Gene Expression in Lung Squamous Cell Carcinoma (LUSC)

Logan Correa

Introduction:

According to the National Cancer Institute, approximately 40.5% of men and women will be diagnosed with cancer at some point during their life (2018). Of these, lung and bronchus cancers account for nearly 12% of all cancer cases. Lung squamous cell carcinoma (LUSC), a subtype of non-small cell lung cancer, remains a significant global health concern with its high mortality rates.

Prior research, including studies by Barbar (2022), has emphasized the heterogeneous nature of LUSC, underscoring its diverse genetic and molecular profiles. While previous findings have identified key genes such as FLRT3, PPP2R2C, and MMP3, which have led to targeted therapies, challenges persist in predicting patient responses and understanding the factors contributing to disease aggressiveness (Ma et al., 2020).

Recent literature also underscores the importance of comprehensive analyses that integrate clinical, genomic, and environmental factors to enhance our understanding of LUSCs progression and treatment outcomes. This study aims to address these challenges by investigating how mutations in specific genes associated with LUSCs can serve as predictors of cancer development. It is hypothesized (I) that there will be a clear distinction between normal and tumor gene expression and (II) that mutations in specific genes linked to LUSCs can function as predictive indicators of cancer development.

Methods:

The clinical dataset for LUSCs was sourced from the Genomic Commons Data Portal TCGA-LUSC project, encompassing a cohort of 551 patients. Each patient's profile includes 56,907 distinct gene expression transcripts, normalized by TPM (transcripts per kilobase million). Notably, the dataset exhibits an imbalance, with 49 patients diagnosed as healthy (normal), while the remaining 502 patients are diagnosed with cancer (tumor).

The initial phase of this study involved conducting data transformation and scaling on the gene expression data. Firstly, logarithmic transformation is applied to normalize skewed distributions, equalize variances across genes, and improve interpretability by making fold changes more intuitive. Additionally, these preprocessing steps assist in meeting statistical assumptions such as normality and equal variances, thereby enhancing the effectiveness of subsequent analyses and machine learning algorithms.

Additionally, scaling the data further diminishes the influence of outliers, ensuring that individual genes do not unduly dominate the analysis solely based on their expression levels. Data was scaled using the R ‘scale()’ function to standardize numeric data. This standardization

involves transforming the data so that it has a mean of zero and a standard deviation of one. This is achieved by subtracting the mean of each variable from its values and then dividing by the standard deviation.

The dataset underwent filtering using a curated list of frequently mutated genes associated with lung cancer from the GCD Portal. This filtering process reduced the number of genes of interest from over 56,000 to approximately 700. Additionally, patients displaying gene expression identified as outliers (below Q1 - 4 x IQR and above Q3 + 4 x IQR) were excluded from the analysis.

A volcano plot was generated to depict gene expression log2 folds against log10 p-values, illustrating significant downregulation, upregulation, and nonsignificant changes in gene expression. Genes lacking significant regulation ($\alpha = 0.05$) were excluded from the analysis. Furthermore, a heatmap was created using a random sample of patients (25 normal, 25 tumor) to visualize gene expression regulation for the top 20 most statistically significant genes identified in the volcano plot.

Principal Component Analysis (PCA) was conducted on the dataset to diminish the dimensionality of the independent variables, which represent genes. A Logistic regression model was developed using the PCA components to analyze the association between the reduced set of gene expression variables and the diagnosis of normal lung tissue versus lung tumor tissue in cases of lung cancer. This approach facilitates the exploration of how the combined effects of principal components contribute to predicting whether a sample corresponds to normal lung tissue or lung tumor tissue based on gene expression profiles specific to lung cancer.

Results:

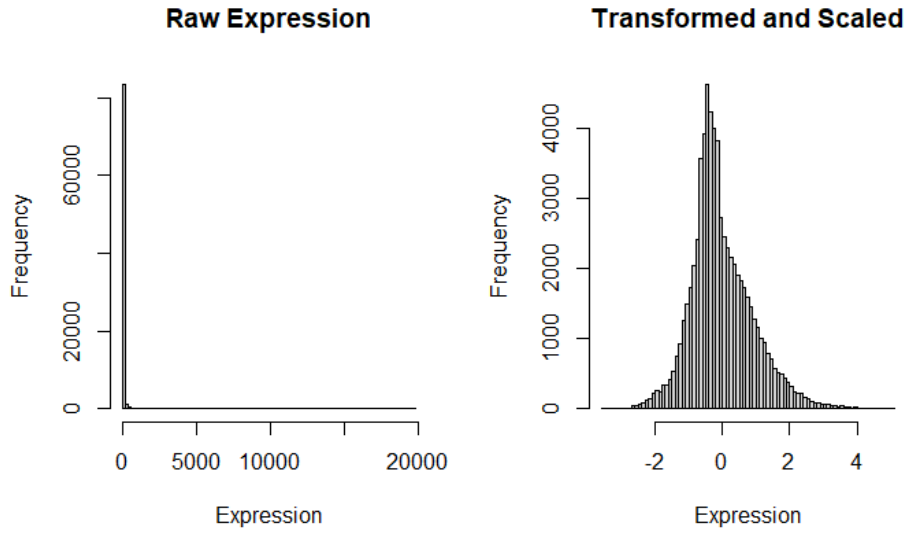

Figure 1. Histograms of raw gene expression (left) and transformed and scaled gene expression (right). The raw expression exhibits a pronounced right skewness and includes notable outliers. In contrast, the transformed and scaled expression demonstrates a normal distribution.

The raw gene expression data exhibits a mean of approximately 22.37 with a wide standard deviation of 137.15, while the scaled data centers around a mean of approximately -0.03 with a more restrained standard deviation of 0.92. Both datasets have similar medians, with the raw data ranging from 0 to 19729.97 and the scaled data ranging from -3.569067 to 4.834242. The raw data displays a skewness of approximately 44.21, indicating a highly skewed distribution, while the scaled data has a skewness of approximately 0.72. Similarly, the kurtosis of the raw data is approximately 5188.35, suggesting heavy-tailedness, whereas the scaled data has a kurtosis of approximately 1.29, indicating a less extreme tail behavior. Overall, scaling gene expression data standardizes the values, making them more comparable by removing differences in scale and distribution, while preserving the underlying statistical properties.

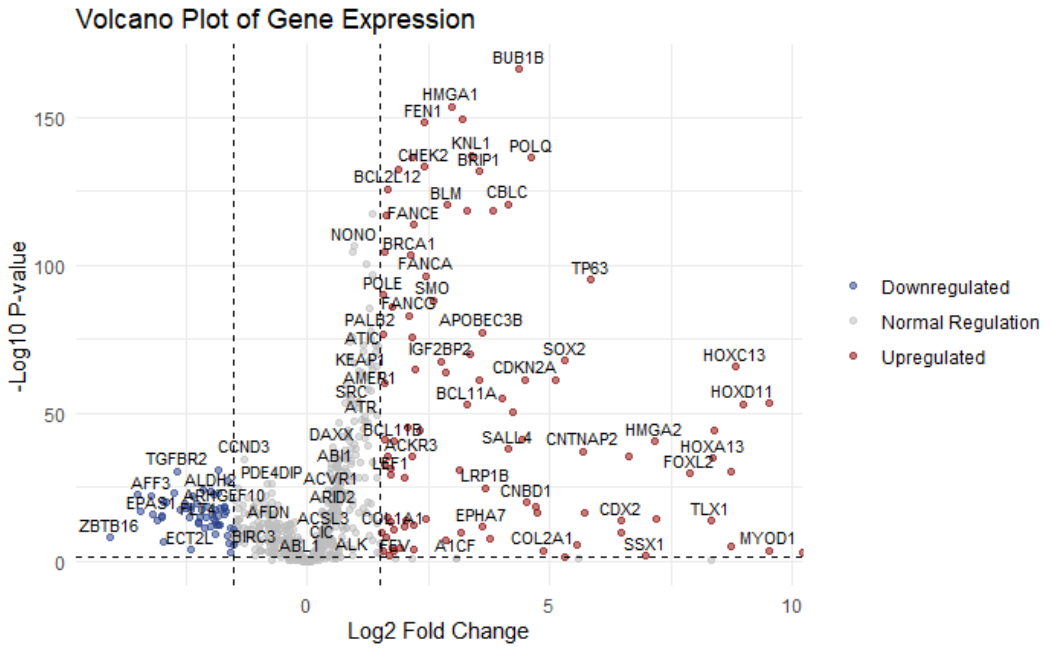

Figure 2. Volcano plot of gene expression. Downregulated genes are showcased in blue, upregulated genes in red, and normally regulated genes in gray. Dotted lines at Log2 Fold Change -1.5 and 1.5 delineate the boundaries for normal regulation.

Among the gene expression data, 7.95% of genes exhibit downregulation, 13.92% display upregulation, while the majority, accounting for 78.12%, show normal regulation. Notably, some of the most significantly regulated genes include BUB1B, HMGA1, STIL, FEN1, and KNL1.The dataset underwent further refinement using insights from the volcano plot, wherein genes with non-significant expression and regulation (p-value >0.05 and log2 fold change between -1.5 and 1.5) were removed.

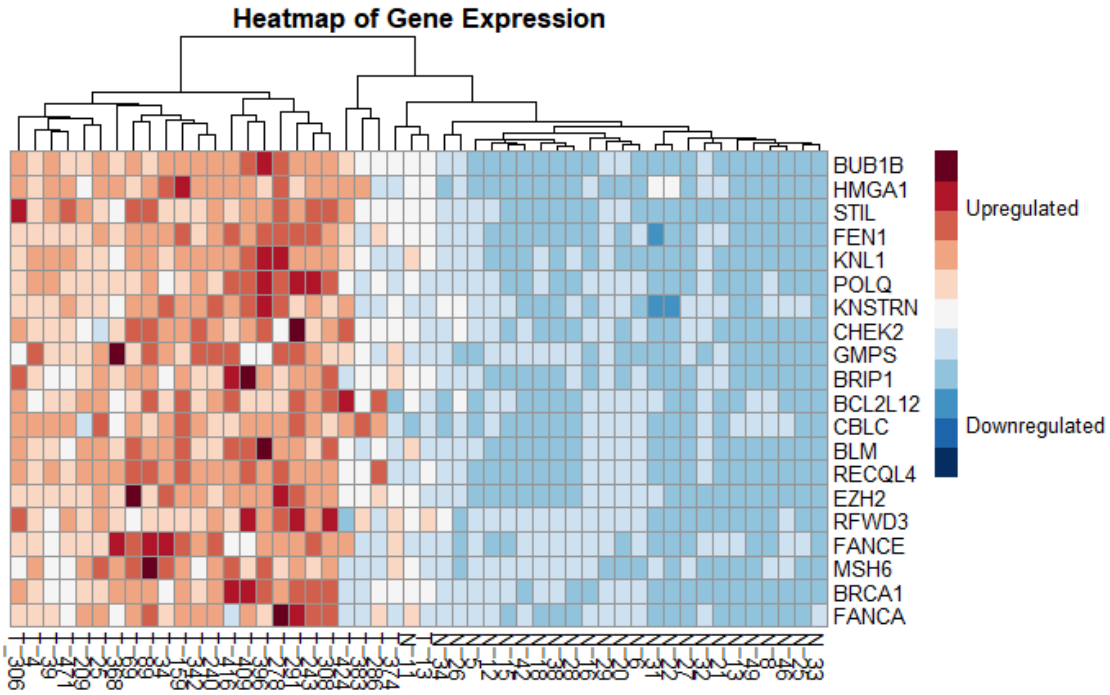

Figure 3. Heatmap of gene expression created using the R Pheatmap package. Gene expression is color-coded to reflect the degree of regulation, with downregulated genes appearing in varying shades of blue and upregulated genes in shades of red. Patient IDs are annotated with a prefix, "N" for normal patients and "T" for tumor patients.

The heatmap reveals a distinct separation between normal and tumor patients, indicating a predominance of upregulated gene expression in tumor samples for the most significantly regulated genes. Clustering of patients was automated by enabling the Pheatmap cluster_col parameter.

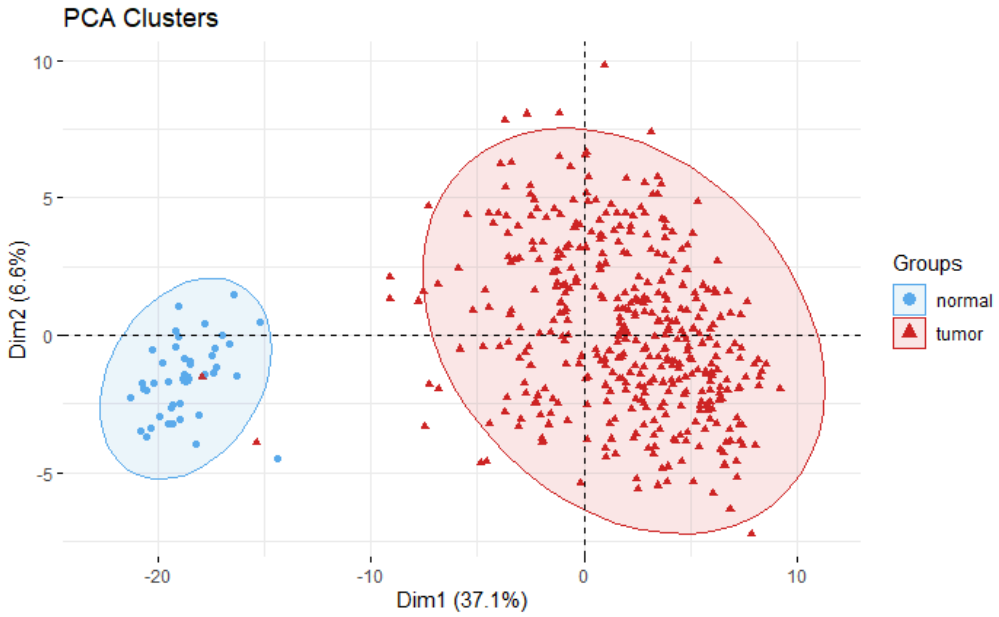

Figure 4. 2D PCA-plot showing clustering of “Normal” and “Tumor” diagnoses. Normal diagnoses are represented with blue circles and tumor diagnoses are represented with red triangles.

After collinear variables were identified and removed, PCA was performed on the scaled dataset. The PCA analysis reveals that the first principal component (PC1) explains a significant 37.1% of the dataset's variance, indicating it captures a major pattern distinguishing between different biological states. Although the second principal component (PC2) accounts for a smaller portion of the variance (6.6%), it still plays a crucial role in differentiating data points, likely reflecting variations not captured by PC1. The leading eigenvectors impacting PC1, which are BUB1B, HMGA1, KNSTRN, EZH2, and FEN1, correspond to the same genes previously identified as significant, demonstrating the effectiveness of PCA in highlighting these key genetic factors.

```
Model 1: Diagnosis ~ PC1
Model 2: Diagnosis ~ PC1 + PC2
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       440     13.557
2       439     13.191  1  0.36616   0.5451

Call:
glm(formula = Diagnosis ~ PC1, family = "binomial", data = pc_scores)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  11.7828     5.2101   2.262  0.02373 *
PC1           0.8377     0.3160   2.651  0.00803 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 268.460  on 441  degrees of freedom
Residual deviance:  13.557  on 440  degrees of freedom
AIC: 17.557

Number of Fisher Scoring iterations: 11

     PC1
2.310982
```

Figure 5. Results from ANOVA Test, binary logistic regression model, and odds ratio calculation.

Using the PCA components, several binary logistic regression models were created to forecast gene expression outcomes. The initial model incorporated PCA components PC1 and PC2 as independent variables, whereas the second model exclusively utilized PC1. To compare the models, an ANOVA test was employed with the likelihood ratio test (LRT) option. The resulting p-value was greater than 0.05 (pr(>Chi) = 0.5451). Consequently, we fail to reject the null hypothesis, which implies that a model employing just one PCA component (PC1) is preferable to a model that includes all PCA components. This suggests that the primary component alone captures sufficient information for predictive modeling.

Using the reduced model, the estimated coefficient for the intercept is 11.7828 with a standard error of 5.2101. The intercept represents the log odds of the outcome when PC1 equals zero. The z-value of 2.262 and a corresponding p-value of 0.02373 indicate that the intercept is significantly different from zero at the 5% significance level. The estimated coefficient for PC1 is 0.8377 with a standard error of 0.3160. The z-value is 2.651, with a p-value of 0.00803, which suggests that PC1 is a statistically significant predictor of 'Diagnosis' at the 1% significance level. The positive coefficient indicates that higher values of PC1 are associated with higher log odds of the outcome variable being “Tumor”. The substantial decrease from the null deviance to the residual deviance (from 268.460 to 13.557) further

suggests that PC1 has a strong effect in predicting 'Diagnosis'. The odds ratio for PC1 is given by exp(coefficient), which is exp(0.8377) = 2.310982. This means that for a one-unit increase in gene expression for PC1, the odds of being diagnosed with LUSC are about 2.31 times higher.

Conclusion:

In conclusion, this analysis provides compelling evidence that certain gene expressions are critical in distinguishing between normal and lung squamous cell carcinoma tumor tissues. The application of PCA yielded a robust principal component (PC1) that explains a substantial portion of the variance in gene expression profiles, and this component's significance is reinforced by the logistic regression analysis. PC1 alone has demonstrated its predictive power, simplifying the model without sacrificing accuracy, as supported by the ANOVA test results. The findings support the hypothesis that mutations in specific genes are linked to cancer development, with genes such as BUB1B, HMGA1, KNSTRN, EZH2, and FEN1 standing out in their association with LUSC. The odds ratio derived from the logistic regression model quantifies this risk, indicating a more than twofold increase in the likelihood of LUSC diagnosis with a unit increase in PC1 gene expression. This study advances our understanding of LUSC's genetic components and underscores the utility of integrated analyses for predicting disease outcomes.

References

Barbar, J., Armach, M., Hodroj, M. H., Assi, S., El Nakib, C., Chamseddine, N., & Assi, H. I. (2022). Emerging genetic biomarkers in lung adenocarcinoma. SAGE open medicine, 10, 20503121221132352. https://doi.org/10.1177/20503121221132352

Creating heatmaps in R. (n.d.). https://igordot.github.io/tutorials/heatmaps-2017-07.nb.html

Ma, X., Ren, H., Peng, R., Li, Y., & Ming, L. (2020). Identification of key genes associated with progression and prognosis for lung squamous cell carcinoma. PeerJ, 8, e9086. https://doi.org/10.7717/peerj.9086

National Cancer Institute. (2018). Cancer of Any Site - Cancer Stat Facts. SEER. https://seer.cancer.gov/statfacts/html/all.html

Genomics Commons Data Portal. Portal.gdc.cancer.gov. Retrieved February 5, 2024, from https://portal.gdc.cancer.gov/projects/TCGA-LUSC

ÖZÇELIk, B. (2024, March 4). TCGA - LUSC: Lung cancer gene expression dataset. Kaggle. https://www.kaggle.com/datasets/noepinefrin/tcga-lusc-lung-cell-squamous-carcinoma-gene-exp?resource=download

Yang P. (2009). Epidemiology of lung cancer prognosis: quantity and quality of life. Methods in molecular biology (Clifton, N.J.), 471, 469–486. https://doi.org/10.1007/978-1-59745-416-2_24

Yang, S., Yu, X., Fan, Y., Shi, X., & Jin, Y. (2018). Clinicopathologic characteristics and survival outcome in patients with advanced lung adenocarcinoma and KRAS mutation. Journal of Cancer, 9(16), 2930–2937. https://doi.org/10.7150/jca.24425