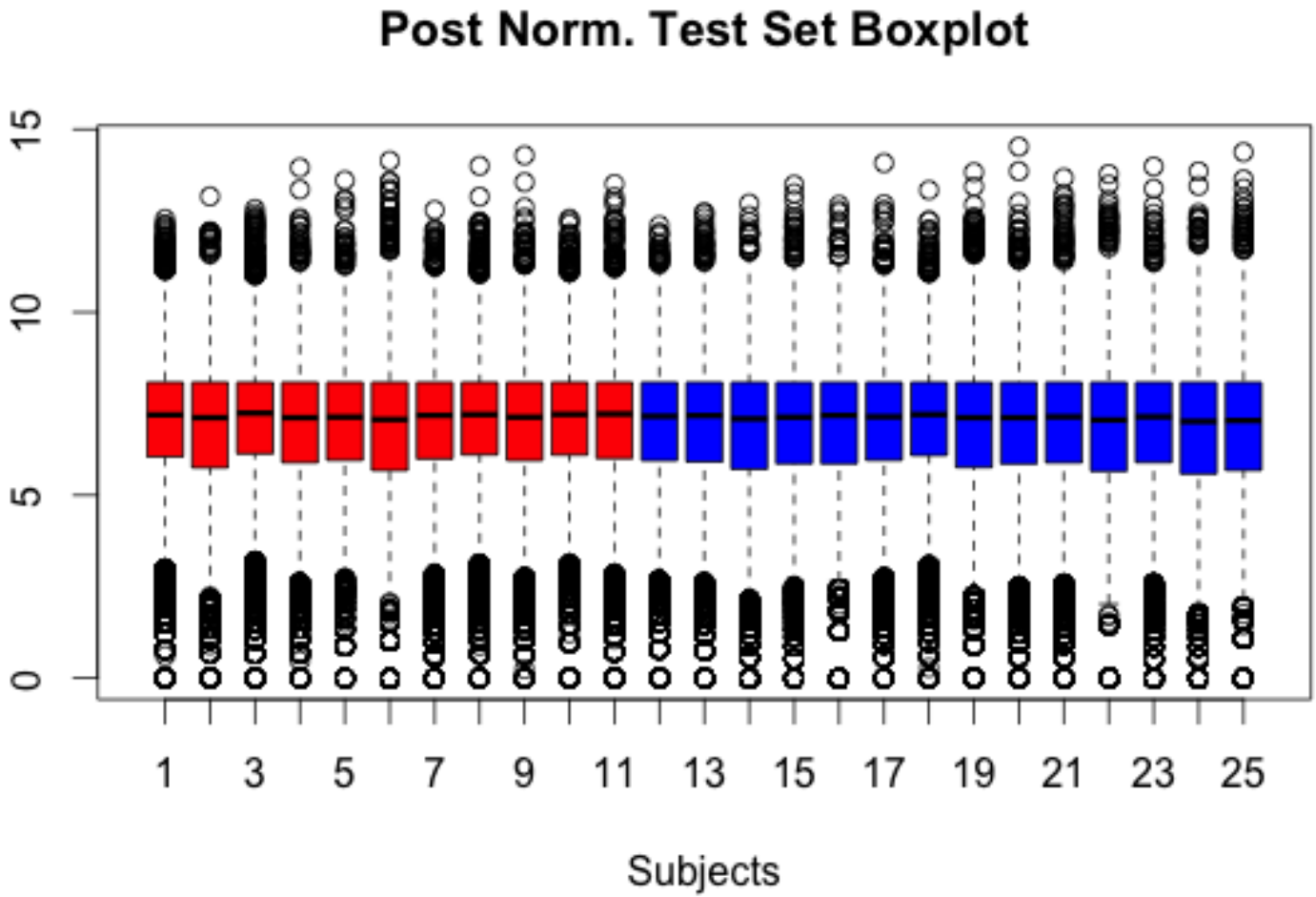




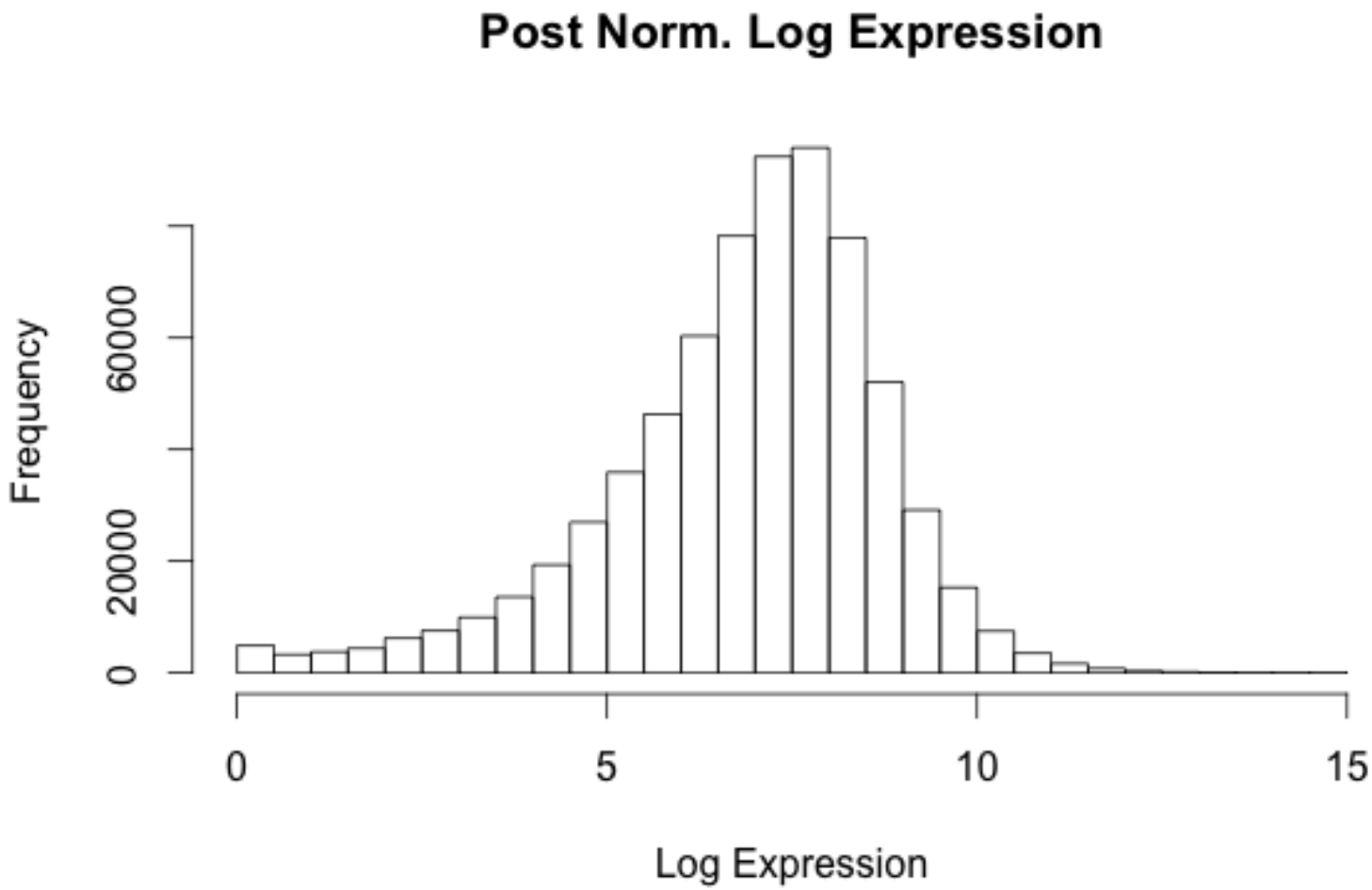
EDA: Normalization



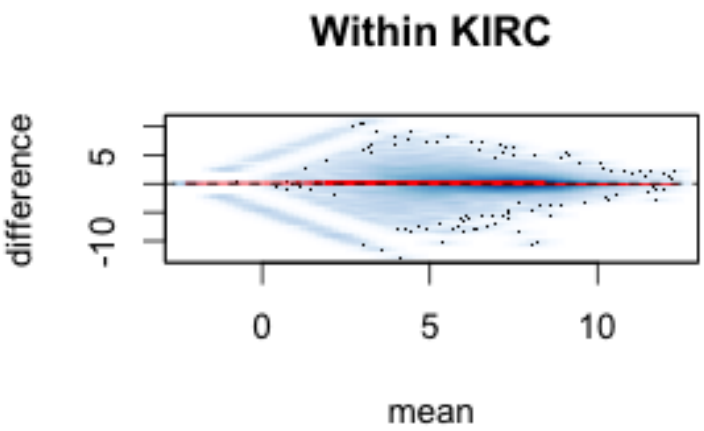
EDA: Normalization

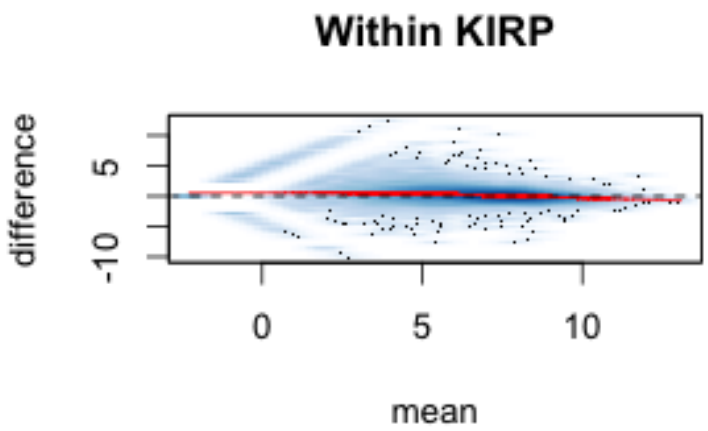


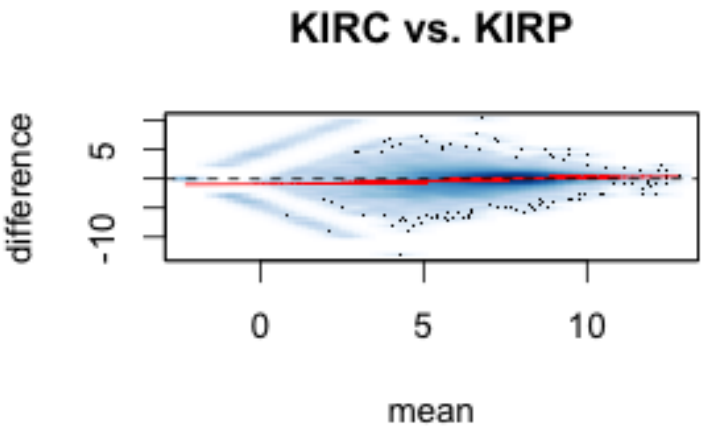
EDA: Normalization

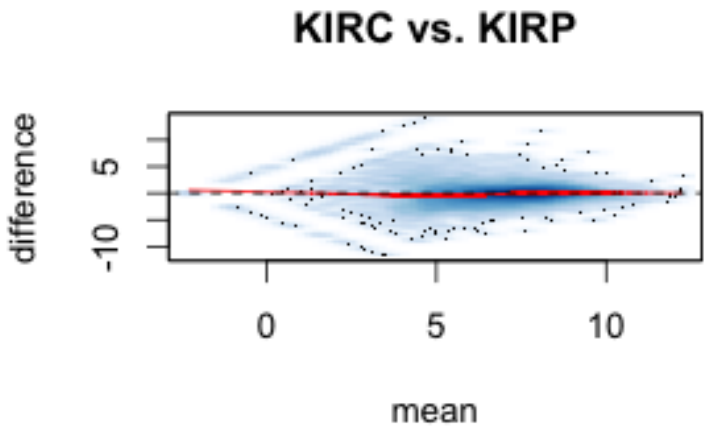


EDA: MD plots

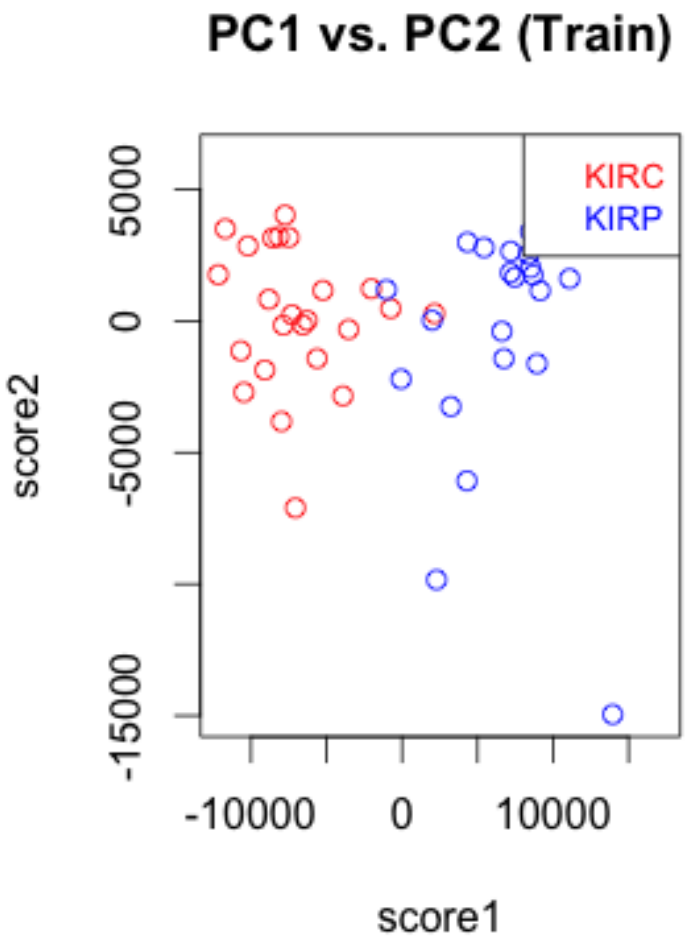

Within KIRC
difference
5
-10
0
5
10
mean

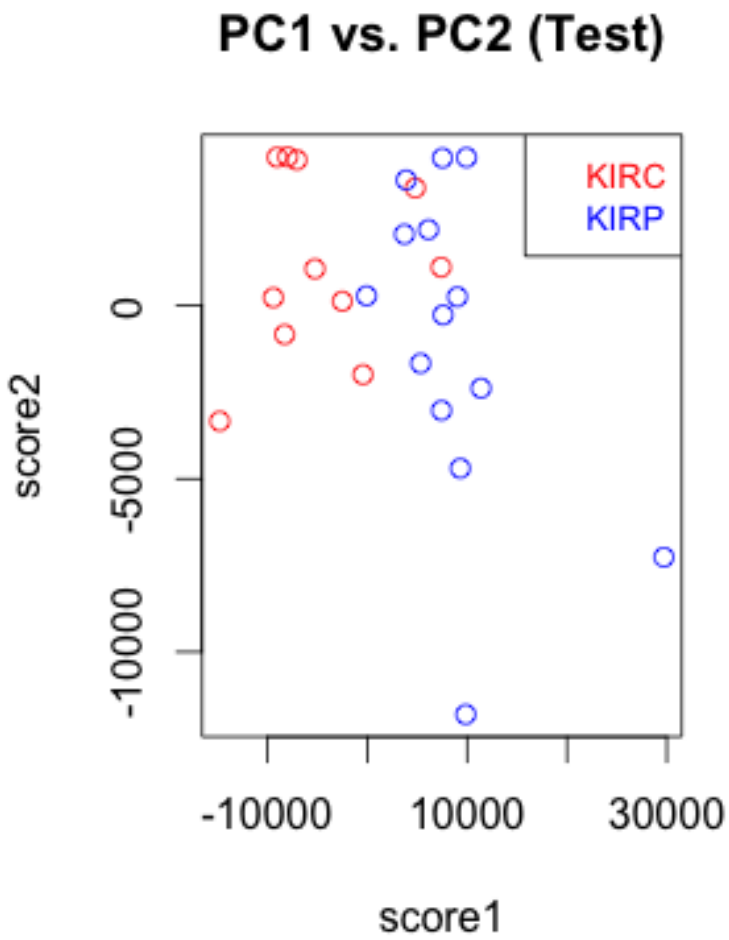

Within KIRP
difference
5
-10
0
5
10
mean


KIRC vs. KIRP
difference
5
-10
0
5
10
mean


KIRC vs. KIRP
difference
5
-10
0
5
10
mean

EDA: PCA


PC1 vs. PC2 (Train)
KIRC
KIRP
score2
5000
0
-5000
-15000
-10000
0
10000
score1


PC1 vs. PC2 (Test)
KIRC
KIRP
score2
0
-5000
-10000
-10000
10000
30000
score1

Method

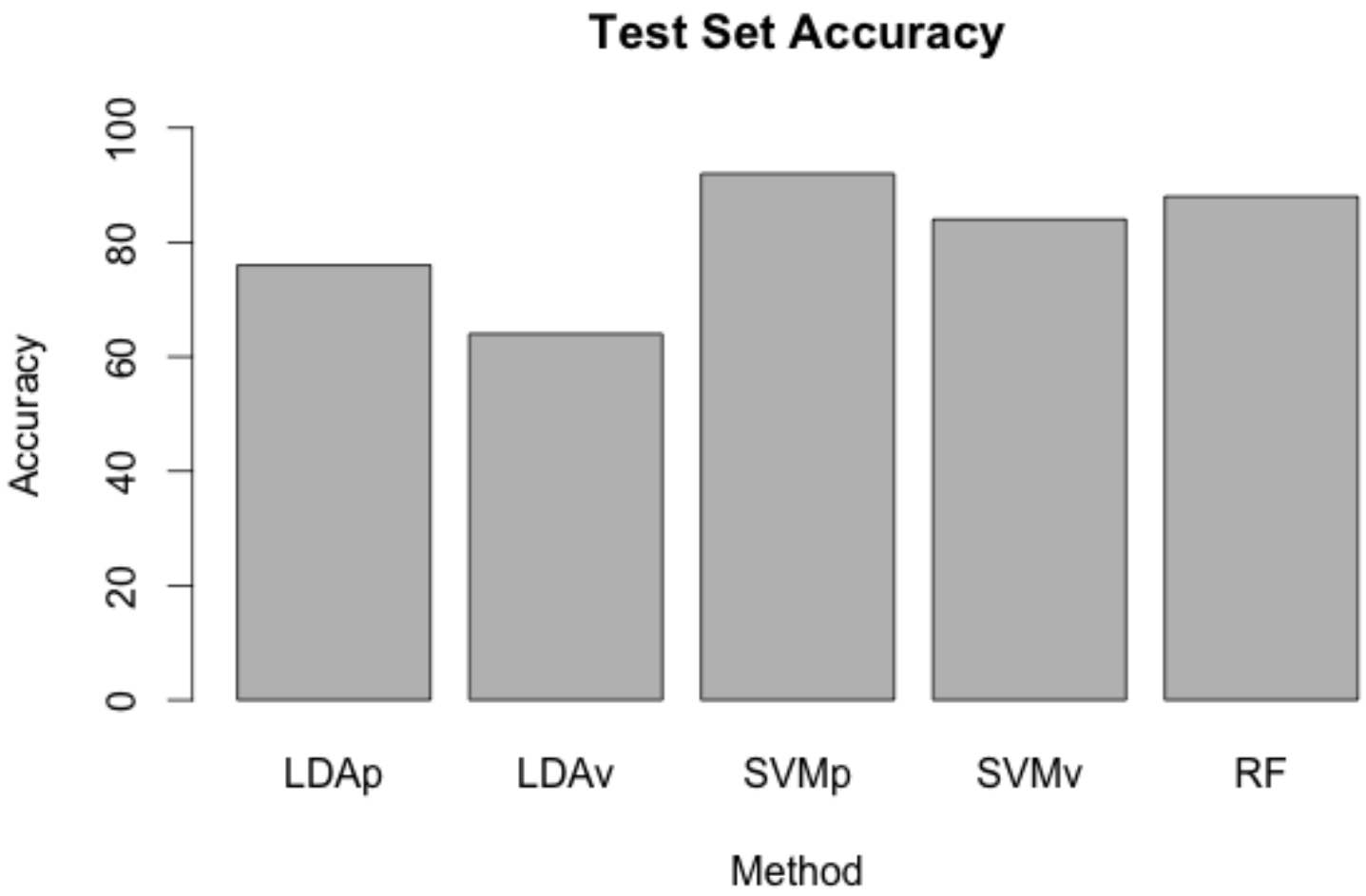
Feature Selection

- Variance
- T-score

Classification methods

- LDA with features selected
- SVM with features selected
- Random forest

Results: Performance



Results: LDA confusion matrix

Variance

pred	true KIRC	KIRP
KIRC	7	4
KIRP	4	10

Tscore

pred	true KIRC	KIRP
KIRC	8	3
KIRP	3	11

Results: SVM confusion matrix

Variance

pred	true KIRC	KIRP
KIRC	7	0
KIRP	4	14

Tscore

pred	true KIRC	KIRP
KIRC	10	0
KIRP	1	14

Results: Random Forest confusion matrix

	true	
pred	KIRC	KIRP
KIRC	8	0
KIRP	3	14

Summary

- Good performance
- Test statistic worked better than variance ranking
- Future directions
  - Technical and/or biological nuisance effects
  - Error bars
  - Biological significance
  - Sensitivity study
  - Non-cancer data