Age-Associated Biomarker Study  
TCGA Lung Cancer RNA-Seq Findings

Derek Chiu

October 12, 2017

# Executive Summary

TCGA Lung cancer data contained 495 clinically annotated cases with gene expression measured via RNA-Seq for 18950 gene targets.

881 gene targets (5%) showed an absolute fold change of 1.25 or larger and an adjusted p-value less than 0.01, demonstrating age-associated expression structure. (The METABRIC series yielded 4335 age-associated targets from 48803 probesets (9%) that showed an absolute fold change of 1.25 or larger and an adjusted p-value less than 0.01.)

# Details

Table 1: Data Sources

|  |  |
| --- | --- |
| Data | File |
| TCGA Lung clinical | data\_bcr\_clinical\_data\_patient.txt |
| RNA-Seq expression | data\_RNA\_Seq\_v2\_expression\_median.rds |
| METABRIC annotations | Annotation\_Illumina\_Human-WG-V3\_hg18\_V1.0.0\_Aug09.rds |
| Age-related genes | AgeRelated\_Accession\_Numbersv02.csv |
| EZH2-related genes | EZH2relatedGenesv08.csv |
| ER binding | ER\_binding\_gene\_hg19\_threshold\_1e-5.txt |

Data are annotated with HUGO gene names. Refseq IDs were not available.

18950 gene targets were available in RNA-Seq expression data.

Clinical data available included age at diagnosis for 495 cases.

Linear regression modeling was carried out using the base-2 logarithms of the RNA-Seq expression data as the dependent variable, and age at diagnosis as the independent variable.

Three regression lines were fitted for each gene target: one for age <= 60 years, one for age > 60 years, and one for all ages. Estimated slope of regression line for each of the three fits was recorded, along with the p-value for the test of regression line slope equal to zero (no correlation between age and expression). P-values were adjusted for multiple comparisons using the Benjamini-Hochberg method.

An association with age was declared for a gene target if at least one of the regression lines indicated an absolute fold change of 1.25 or greater over the age range for the regression fit, with an adjusted p-value less than 0.01. The fold change requirement is to ensure that some degree of biological relevance is present in the findings. With such a large data set (495 cases), trivially small changes in expression can yield a small p-value. These statistical artifacts are not generally useful in attempting to better understand the underlying biological mechanisms occurring in the data. Inclusion of a minimum absolute fold change criterion helps to ensure that statistically significant findings also have some degree of biological relevance.

There is insufficient evidence to claim a statistically significant association between ER binding and age-associated expression, as seen in the Fisher's Exact Test.

Table 2: ER binding and age associated gene counts for the TCGA Lung cancer data

|  |  |  |  |
| --- | --- | --- | --- |
| ER binding | Age association Age associated | Not age associated | Total |
| **ER binding** N Column(%) | 92 10.4427% | 1586 8.7775% | 1678 |
| **Not ER binding** N Column(%) | 789 89.5573% | 16483 91.2225% | 17272 |
| Total | 881 4.6491% | 18069 95.3509% | 18950 |

##   
## Fisher's Exact Test for Count Data  
##   
## data: tab  
## p-value = 0.08926  
## alternative hypothesis: true odds ratio is not equal to 1  
## 95 percent confidence interval:  
## 0.959734 1.515485  
## sample estimates:  
## odds ratio   
## 1.211822

Additional data on genes showing an ER binding affinity are available from a ChIP-Seq study (Carroll et al.).

Of the 18950 gene targets available, 1678 were identified as having an ER binding affinity from the ChIP-Seq findings (9%).

Of the 881 age-associated gene targets, 92 were in the ER binding set (10%) showing a degree of enrichment in the age-associated set.

# Figures

Representative age-associated gene target plots are shown below for different patterns of association. LOESS smoothing curves are shown on each graph since expression trends can be more complex than linear trends.

