Identification of a Unique Set of Metastatic and Proliferation Genes Associated with Significantly Altered Lung Squamous Cell Carcinoma Patient Survival

Timothy Lee / Michael Jin

June 3, 2016

Abstract

Lung cancer is one of the most deadly human diseases in the world, accounting for more deaths than any other cancer. While 5-year survival rates for patients diagnosed at an early stage are relatively high at ~55%, only ~16% of lung cancer cases are diagnosed at a localizes stage (Molina 2008). The vast majority of patients are diagnosed after metastasis, at a stage when suvival is drastically reduced. This is particularly devastating for patients with lung squamous cell carcinoma, as current chemotherapeutic regimens designed to treat metastatic stage patients are highly primitive and incur a significant number of deleterious side effects. In our analysis, we screened for 76 genes whose expression correlates with patient survival. We followed up with an evaluation of possible molecular interactions with one of the most well characterized and researched oncogenes, KRAS, to investigate whether any of our mutations impact metastatic potential. We also analyze the impact of mutations in our candidate gene set on three proliferation markers shown to vary directly with proliferation potential. Our analyses yielded a set of 9 genes we believe significantly impact patient survival outcome through either modulation of metastasis or proliferation.

Introduction/Background

Lung cancer remains one of the most commonly diagnosed cancers in the United States, with over 220,000 new cases expected over the course of 2016. More striking, however, is that the American Cancer Society estimates over 150,000 patients currently living with lung cancer will die over the course of 2016. This number constitutes more than 25% of the predicted 595,000 total deaths resulting from cancer (Molina 2008). While efforts have been made to reduce these numbers, lung cancer remains difficult to treat unless diagnosed at an early stage.

Lung cancers are primarily distributed between two types according to morphological features: non-small cell lung cancer and small cell lung cancer. Non-small cell lung cancer accounts for over 80% of the lung cancer cases in the United States, representing the vast majority of patients currently living with lung cancer. Patients with non-small cell lung cancer carry a 5-year survival rate of around 15%, with many within that fraction developing new respiratory tract cancers, such as subsequent primary lung cancer (SPLC) (Molina 2008).

Non-small cell lung cancer are predominantly of the following two forms: adenocarcinoma (~50%) and squamous cell carcinoma (~30%). Adenocarcinoma usually arises in the periphery of the lung tissue while squamous cell carcinoma typically develops more centrally in the bronical tubes. Recent advances in our understanding of cancer biology pathways has led to the development of novel therapies targeting key proteins commonly mutated in non-small cell lung cancers, such as epidermal growth factor receptor (EGFR), anaplastic lymphoma kinase (ALK), and vascular endothelial growth factor receptor (VEGFR). Some of the most commonly prescribed compounds for use in

chemotherapy of metastatic non-small cell lung cancer cells include gefitinib, crizotinib, bevacizumab, which inhibit the activity of hyperactive EGFR, ALK, and VEGFR mutants respectively (Maemondo et al., 2010). In numerous clinical trials, these compounds have been shown to prolong overall survival in late stage lung cancer patients, with the addition of gefitinib to chemotherapy regimens doubling patients' median progression-free survival (10.8 months vs. 5.4 months in a control samples treated with standard chemotherapy).

However, none of these treatments have been approved for use in treatment of squamous cell carcinoma. While extensive genomic analysis has revealed a number of druggable targets, such as those mentioned above, in lung adenocarcinomas, no such targets currently exist for lung squamous cell carcinomas. Presently, chemotherapies for patients with squamous cell carcinomas include mostly cytotoxic agents, such as cisplatin, which have been shown to have extremely severe side effects, including high levels of nephrotoxicity and cardiotoxicity (Demkow et al., 2013).

Clearly, more advanced and effective therapies for squamous cell carcinoma are required in order to effectively treat patients without incurring significant side effects. Comprehensive genomic analysis of clinical tumor samples extracted from squamous cell carcinoma patients would provide additional insight regarding potential drug targets for future therapies. Combining extensive longitudinal analysis with an understanding of commonly mutated protein pathways, we hope to screen for genes that significantly alter patient longevity.

Subsequently, we hope to more deeply inspect our candidate genes, investigating whether any mutations in our candidate genes affects metastasis or proliferation. Literature suggests that high KRAS expression is associated with highly metastatic lung cancer strains while high MYBL2, BUB1, and PLK1 expression reflects increased proliferation (Reily et al., 2009; Whitfield et al., 2006). By determining which genes, whem mutated, significantly increased expression of our marker genes, we are able to narrow our initial screen results to a smaller list of genes we believe to affect metastasis and proliferation. We hope that this information will provide valuable knowledge regarding genetic loci that can be targeted to reduce cancer phenotypes in patients with highly metastatic lung squamous cell carcinoma.

Data Collection

The clinical data that we used came from the TCGA ("The Cancer Genome Atlas") data portal. The BioPortal for Cancer Genomics (www.cbioportal.org) provides mutation data and TCGA gene expression. The "TCGA2STAT" package in R provided simple TCGA data access for integrated analysis from these two sources. The particular disease we looked at, lung squamous cell carcinoma, has a disease accession code of "LUSC". This dataset provides 504 patient histories along with their age, gender, ethnicity, and smoking history. Expression data for 20501 genes from RNASeq is available 501 of these patients. The expression furthermore is log 2 normalized since it is given in the Level III format. Furthermore, 178 of these patients had their genomes sequenced for 13665 mutations.

Data Cleanup

Before performing downstream analysis, it was necessary to remove patients with inconsistent attributes from the dataset. For example, the "vital status" attribute, which is a binary variable stating whether the patient has passed away, did not have a corresponding "daystodeath" attribute for some patients. (The "daystodeath" attribute denotes number of days after the clinical study started.) Patients that supposedly had their last check-in with their doctor (the "daystolastfollowup" attribute) after their time of death also were removed. This helped provide a cleaner basis for longevity analysis.

To accurately perform longevity analysis of the patients, it was necessary to perform right censoring on the patient history. To reduce confounding from unobserved events in the censored data, such as loss to follow-up, drop-put, and study termination, we thresholded all the critical event times to 6 years after the study. Each patient has a 'time of last critical event' attribute, and if this time was after the 6 years, the 'time of last critical event' was changed to exactly 6 years after the study. Patients who died after the study terminated had their status changed to living, because it may have been possible that their death was not directly attributed to lung squamous cell carcinoma.

In addition, we only wanted to assess the genes where an insignificant number of patients had a mutation in that gene. If more than 10 patients had a mutation in a corresponding gene, then the gene was kept for downstream analysis. In the end, we ended up with 501 patients with proper right-censored survival data and their RNA expression levels for 539 genes, where there were more than 10 patients with mutations in each gene.

Aim 1: Survival Analysis

We performed a longevity analysis of the patient data, to see whether the RNA expression of each gene is correlated to the survival of a patient. Using the univariate Cox proportional hazards model, we used the wald test,likelihood ratio test, and log-rank test to determine whether each gene was correlated to survival.

The Cox proportional hazards model assumes that there is an underlying hazard function, describing the change of risk that the critical event occurs over time given baseline covariate levels (the covariates in this case are the gene expression levels). Under the proportional hazards condition, which states that covariate levels are multiplicatively related to the hazard, allowing the hazard to respond exponentially to each unit increase in the covariate value. The formula is as follows:

$$\lambda(t|X) = \lambda_0(t) * exp(\beta X)$$

where $\lambda(t|X)$ is the hazard at time t given the covariate X, $\lambda_0(t)$ is the baseline hazard function at time t, B is the coefficient that determines how to scale the baseline hazard given the covariate level.

After fitting the Cox model to the levels of gene expression for each patient and their survivals, we used a series of statistical tests to determine whether the expression of a gene is correlated.

Wald test

The Wald test is a parametric test that tests the true value of a parameter in a model that describes a relationship between data. We use the Wald test to test whether the B coefficient for the covariate is equal to 0 (null hypothesis), meaning that the covariate has no relationship with the hazard. The Wald test statistic is:

$$\frac{\hat{\beta}^2}{Var(\hat{\beta})}$$

where $\hat{\beta}$ describes the maximum likelihood estimate the the beta parameter.If the null hypothesis is true, then the statistic is assumed to follow a chi-squared distribution with 1 degree of freedom. Large chi-square values support the alternative hypothesis that the B is not zero. Incidentally, the parameter B has a direct relationship with the hazard ratio with respect to the baseline ratio.

Likelihood ratio test

The likelihood ratio test is another parametric test that tests the true value of a parameter in a model that describes a relationship between data. As in the wald test, We use the likelihood ratio test to test whether the B coefficient for the covariate is equal to 0 (null hypothesis), meaning that the covariate has no relationship with the hazard. In effect the test statistic is the log ratio between the likelihoods of the estimate with maximum likelihood given the sample and the likelihood that the true parameter is 0. The likelihood ratio test statistic is:

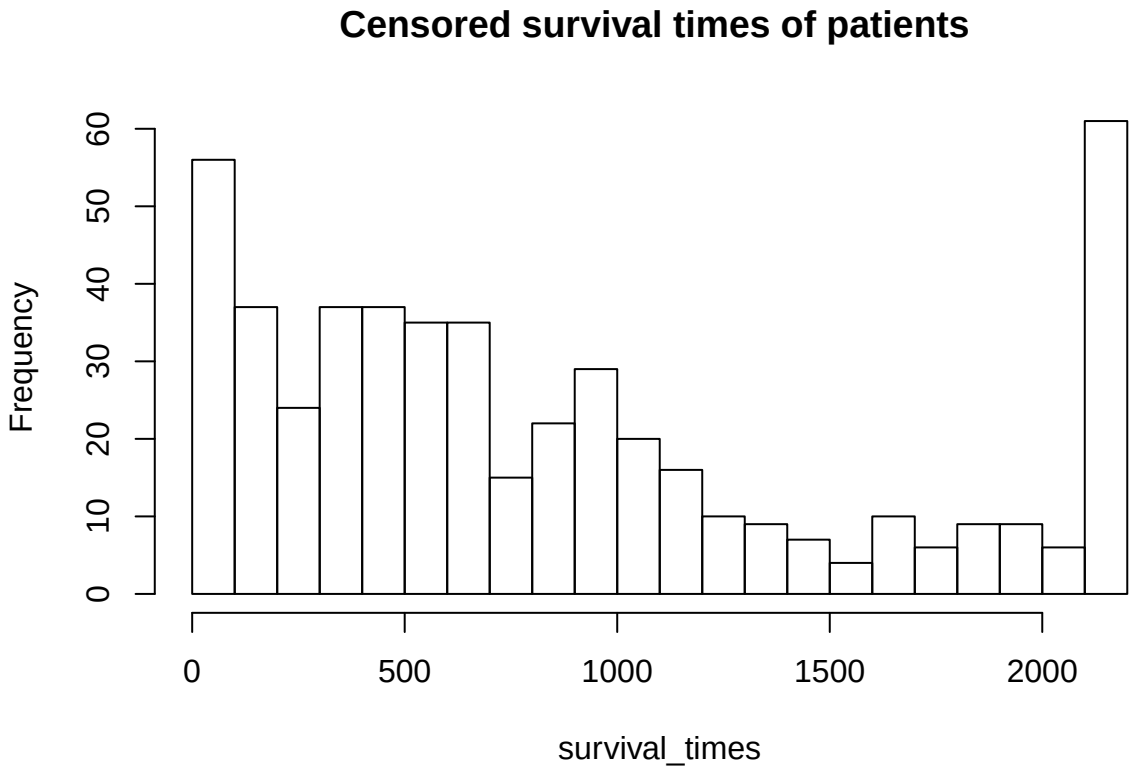
$$2[l(\hat{\beta}) - l(0)]$$

$l(\hat{\beta})$ = the log likelihood of the maximum likelihood estimate of the B parameter $l(0)$ = the log likelihood when B = 0

where if the null hypothesis is true, then the statistic is assumed to follow a chi-squared distribution with 1 degree of freedom. Large chi-square values support the alternative hypothesis that the B is not zero.

Log-Rank test

The log-rank test is a non-parametric test that compares the survival functions of two sets of data over time. It is appropriate to use when the survival times are right skewed and censored (as done previously). Here is a distribution of survival times to show that the survival times are right skewed. The large bar in the end at 2000 days is due to right-censoring of the survival times.



The null hypothesis of this test is that the Kaplan-Meier estimate of the two data sets is equal throughout the study. The Kaplan-Meier estimate is a nonparametric estimator of the survival data. When the two sets are compared, rather than depending on the critical event times themselves, this test depends on the ranks of the event times themselves.

The log-rank test statistic is computed from creating a contingency table at each point of failure, assuming that the number of deaths from one group follows a hypergeometric distribution conditional on the margins to get the number of expected deaths for that group. The statistic itself follows a chi-square distribution with one degree of freedom. Higher chi-square values support the alternative hypothesis that the Kaplan-Meier estimate is not equal for the two populations at all periods of time.

To split up the two different groups, we used the median of the level of gene expression as a prognostic indicator to separate the two patients into one group denoted as high expression and another as low expression. A 95% significance was used to determine significance.

With all three tests over the plethora of statistical hypotheses being tested, it was necessary to convert each p-value to a corresponding q-value to account for multiple hypothesis testing.

```
##  [1] "OR4C6"    "ABCA13"   "ZEB2"     "FLNC"     "ODZ4"     "VWF"
##  [7] "AKAP6"    "FLG2"     "CCDC141"  "ERBB4"    "ITPR2"    "PTPRB"
## [13] "NEB"      "ROS1"     "ADAMTS16" "FBN2"     "CADPS2"   "MYH1"
## [19] "RP1"      "NLRP5"    "C7"       "CPS1"     "ZAN"      "CD163"
## [25] "MYH2"     "LRRK2"    "XDH"      "DMBT1"    "EPHA6"    "ATP10A"
## [31] "MAGI2"    "ACSM2B"   "TRIM58"   "SLC17A8"  "C6"       "GRM3"
## [37] "CSMD3"    "PTEN"     "F5"       "OR5L2"    "KIRREL2"  "FN1"
## [43] "NAV3"     "FLT1"     "KLHL1"    "SYNE2"    "NLRP12"   "PEG3"
## [49] "ANK2"     "MYT1L"    "ADCY2"    "DPP6"     "LRRC4C"   "CACNA1C"
## [55] "TMPRSS15" "C15orf2"  "CNTNAP4"  "UNC13A"   "MED12L"   "MDGA2"
## [61] "TAF1"     "GPR112"   "FAM47A"   "NALCN"    "LRFN5"    "ZNF208"
## [67] "SLITRK5"  "ZNF560"   "GRID1"    "COL5A2"   "DLG2"     "STAB2"
## [73] "RALGAPA2" "LCT"      "SORCS3"   "EPHB1"    "SPHKAP"   "GRM5"
```

In the end, we were able to procure 76 genes that are deemed significant in changing the survival of patients. We will characterize the relationship of these 76 genes with a well-known cancer gene, KRAS, in the next aim. KRAS expression has been associated closely with highly metastatic lung cancer variants. We hope to demonstrate a relationship between a subset of our 76 candidates and an metastatic phenotype.

Aim 2: Finding genes linked to expression of KRAS

From these genes, we wanted to see which genes, either through expression levels or mutations, were correlated with KRAS, a notable cancer gene. This would suggest that possibly the genes are in the same pathway as KRAS.

Finding gene mutations that are linked to KRAS expression

In this section, we used a mix of parametric and non-parametric tests, as well as tests from literature, that test whether a mutation in a gene significantly alters KRAS expression.

- *Tests derived from literature*

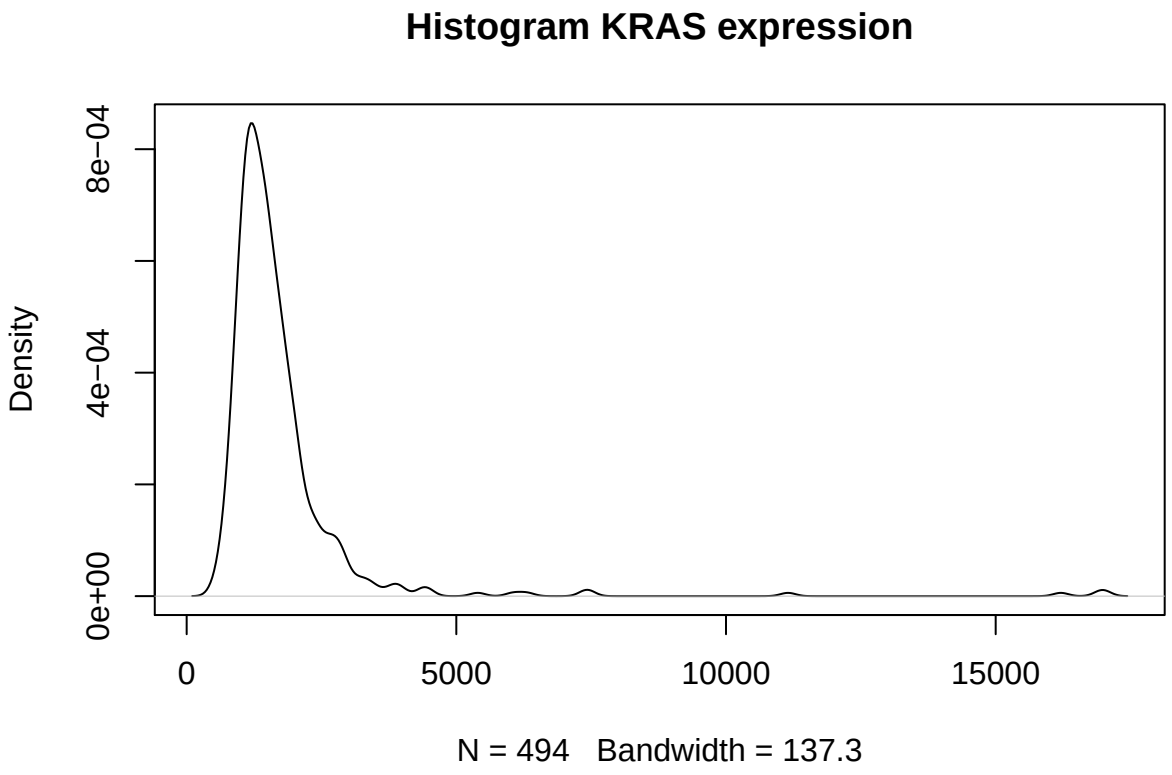
In Brodie et al, the paper uses the same dataset to analyze whether a lung cancer specific mutation is related to CHFR expression. CHFR expression is often used as powerful predictor for response to taxane, a drug used in first-line chemotherapy. They classified patients based on two binary

categorical variables: whether they had the mutation and whether they expressed high or low CHFR expression (threshold as determined by the median). Creating contigency tables for each mutation, they used the odds ratio and chi square tests to characterize the relationship between the two variables. We will begin by attempting the same approach as Brodie et al. using high/ low KRAS expression as our categorical variable (Brodie et al., 2015). The main criticism of this approach is that it loses valuable information about the level of gene expression among the patients, by treating it as a categorical variable. Furthermore, pairs of patients that have gene expression levels close to the median but happen to fall on different sides of the median, are carelessly treated as patients that have vastly different gene expression levels.

- *Mann-Whitney and t-test*

In an effort to keep the gene expression of KRAS as a continuous variable, we will use the two sample t-test and Mann-Whitney test, a nonparametric alternative to the t-test, to characterize this relationship. These approaches test whether the two samples, one with the mutation and another without the mutation, have are drawn from the same distribution. The two sample t-test assumes that the gene expression levels follow the normal distribution, but the Mann-Whitney test more simply assumes that the distribution is symmetric. The t-test with its stronger assumption holds a stronger test than Mann-Whitney. It is unclear whether the KRAS expression of patients is normally distributed for each mutation (after splitting up patients by whether they have the mutation or not.) Here is the distribution of KRAS expression levels and a line that defines the probability density function of the normal distribution.

```
plot(density(KRAS_expression), main = "Histogram KRAS expression")
```



The distribution is right-skewed, suggesting that the Mann-Whitney test may be more appropriate.

Taking those mutations that have a pvalue < 0.05 in either the Mann-Whitney or t-test, we can accept the alternative hypothesis that the samples with and without mutations have significantly different levels of KRAS expression.

```
##                mann        ttest
## ZNF560   0.01996939 0.0013106531
## TRIM58   0.02752699 0.0017289965
## KIRREL2  0.03418362 0.1731831669
## NLRP12   0.03583264 0.3800124660
## DMBT1    0.04343164 0.1416818857
## C7       0.04672369 0.2261754449
## RALGAPA2 0.04827905 0.0002032706
## FLNC     0.06995103 0.0040460189
## ANK2     0.07562179 0.0161707785
## MED12L   0.24000847 0.0009756122
## GRM3     0.66336078 0.0373916590
```

These are the genes that when mutated, will have a significant difference in the gene expression of KRAS.

```
##          no mut KRAS exp mean has mut KRAS exp mean
## ZNF560                1729.844               1204.001
## TRIM58                1733.841               1301.597
## KIRREL2               1676.991               1991.992
## NLRP12                1688.828               1828.181
## DMBT1                 1672.743               2055.330
## C7                    1676.315               2002.075
## RALGAPA2              1727.543               1279.084
## FLNC                  1730.168               1312.963
## ANK2                  1745.399               1418.233
## MED12L                1724.511               1351.365
## GRM3                  1716.874               1446.531
```

Finding correlations between KRAS expression and gene expression

In addition to correlating gene mutations to KRAS expression, we would like to characterize the relationship of the gene expression levels derived from the survival analysis of Aim 1 to the gene expression of KRAS. Using a mix of parametric and nonparametric measures to characterize the correlation coefficient, we can determine whether the level of gene expression and level of KRAS expression is dependent.

The correlation coefficient measurements we will use are the Pearson, Kendall Tk, and Spearman coefficient.

Pearson coefficient

The pearson coefficient is the tradition correlation coefficient that assumes that the best model/relationship between two variables is one that is exactly linear, since its measure is directly derived from the least squares linear regression model. The measurement is highly interpretable

and well known, with r^2 close to 1 or -1 meaning that it is close to being a linear fit, and r^2 close to 0 meaning that there is no noticeable relationship between the two. Given that it may not be certain that the relationship between the gene levels, if there is one, is linear, the Pearson coefficient has its disadvantages.

Th test statistic is based on Pearson's product moment correlation coefficient between the two variables and follows a t distribution with the number of samples - 2 degrees of freedom if the samples follow independent normal distributions.

Kendall's rank correlation

The Kendall Tk provides a distribution free measure of monotonicity between the two variables. Given a random sample of (X,Y) variables, it takes a random pairs of these variables and classifies them as concordant or discordant. Concordant pairs have the property that the pairs' X and Y variables are monotonically increasing, while discordant pairs have the X and Y variables in different directions. A hypothesis test on whether the Kendall rank correlation is 0 can be done on the exact distribution. While it is distribution-free, the Kendall correlation may underestimate the correlation between two variables because when it takes a pair of observations to determine concordancy, it makes no distinction of whether the X values are close together. When X values are close together, concordancy should not matter, because the Y values should be more close together than when the X values are farther apart. The Kendall correlation heavily penalizes for monotonicity's sake against random pairs of observations that are close together, that happen to be discordant due to variable error.

An exact p-value can be found by taking the finite values of the variables.

Spearman rank correlation

Like the Kendall correlation, the Spearman rank correlation is distribution-free. It orders the two variables and uses the rank of the variables to determine correlation. Rather than assuming a linear model, monotonicity instead is assumed. A test of independence could be done using the exact distribution. A disadvantage is that a p-value cannot be exactly quantified with the data has ties.

For each gene, we correlated the gene expression level with KRAS expression level through these 3 metrics, doing a hypothesis test of whether the correlation coefficient = 0, that the data is totally uncorrelated. We will use q values to account for multiple hypothesis testing.

```
head(significant_pearson)
```

```
##          estimate      p.value
## NLRP5   0.4155847            0
## CNTNAP4 0.2428081 4.629068e-08
## PEG3    0.2280624 2.994763e-07
## ITPR2   0.2184169 9.507748e-07
## CPS1    0.1429466  0.001445187
## MED12L  0.1331716  0.003020855
```

```
head(significant_spearman)
```

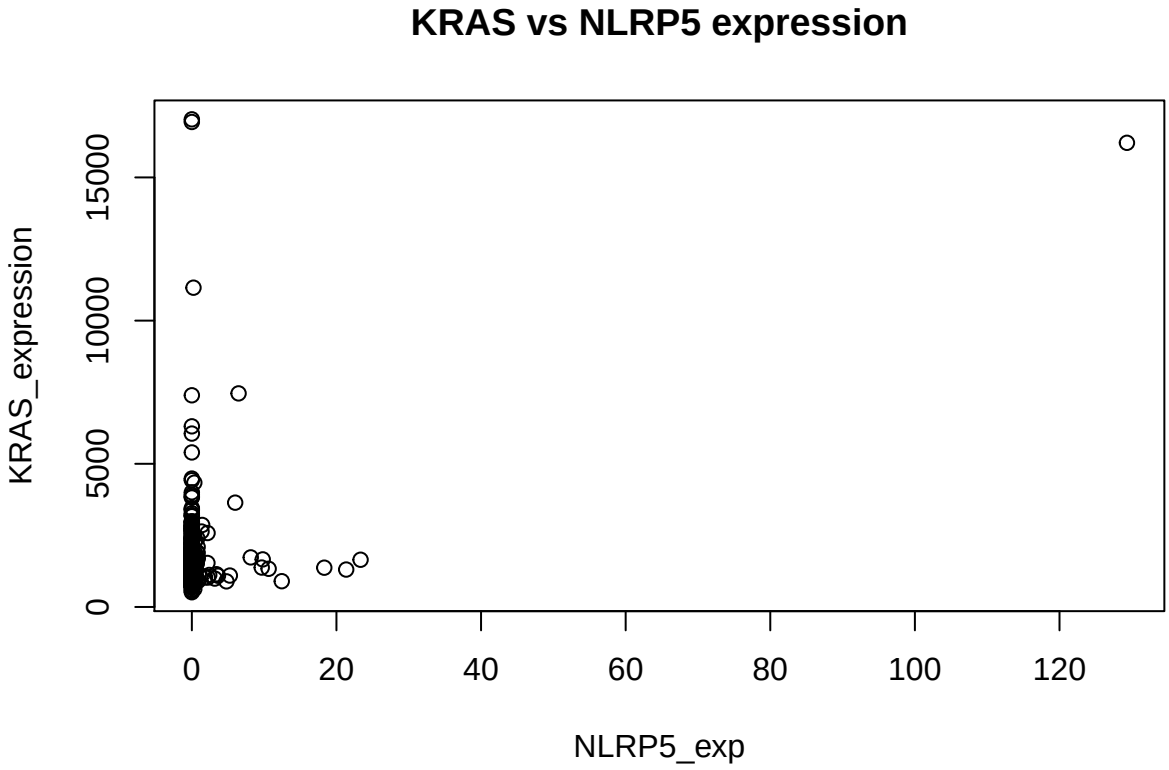
```
##          estimate      p.value
```

```
## ABCA13  0.2927022 4.022097e-11
## MED12L  0.2624293 3.646961e-09
## SORCS3   0.243267 4.359045e-08
## FBN2    0.2316827  2.07741e-07
## XDH    -0.2166066 1.248036e-06
## NLRP12 -0.2066666 3.624037e-06
```

```
head(significant_kendall)
```

```
##          estimate      p.value
## ABCA13  0.1940281 1.149196e-10
## MED12L  0.1763392 4.677885e-09
## SORCS3   0.175716 3.978058e-08
## FBN2    0.1550698 2.582154e-07
## XDH    -0.1462992 1.172249e-06
## NLRP12 -0.1375046 5.104869e-06
```

From the results above which show the correlation coefficient and corresponding p-value of the hypothesis test, it seems that some genes do in fact have a correlation in gene expression with the KRAS gene expression. There is not one particular gene that seems to have a exactly perfect relationship with KRAS expression, which is acceptable because we simplified the model to just one explanatory variable (the gene expression). The Kendall and Spearman coefficients have the same order but without the same p values and estimate, which is unsurpring since they are both nonparametric, but are different in how they calculate the estimate. There were 45 genes that were deemed as having a correlation. There were 19 genes deemed as significant by Pearson correlation coefficient, lower than the 45 genes in the other two metrics perhaps because the pearson correlation coefficient assumes linearity of the model, underestimating the true correlation if the bivariate data fit some other monotonic relationship. The seemingly well-correlated gene NLRP5 from the Pearson coefficient test is actually rather very uncorrelated, with many data points showing no NLRP5 expression. This shows that the Pearson coefficient test is not always the best measure of correlation.




```
exp_gene_significant
```

```
##  [1] "CNTNAP4" "PEG3"    "MED12L"  "ATP10A"  "ZEB2"    "CADPS2"  "PTEN"
##  [8] "XDH"     "COL5A2"  "PTPRB"   "C7"      "TAF1"    "FLT1"    "NEB"
## [15] "FBN2"    "NAV3"
```

There were 16 genes that passed all three significance tests. Some of these genes have been further characterized in literature in their relationship with KRAS expression.

We want to focus specifically on genes where the presence of mutations correlates directly with high KRAS gene expression. Given high KRAS expression is a hallmark trait of aggressive metastatic lung cancer, we would expect mutations of these genes to also drive metastatic potential (Reily et al., 2009).

We ended up with list of 4 genes that match this criteria: KIRREL2, NLRP12, DMBT1, and C7. Throrough literature analysis was completed to investigate any known phenotypes associated with mutations in these 4 genes.A brief summary of our findings is described below:

KIRREL2 (Kin of IRRE-like Protein 2) - Regulates basal insulin secretion. Specifically expressed in beta-pancreatic cells. Loss of KIRREL2 function is associated with dramatically increased basal insulin production (Yesildag et al., 2015).

NLRP12 (NLR Family, Pyrin Domain Containing 12) - Associated with suppression of colon inflammation and tumorigenesis. Given that NLRP12 is known to be inhibitory towards metastasis, it is likely that the mutations documented in our patient samples were loss of function mutations resulting in increased metastasis. (Allen et al., 2012)

DMBT1 (deleted in malignant brain tumors 1) - Demonstrated to be downregulated by in metastatic bladder carcinoma. DMBT1 expression inversely correlates with cancer stage. (Dodurga et al., 2011)

C7 (complement component 7) - Loss of C7 expression is associated with reduced immune function and inhibiton of the the immune complement system. (Würzner et al., 2003)

Analyzing the above findings, loss of function mutations in each of these seems likely to increase growth or even full dysregulation. Therefore, our findings that mutations in these genes are linked to increased metastatic potential seems reasonable.

Aim 3: Characterize relationship of candidate genes with known markers of cell proliferation/

To better understand the role of our candidate genes in promoting cancer phenotypes, we sought to investigate whether mutations of any of our genes was associated with statistically significant changes in proliferation. We used expression levels of three established genes, MYBL2, PLK1, and BUB1, as indicators for proliferation potential of our tumor samples. Previously published by Whitfield et al. in Nature Genetics, these three genes were found to be a reliable gene signature for evaluation of proliferation in numerous cancer types (Whitfield et al., 2006). By determining the genes, that, when mutated, significantly increase expression of our proliferation signature, we are able to identify genes from our initial list of 76 candidates that are specifically associated with proliferation.

```
##          no mut MYBL2 exp mean has mut MYBL2 exp mean
## FBN2                   863.3312              1056.9779
## NAV3                   886.0515               894.9320
## SLC17A8                879.3611               991.2835
## RALGAPA2               901.1296               704.9217
## ZNF560                 880.6426               992.5271
## CADPS2                 884.7402               921.4290
## GRM5                   871.8801              1102.2265
## AKAP6                  874.1914              1031.5035
## VWF                    891.6944               847.7360
## UNC13A                 890.9418               843.3063
## TMPRSS15               896.2852               759.3097
## F5                     900.8933               764.8260
```

```
##          no mut PLK1 exp mean has mut PLK1 exp mean
## AKAP6                 874.1914             1031.5035
## LRFN5                 861.0623             1082.8374
## RALGAPA2              901.1296              704.9217
## CADPS2                884.7402              921.4290
```

```
##          no mut BUB1 exp_mean has mut BUB1 exp_mean
## GRM5                  871.8801             1102.2265
## LRFN5                 861.0623             1082.8374
## FBN2                  863.3312             1056.9779
## NALCN                 864.0700             1139.4653
## MDGA2                 868.5625             1077.6092
```

```
## RALGAPA2            901.1296            704.9217
## NEB                 911.2225            758.6022
```

Our results indicate that 5 genes are associated with high expression levels of at least 2 our of 3 of our proliferation signature genes: AKAP6, CADP52, FBN2, GRM5, and LRFN5. Further characterization of the mechanisms behind these genes is required but we are reasonably confident mutations in these genes significantly increases proliferation in cancer cells.

Discussion

In our analysis, we were able to determine a unique list of 76 candidate genes whose expression levels significantly affect patient survival outcome. Further analysis to characterize the biological basis for this phenotypic variation was done by doing association tests with a number of marker genes. High KRAS expression has been shown to correlate with extremely aggressive lung cancers phenotypes (Reily et al., 2009). By determining a list of genes that, when mutated, correlate with high KRAS expression levels, we are able to narrow our candidate list to only include genes associated with extremely aggressive tumors. Similarly, we utilized a three gene signature previously found to reliably track proliferation potential to determine genes whose mutations are associated with rapidly proliferating cancer types (Whitfield et al., 2006). From this, we were able to narrow our list of 76 candidates down to a final set of 9 high potential genes with documented mutations associated with either increased metastatic aggression or increased proliferation.

Future Directions

Future directions include expanding our analyses' depth or breadth. To more deeply analyze the functions of our final set of genes, statistical analysis of microarray data would allow us to further explore expression levels of these genes in a variety of distinct genotypic settings. In vitro biochemical and biological assays such as in vitro kinase assays or immunoprecipitation analysis would allow us to evaluate both enzymatic kinetics and putative protein-protein or protein-ligand interactions. Additionally, creation of fluorescent fusion proteins for our genes of interest would allow us to investigate protein localization and compoartmentalization.

Additionally, we can expand our analyzes laterally by including other gene signatures associated with a variety of common cancer trait. The two traits we focused on in our analysis were metastatic aggression and proliferation. However, a large number of gene sigaures have been discovered corresponding to phenotypes such as cellular senescence, loss of cell polarity, and angiogenesis. Expanding the number of sigatures used to sort our candidates would allow us to allocate each of our initial 76 genes into broad functional buckets, providing insight into potential binding partners or mechanistic relationships.

We parsed through an extensive dataset documenting clinical samples extracted from lung squamous cell carcinoma patients, identifying a small set of genes which were statistically significant for our predicted covariates. Future characterization of these genes and the proteins they encode will likely reveal interesting biological interactions that could potentially be the target for future cancer therapies.