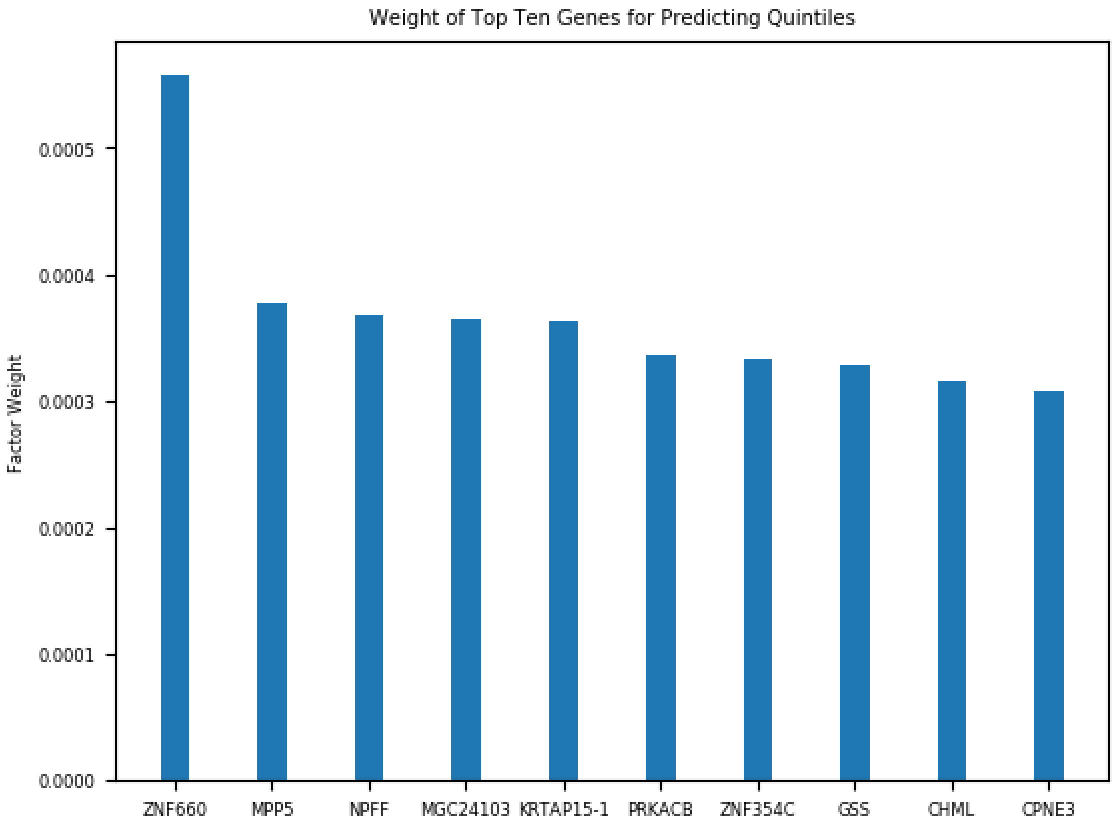
Report on Quintile Machine Learning Predictions

We were able to get some success with your sugestion of sperating into quintles and predicting which quintile the paitent falls into. These are the accuracy of diffrent machine learning models after removing the patietns who survived. We believe with a larger patient sample pool we will be able to increase the accuracy

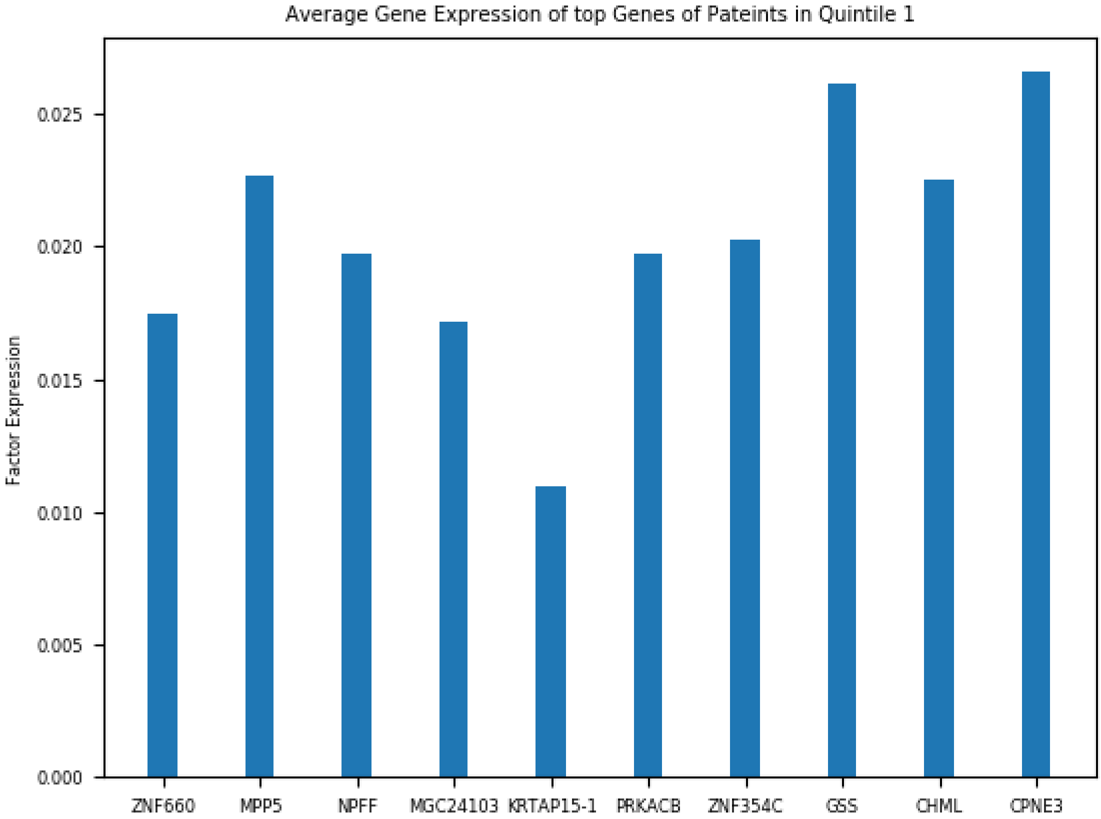
```
Total Number of Patients: N = 258

Accuracy of Random Forest is : 28%

Accuracy of Trees is : 33%

Accuracy of Support Vector Classification is : 16%

Accuracy of Logistical Regression is : 26%
```

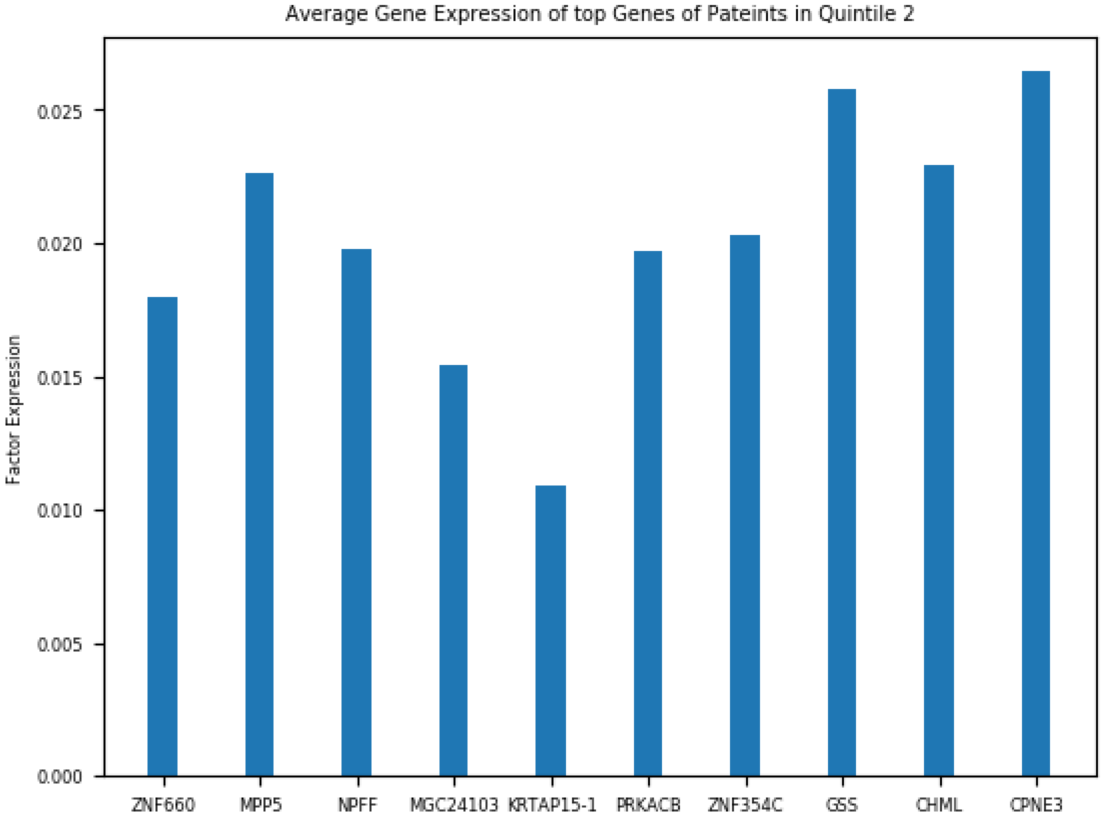
Below is a bargraph showing the top ten most important genes the model used for prediction and there weights

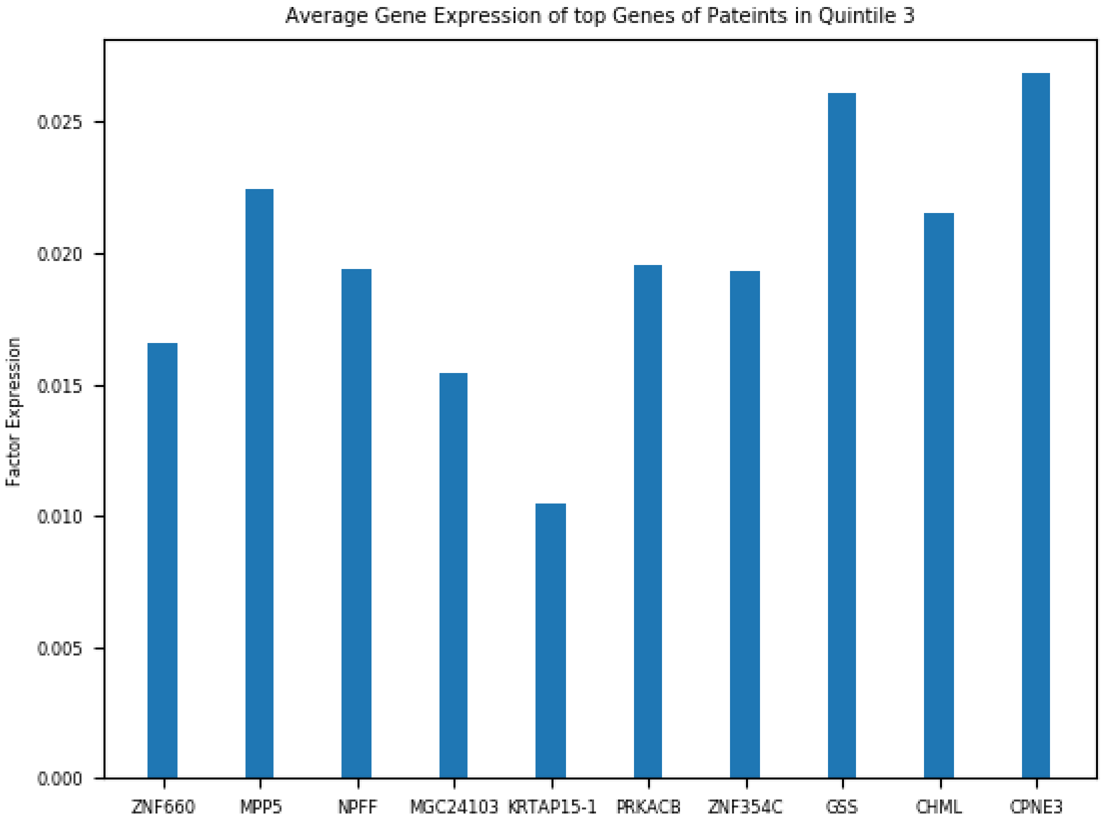


We tried to then further isolate the genes used for prediction by classifiying each quintile individualy. Every patient in the quintile was given a value of 1 and every patient not in the quintile was given a value of 0. The model was asked to predict the patients value. We ran into issues as each quintile has roughly 50 patients in them versus the 200 not in them. We tried methods to compensate for the skewed sample distribution however they began to break the model.

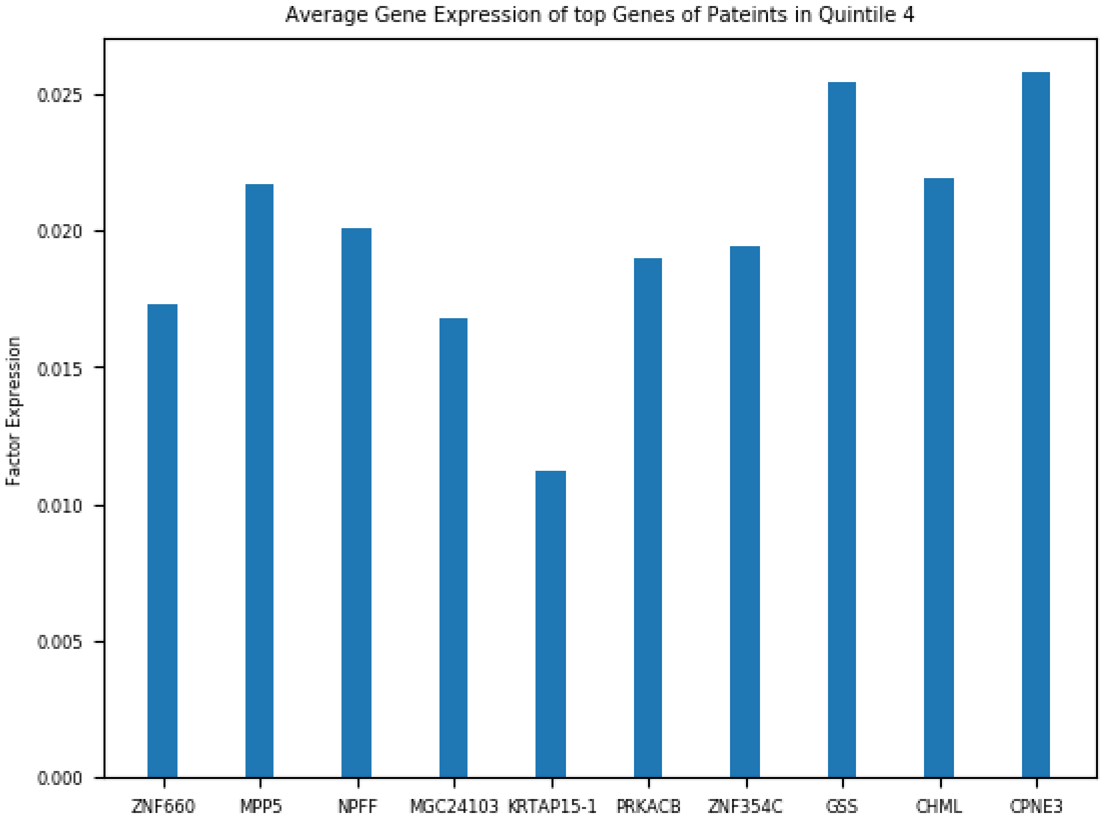
Instead we Looked at the gene Expression of the top ten genes in each patient when the patients were grouped by percentile they fall into. Below are graphs showing gene expression per quintile group.



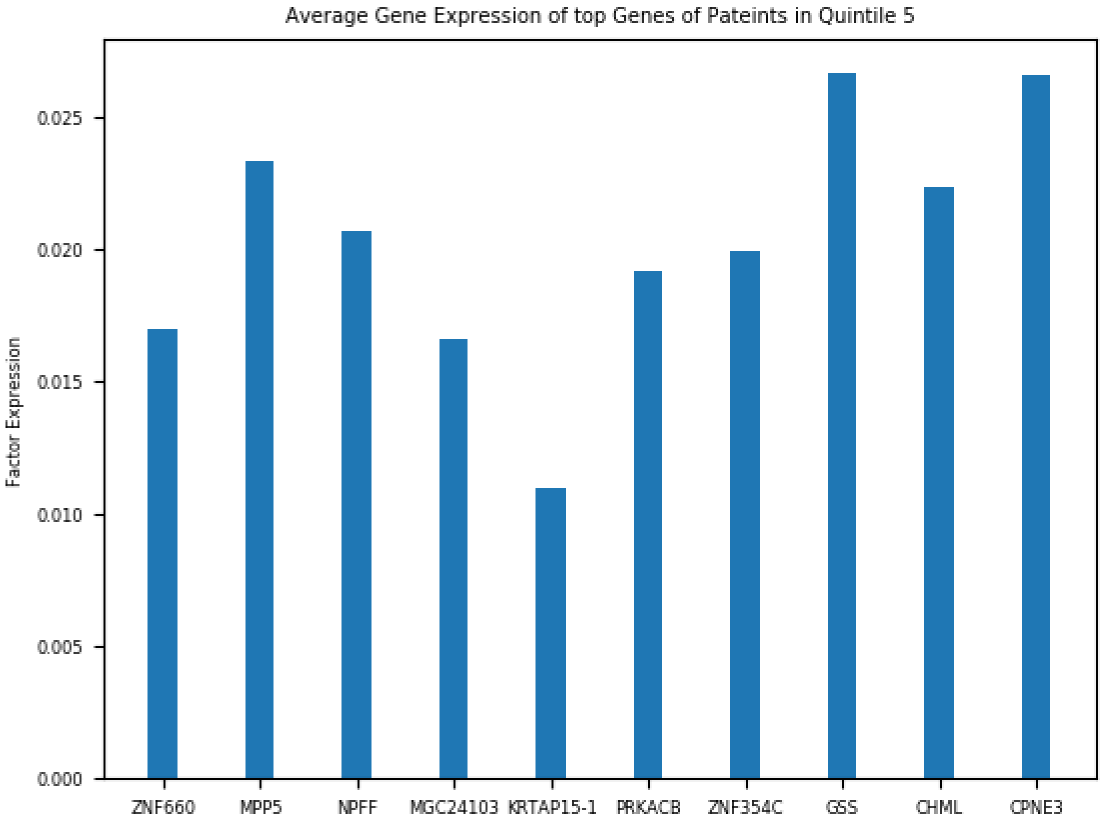




Average Gene Expression of top Genes of Pateints in Quintile 3



Average Gene Expression of top Genes of Pateints in Quintile 4



They all look very similar which leads us to believe that the best way to increase accuracy of prediction would be to get new variables. Methelation and Histone modification are two that readily come to mind. Using FireBrowse we were able to find a kindey cancer database of 891 usable patients. This database includes clinical data, gene expression, and methelation. We believe moving to this data set would be the best next step.