Evaluation of the Multi-Omic Profile of Smoking and Non-Smoking Patients with Lung Adenocarcinoma

Joshua Lin, Ethen Chen, Changshen Chen

Department of Quantitative and Computational Biology, University of Southern California

QBIO490x: Multi-Omic Data Analysis

Wade Boohar and Mahija Mogalipuvvu

December 6, 2024

Introduction

Lung adenocarcinoma, the most common subtype of non-small cell lung cancer (NSCLC), represents a significant global health burden and is responsible for a substantial proportion of cancer-related morbidity and mortality worldwide (*Lung Adenocarcinoma Study - NCI*, 2018; The Cancer Genome Atlas Research Network, 2014). Lung adenocarcinoma has been closely associated with smoking with numerous studies identifying tobacco-related carcinogens as a primary driver of mutagenesis and tumorigenesis in this disease (Yang, P et al, 2002). Individuals with a history of smoking are 15–30 times more likely to develop lung cancer compared with those who do not smoke (Hu, Yunqian, and Guohan Chen, 2015). However, according to The Cancer Genome Atlas Research Network (2014), an increasing number of cases are being diagnosed in individuals who have never smoked. These non-smoking yet heavily exposed individuals constitute a distinct and underexplored subgroup of lung adenocarcinoma patients.

Emerging evidence suggests that the molecular pathogenesis of lung adenocarcinoma in non-smoking patients differs significantly from that of heavily exposed patients. According to Govindan et al. (2012), there have been discovered differences between smoking and never-smoking patients, in terms of differences in point mutation frequencies (10x), mutation spectrums, and mutation sets, along with fusion genes and structural variants. Furthermore, somatic mutations, RNA, and miRNA differences have been discovered between smoking and non-smoking patients (Sui et al., 2020). However, prognostic trends in the context of these discovered differences between smoking and never-smoking patients with lung adenocarcinoma should be further studied. Understanding the heterogeneity of somatic mutations, gene expression profiles, and epigenetic alterations is critical for elucidating the biological

mechanisms driving cancer development in this subgroup, improving diagnostic precision, and identifying novel therapeutic targets tailored to their specific molecular and environmental risk factors.

Overall, this paper aims to provide a comprehensive analysis of the molecular characteristics and clinical outcomes of non-smoking and patients with heavy exposure to smoking. Utilizing the TCGA database, we integrate multi-omics data, including gene mutation landscapes, differential gene expression profiles, and epigenetic modifications. Using statistical and bioinformatics tools in R, we conduct in-depth analyses to identify unique biomarkers and molecular pathways that characterize this distinct patient subgroup. Additionally, we list our interest genes and evaluate their potential to potentially be used to predict clinical outcomes and guide personalized treatment strategies through further investigation and study in the future.

Methods

Using the clinical data from Genomic Data Commons (GDC)'s TCGA-LUAD database, we sorted patients into five groups - "Non-smoker," "Minimal Exposure," "Mild Exposure," "Moderate Exposure," and "Heavy Exposure" based on the tobacco smoking pack years smoked data frame. A pack-year is defined as smoking one pack of cigarettes (20 cigarettes) per day for one year (US Preventive Services Task Force et al., 2021). This indicator provides a quantitative measure of cumulative smoking exposure, allowing us to assess the relationship between smoking intensity, duration, and the development or progression of lung cancer better. Patients in the Non-smoker category either have a smoking pack years value of 0 or have a smoking history indicator of 1 and "not available" for smoking pack years. The "Minimal Exposure" group has 1-5 smoking pack years, "Mild Exposure" has 6-10 smoking pack years, and "Moderate

Exposure" has 11-20 smoking pack years. Finally, patients in the "Heavy Exposure" group have more than 20 smoking pack years. Our grouping criteria were defined using guidelines from the US Preventive Services Task Force, as outlined in their recommendation statement (US Preventive Services Task Force et al., 2021). Based on the five groups of patients, we then conducted a Kaplan-Meier survival analysis on those four groups of patients based on their survival time and potential death events.

Lung Adenocarcinoma MAF data were accessed from TCGA with the accession code "TCGA-LUAD". The R package maftools were used to analyze the MAF data. First, a co-oncoplot was generated to find out the top 10 most mutated genes between the two patient groups of interest - "Non-smoker" and "Heavy Exposure." Then we did a Contingency table using a mosaic plot and Fisher's Exact Test on TP53 - one of the highest mutated genes determined in the co-oncoplot. Later, we generated a co-lollipop plot for TP53, comparing the two patient groups of interest. Lastly, a MafSurvival KM plot was generated based on the mutations of the TP53 gene among the two patient groups.

Moreover, rna_se and methylation450 data were accessed from TCGA with the accession code "TCGA-LUAD" to analyze the transcription and methylation data for Lung Adenocarcinoma. First, the clinical data frame linked with RNA data was adapted to the clinic data frame linked with the general clinical data since we used the column that was only present in the general clinical but not present in the RNA clinical data frame. RNA genes data frame was then generated from rna_se, showing the gene information for mutation. Following that, the rna counts data frame was generated using the Tumor_sample_barcode column from the adapted RNA clinical data frame as column names and the gene id column from the RNA genes data frame. The RNA counts data frameworks as a correlation between patients and genetic

mutations. For data processing, all patients with NA values either in the smoking category, vital status, race, and gender were filtered out, and those patients who were not in the non-smoker and heavy exposure groups were also filtered out. The genes with a total mutation count sum of less than 20 were filtered out and the counts data frame was filtered based on the filtered clinical and genes data frame. From there, we conducted a DESeq2 analysis using preprocessed clinical, genes, and counts data frame, focusing on the difference in expression between the non-smoker and heavy exposure patient groups. We used the cleaned smoking category as variates and vital status, gender, and race as covariates. A results data frame was created to store the data generated from the DESeq2 analysis and an enhanced volcano plot was generated to show the results from the DESeq2 analysis, essentially showing which genes are either upregulated or downregulated comparing the heavy exposure group to the non-smoker group.

The clinical data frame for methylation, beta, and CpG sites was generated from the methylation450 data, those methylation that are not interested in this study are filtered out from the betas and CpG sites data frame. The data frames are then preprocessed to fit the general clinic data frame for reasons stated in the former paragraph. After that, a differential methylation analysis was conducted and a ggplot was generated to show the different methylation levels for the heavy exposure and non-smoker groups. After that, both expression and methylation were taken into account and four groups of interested genes were generated based on the intersected genes between the four groups: upregulated(log2foldchange > 1.25), downregulated(log2foldchange < -1.25), hypermethylated(log2foldchange > 0.75), and hypomethylated(log2foldchange < -0.75). Later, .txt files including the hyper- or hypo-methylation of CpG sites of different interest genes were generated; those txt files were submitted to the UCSC genome browser to analyze potential overlap between the CpG sites and

promoter region. A note here is that for expression, rna_tpm is used other than rna_counts this time. This is because the TPM data are normalized so that we can see the true difference in expression, but the counts' data were not. We still processed the tpm data frame the same way we processed the counts' data frame so that our analysis stays constant.

Results

First, the TCGA lung adenocarcinoma patients were divided into groups based on available smoking history (smoking pack years), with a table showing each group's sample sizes below.

Table 1. Smoking Group Table illustrating each group's sample sizes (number of patients) and the defined smoking pack year bins for each group

Group	Smoking Pack Years	Number of Patients
Non-Smoker	0	70
Minimal Exposure	1-5	12
Mild Exposure	6-10	15
Moderate Exposure	11-20	56
Heavy Exposure	> 20	260

A Kaplan-Meier plot was generated to compare the differences in survival probability based on patient's smoking pack years between smoking categories (Non-smoker, Minimal Exposure, Mild Exposure, Moderate Exposure, and Heavy Exposure):

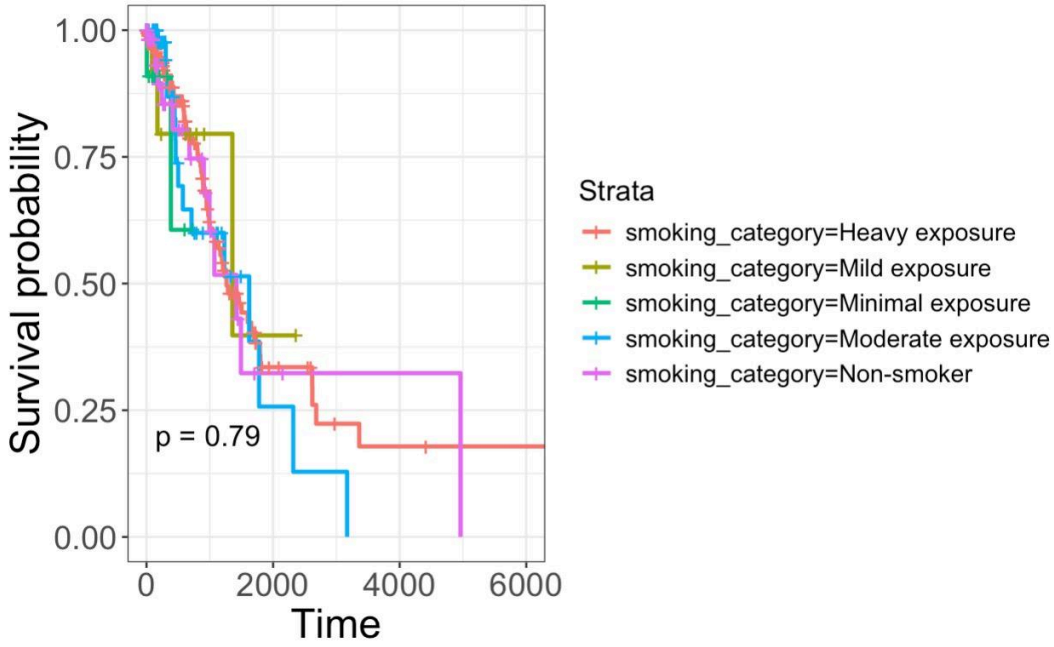

Figure 1. Kaplan-Meier Survival plot showing the different survival probability between the five patient groups. The P-value of 0.79 denotes the statistical insignificance of the differences in survival probability between strata.

Since the p-value is 0.79 > 0.05, the differences in survival probability between defined smoking strata were shown to be non-significant. This is primarily due to the heavily skewed number of patients toward the heavy exposure group, which reduces the statistical power of detecting differences across all groups. To address this, focusing on the comparison between the most distinct groups—non-smokers (0 smoking pack-years) and heavy smokers (smoking pack-years > 20)—can provide a clearer understanding of the extremes of smoking exposure. This targeted analysis avoids dilution caused by intermediate strata, enhances statistical power, and offers clinically relevant insights by emphasizing the risks of heavy smoking relative to no exposure. It also allows for a more precise hypothesis test, reducing noise and variability that may obscure meaningful findings in broader group comparisons.

A co-oncoplot was generated to investigate genomic differences between such two interest groups. The top 10 most mutated genes for the Non-smoker and Heavy exposure groups are shown below.

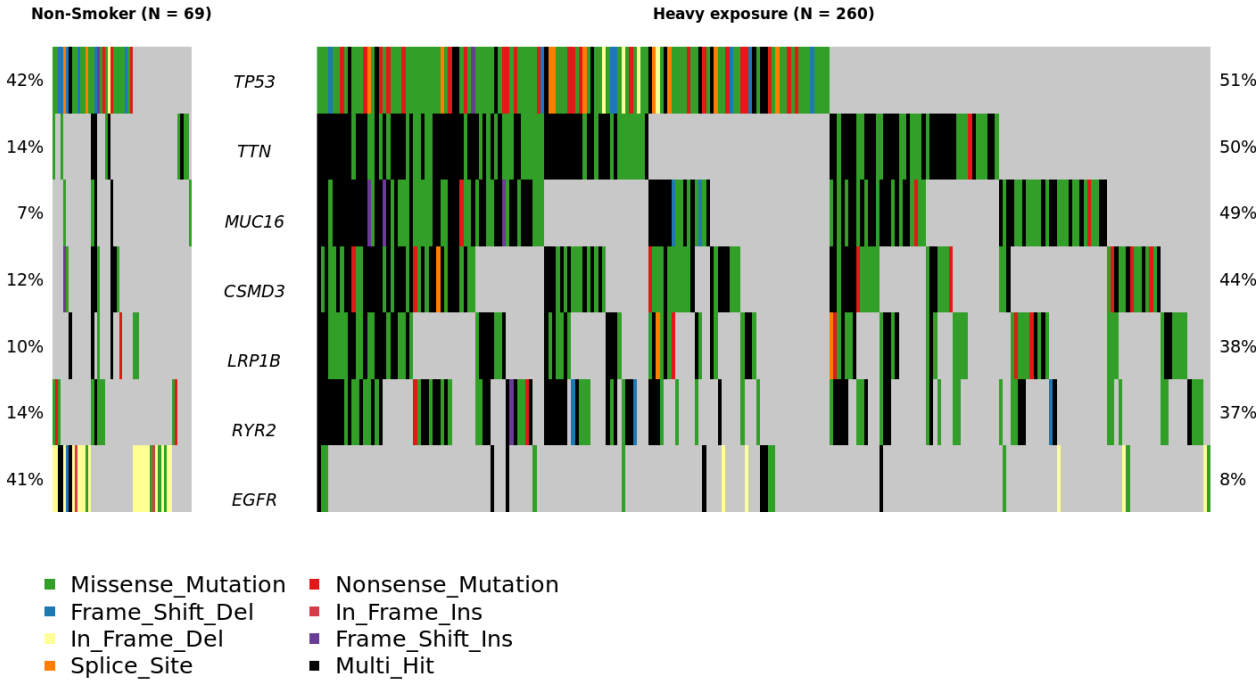

Figure 2. Co-Oncoplot showing the difference in mutation between the Non-smoker group and the Heavy Exposure group among the top 7 most mutated genes.

Although the 2 groups have different sample sizes, TP53 was determined to have the highest number of mutations between the two groups. Furthermore, the following mosaic plot was generated to visualize a contingency table between the two groups and the TP53 gene.

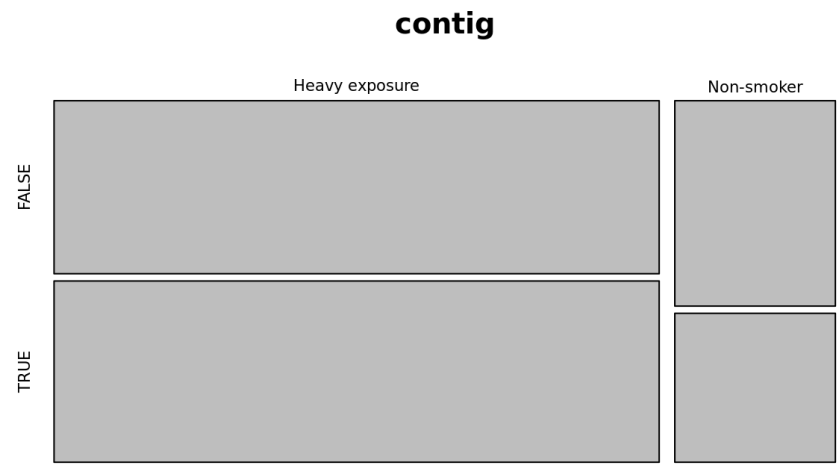

Figure 3. Mosaic Plot depicting a contingency table between the Heavy Exposure and Non-smoker group for the TP53 gene.

After performing Fisher's exact test on this contingency table, the p-value was 0.2227 > 0.05, so the alternative hypothesis that the true odds ratio is not equal to 1 is determined to be false. Moreover, the following co-lollipop plot was generated for the two populations and the TP53 selected gene.

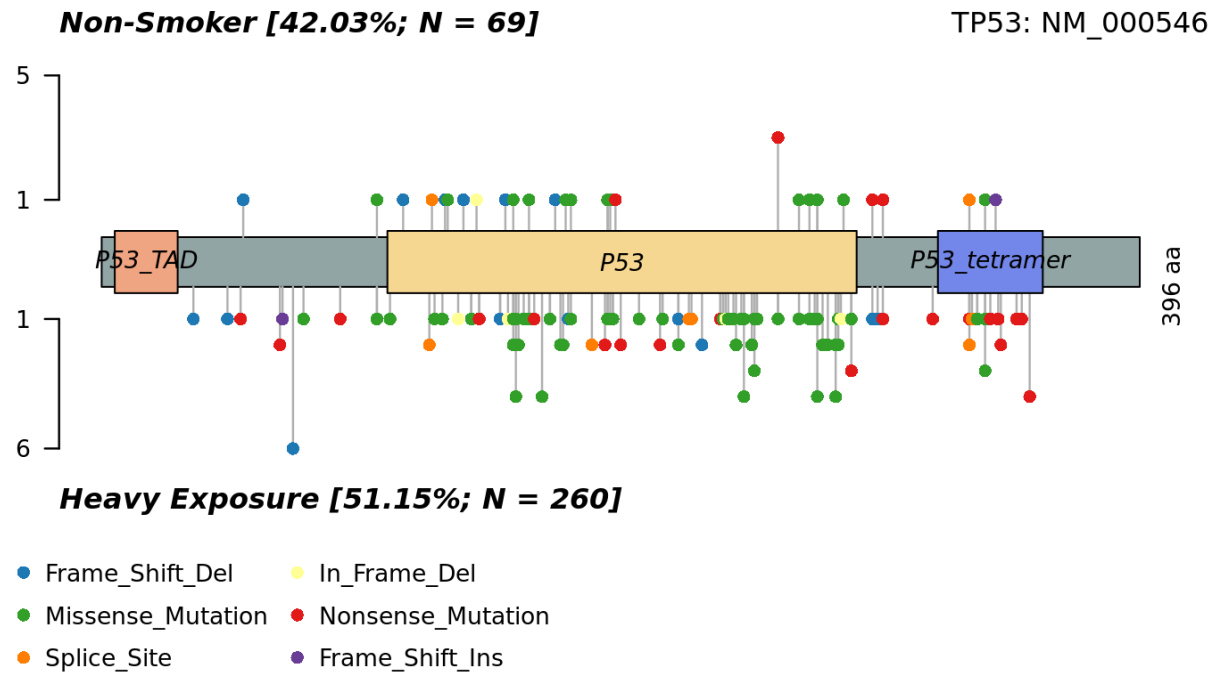

Figure 4. Co-lollipop Plot shows the different types and sites of mutation between the Heavy Exposure and Non-smoker groups on TP53.

Finally, the following mafSurvival KM Plot was generated to investigate differences between non-mutated (wild-type (WT)) and mutated (mutant) groups regarding the TP53 gene.

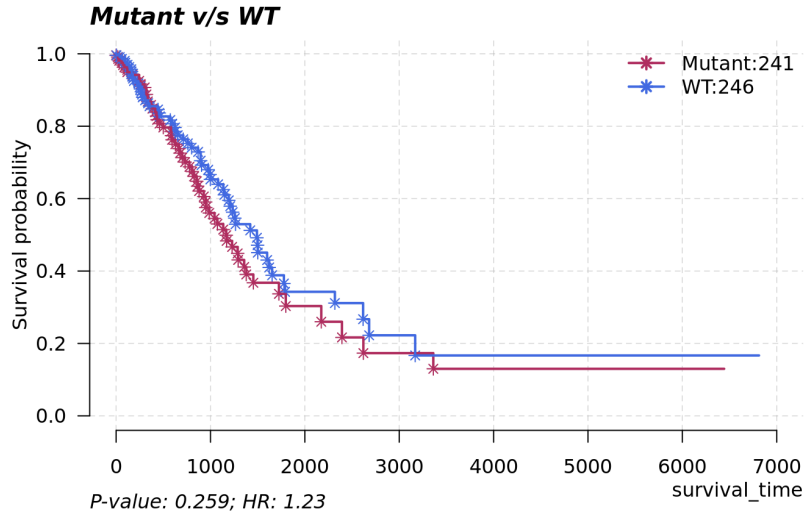

Figure 5. mafSurvival KM Plot showing that the mutant and WT group for the TP53 gene does not have a huge difference in survival probability. The P-value of 0.259 denotes the statistical insignificance of the found results.

With a p-value of 0.259 > 0.05, there is insufficient evidence to determine significant differences in survival between mutant and wild-type groups for the TP53 gene in TCGA lung adenocarcinoma patients.

For transcription data, DESeq2 analysis was conducted and an enhanced Volcano graph was generated.

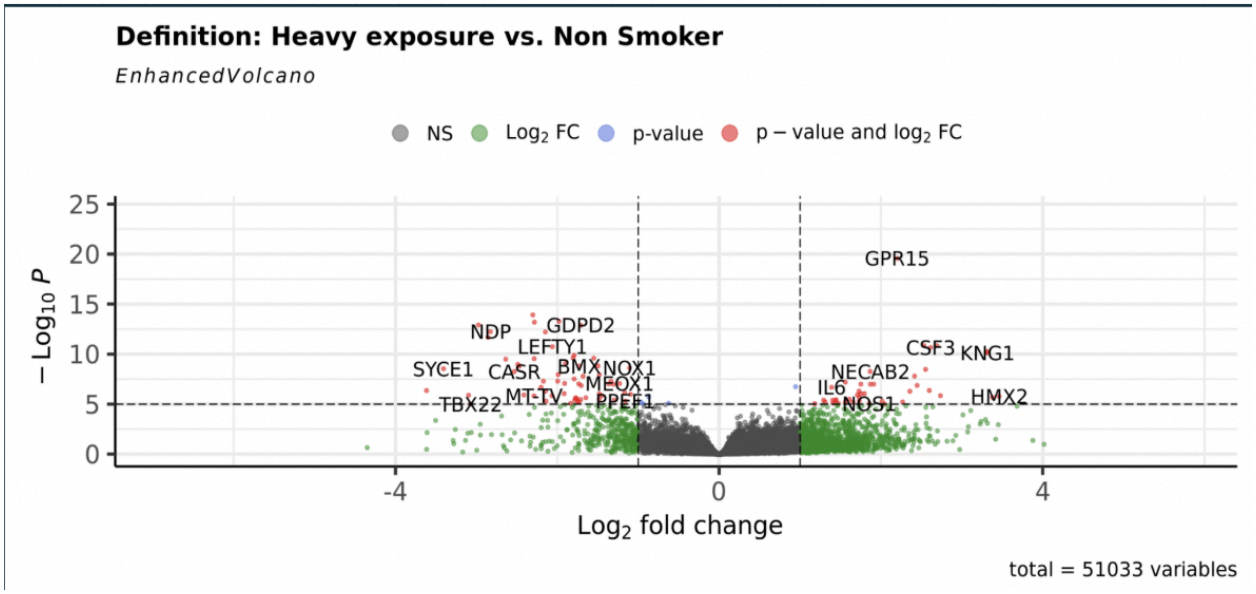

Figure 5. Enhanced Volcano Plot based on the differential expression analysis.

Genes such as HMX2 and GPR15 in the upper right corner (red dots) are significantly upregulated in the Heavy Exposure Group compared to the Non-smoker group, and genes such as NDP and SYCE1 in the upper left corner (red dots) are significantly downregulated in the Heavy Exposure Group compared to the Non-smoker group.

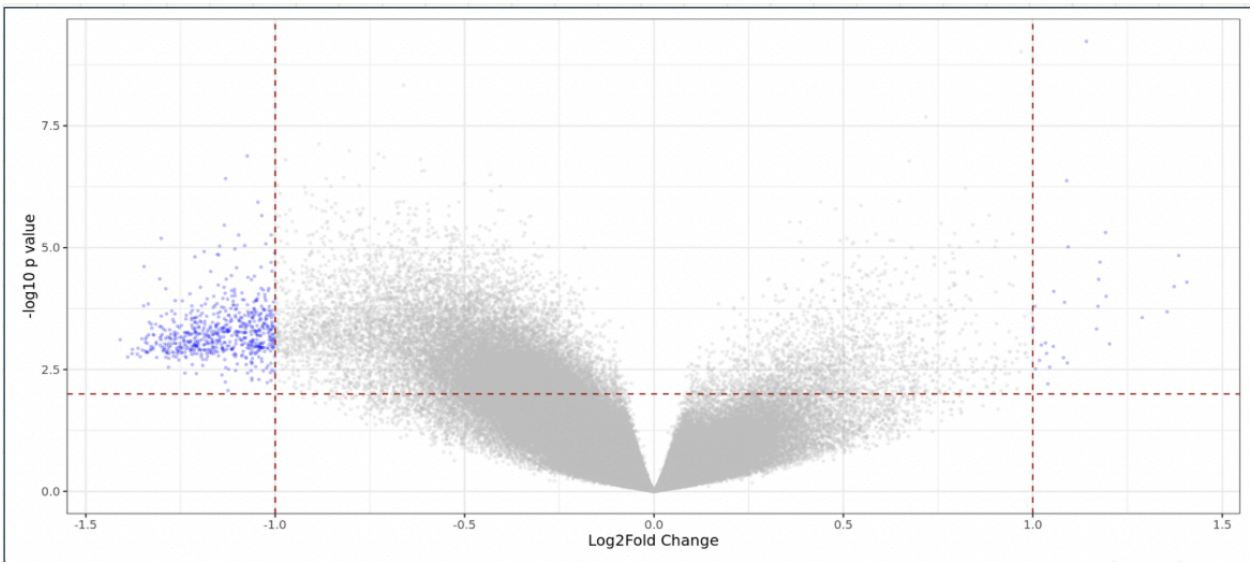

Figure 6. Enhanced Volcano Plot based on the differential expression analysis. This volcano plot for methylation showcases the difference in methylation between heavy smokers and non-smokers. The blue dots on the right showcase the genes that are significantly hypermethylated in the heavy-exposure smoking group compared to the non-smokers, and the blue dots on the left showcase the genes that are significantly hypomethylated in the heavy-exposure smoking group compared to the non-smokers.

As for interest genes, the following code block is used to find out the genes that are intersections between the four groups: upregulated, downregulated, hypermethylated, and hypomethylated.

```
upregulated <- results[(results$log2FoldChange > 1.25), 'gene_name']
hypomethylated <- dat[dat$foldchange < -0.75, 'geneName']
downregulated <- results[(results$log2FoldChange < -1.25),
'gene_name']
hypermethylated <- dat[dat$foldchange > 0.75, 'geneName']
interest_genes1 <- intersect(upregulated, hypomethylated)
interest_genes2 <- intersect(downregulated, hypomethylated)
interest_genes3 <- intersect(upregulated, hypermethylated)
interest_genes4 <- intersect(downregulated, hypermethylated)
```

The resulting interest genes were grouped based on shared significant differences in expression and methylation:

Table 2. **Table of Interest Genes** illustrating the shared significant differences in differential expression and methylation.

	Upregulated	Downregulated
Hypermethylated	SHOX2	ASIC2
Hypomethylated	ESX1, MYT1L, RIPPLY2, TDRD15	CTTNA2, LHX3, EBF3

Using the UCSC Genome Browser, the following figures were generated for specific interest gene groups:

For the upregulated and hypomethylated group:

ESX1:

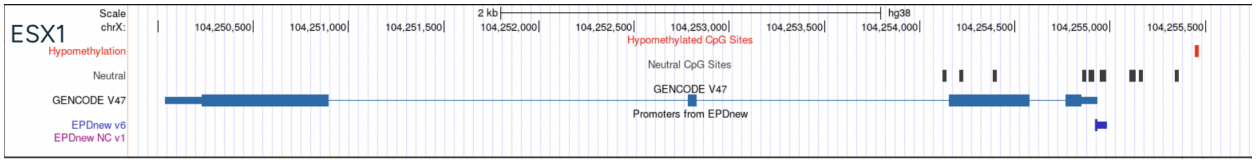

Figure 6. UCSC genome Browser picture for gene ESX1 does not appear significant since there are no hypomethylated CpG sites that overlap with any promoter region.

MYT1L:

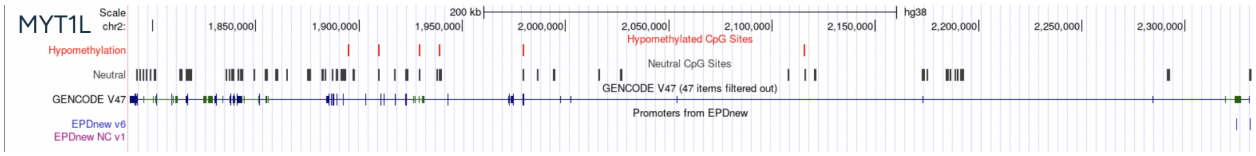

Figure 7. UCSC genome Browser picture for gene MYT1L does not appear significant since there are no hypomethylated CpG sites that overlap with any promoter region.

RIPPLY2:

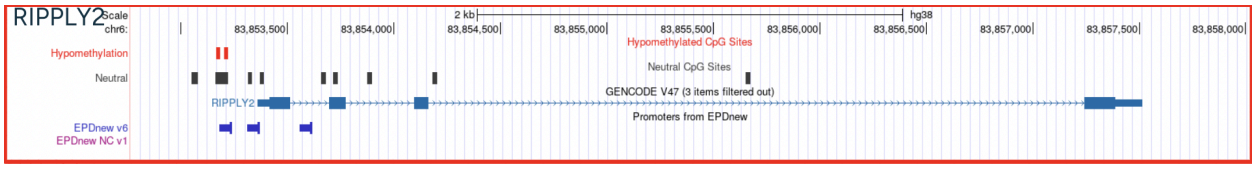

Figure 7. UCSC genome Browser picture for gene RIPPLY2 does appear significant since two hypomethylated CpG sites overlap with an EDPnew v6 promoter region. This might imply that the hypomethylation of the CpG sites near that promoter region leads to more expression of the protein coded by that promoter region. This might further imply that RIPPLY2 might be an oncogene in LUAD.

TDRD15:

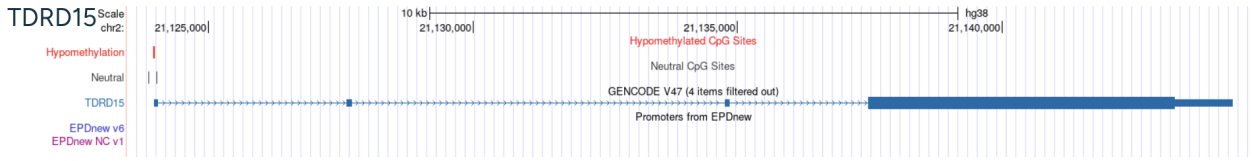

Figure 8. UCSC genome Browser picture for gene TDRD15 does not appear significant since there are no hypomethylated CpG sites that overlap with any promoter region.

For the downregulated and hypomethylated group:

CTNNA2:

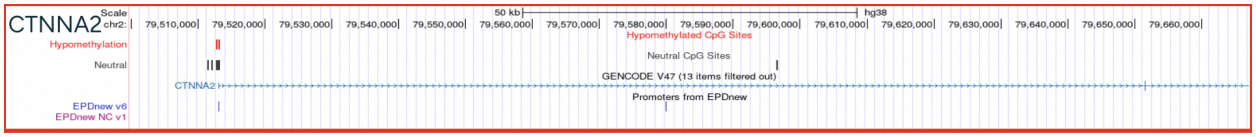

Figure 9. UCSC genome Browser picture for gene CTNNA2 does appear significant since one hypomethylated CpG site overlaps with an EDPnew v6 promoter region. This might imply that the hypomethylation of the CpG sites near that promoter region leads to less expression of the

protein coded by that promoter region. This is an interesting discovery since we have expected that less methylation would lead to more expression, but in this gene, it went the opposite way.

LHX3:

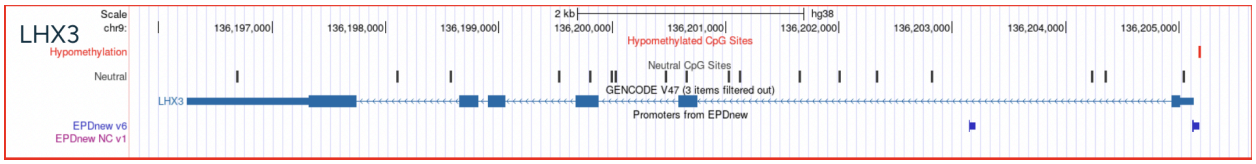

Figure 10. UCSC genome Browser picture for gene LHX3 does appear significant since one hypomethylated CpG sites overlap with an EDPnew v6 promoter region. This might imply that the hypomethylation of the CpG sites near that promoter region leads to less expression of the protein coded by that promoter region. This is an interesting discovery since we have expected that less methylation would lead to more expression, but in this gene, it went the opposite way.

EBF3:

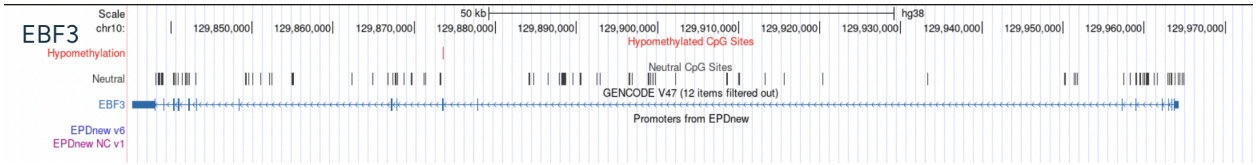

Figure 11. UCSC genome Browser picture for gene EBF3 does not appear significant since there are no hypomethylated CpG sites that overlap with any promoter region.

For the upregulated and hypermethylated group:

SHOX2:

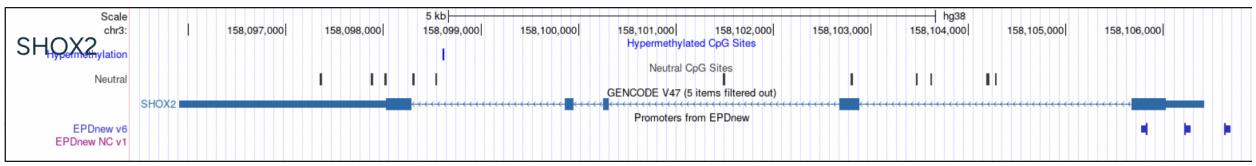

Figure 12. UCSC genome Browser picture for gene SHOX2 does not appear significant since there are no hypomethylated CpG sites that overlap with any promoter region.

For the downregulated and hypermethylated group:

ASIC2:

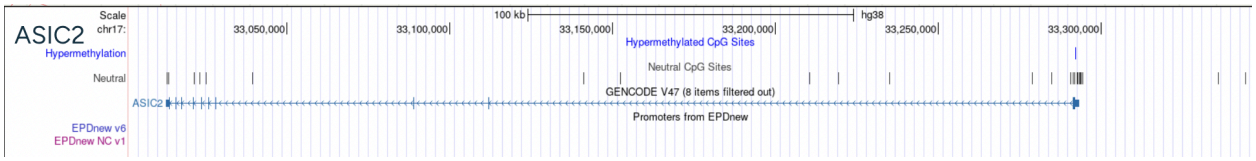

Figure 13. UCSC genome Browser picture for gene ASIC2 does not appear significant since there are no hypomethylated CpG sites that overlap with any promoter region.

Discussion

After the TCGA lung adenocarcinoma patients were subdivided into groups based on available smoking pack-year history, Figure 1 illustrates how there was not sufficient evidence to suggest significant differences between smoking LUAD groups in survival probability with a p-value of $0.79 > 0.05$. With a desire to investigate smoking differences, the heavy-exposure and non-smoker groups were analyzed further based on the conclusions of groups with the largest differences in smoking exposure. Based on Figure 2's co-oncoplot, TP53 was selected for further genomic analysis between the heavy exposure and non-smoking groups based on the highest number of mutations between the groups. With a lack of significant results after investigating the TP53 gene between the heavy-exposure and non-smoking groups as seen in Figures 3, 4, and 5, analysis of differential expression and methylation were performed on the two groups. Figure 5 illustrates how there were more significantly downregulated genes than significantly upregulated genes in the heavy-exposure (smoking) group, using non-smokers as the comparison baseline

group in the DESeq2 analysis. Furthermore, the volcano plot in Figure 6 demonstrates how there were more significantly hypomethylated CpG sites than significantly hypermethylated CpG sites between heavy-exposure and non-smoking groups (using the non-smoking group also as the baseline comparison). Table 2 was generated to show the found interest genes that had significance in differential expression and methylation and further investigated using the UCSC Genome Browser.

Specifically focusing on potentially significant interest genes, the ripply transcriptional repressor 2 (RIPPLY2) gene was analyzed to have significant upregulation and significant hypomethylation in the heavy-exposure group in comparison to the non-smoking group. Although there is likely not too much prior research on correlations between RIPPLY2 and lung adenocarcinoma (and specifically prior smoking history), RIPPLY2 has been seen to be associated with the prognosis of endometrial cancer (Li et al., 2020). Therefore, further multi-omic analyses should be performed focusing on RIPPLY2 and LUAD patients with heavy exposure to smoking to determine if RIPPLY2 can be used as a potential prognostic or even therapeutic factor for lung adenocarcinoma in patients with heavy exposure to smoking.

The catenin alpha-2 (CTNNA2) gene was analyzed to have significant downregulation and significant hypomethylation in the heavy-exposure group in comparison to the non-smoking group. According to Yang et al. (2021), CTNNA2 mutants had higher amounts of tumor neoantigens and tumor mutation burden, where longer overall survival is seen in lung adenocarcinoma patients with CTNNA2 mutation. Since this study primarily compares LUAD wild-type to mutants, the significant downregulation and hypomethylation in the LUAD smoking (heavy-exposure) group indicates how further studies such as genomic analysis of the CTNNA2 gene between smoking and non-smoking LUAD groups should be performed to determine if

CTNNA2 can be used as a potential indicator for the prognosis of LUAD patients with heavy-exposure to smoking.

The LIM homeobox 3 (LHX3) gene was analyzed to also have significant downregulation and significant hypomethylation in the heavy-exposure group in comparison to the non-smoking group. According to Lin et al. (2017), LHX3, with its increased expression in lung cancer "promotes cell proliferation and invasion," has been analyzed to determine an association "with unfavorable survival" in lung adenocarcinoma and can be used as "an early-stage and radiosensitivity prognostic factor" for LUAD. Similarly to CTNNA2, further genomic analysis should be used to determine if the significant downregulation and hypomethylation for heavy-exposure LUAD patients can be used to indicate if LHX3 can be used as a prognostic biomarker for patients that have heavy exposure to smoking and have lung adenocarcinoma.

In terms of limitations to the study, this paper was an observational study of sampled lung adenocarcinoma patients from the TCGA database, so only a correlation can be determined with the need for proper and adequate experimentation to potentially determine causation. Since this study focused on TCGA-LUAD patients, there are potentially some demographic discrepancies, such as a White racial majority of the sampled patients that may not reflect the overall population; therefore, the results and conclusions in the study may be only applicable to the populations similar to the patients sampled in the TCGA-LUAD database. With somewhat large differences in the smoking (heavy-exposure) and non-smoking group sample sizes (260 vs 70 samples respectively as shown in Table 1), there could be potential discrepancies in the conclusiveness of the results, so further investigation and study are needed for verification of the conclusions.

For future study and research, the significantly hypomethylated interest genes of RIPPLY2, CTNNA2, and LHX3 should be investigated further on potential correlations between hypomethylation, differential expression, and other multi-omics focuses for the heavy-exposure lung adenocarcinoma group. Furthermore, studying the multi-omics of certain subpopulations of lung adenocarcinoma TCGA patients based on age at diagnosis ranges and gender could be used to find potential diagnostic and/or prognostic indicators between smoking and non-smoking groups. With the potential limitations of the differences in smoking group sample sizes, comparing this study's conclusions to multi-omic analysis on other public databases such as the Gene Expression Omnibus and the Sequence Read Archive could potentially prove helpful in determining specific genetic indicators between lung adenocarcinoma patients with heavy-exposure (smoking) and non-smoking histories. Moreover, performing similar multi-omic analysis on heavy-exposure and non-smoker groups for other non-small cell lung cancers like squamous cell carcinoma (using the TCGA-LUSC database) could be used to compare results and identify potentially common target genes.

By identifying potential biomarkers or genes correlated with lung adenocarcinoma and smoking history, new therapeutic targets could emerge for patients with prior heavy exposure to smoking, potentially leading to more personalized prognoses and effective treatments. Such discoveries may allow for the development of prognostic tools to predict LUAD progression more accurately, enabling earlier intervention and better patient outcomes, specifically focusing on lung adenocarcinoma patients with heavy exposure to smoking. Although further investigation is needed, this study could contribute to potentially improving the prognosis of heavy-exposure patients with lung adenocarcinoma, to reduce the high mortality rate associated with LUAD by informing potential targeted therapy and prognostic indicators.

References

Govindan, R., Ding, L., Griffith, M., Subramanian, J., Dees, Nathan D., Kanchi, Krishna L., Maher, Christopher A., Fulton, R., Fulton, L., Wallis, J., Chen, K., Walker, J., McDonald, S., Bose, R., Ornitz, D., Xiong, D., You, M., Dooling, David J., Watson, M., & Mardis, Elaine R. (2012). Genomic Landscape of Non-Small Cell Lung Cancer in Smokers and Never-Smokers. *Cell*, *150*(6), 1121–1134. https://doi.org/10.1016/j.cell.2012.08.024

Hu, Yunqian, and Guohan Chen. "Pathogenic mechanisms of lung adenocarcinoma in smokers and non-smokers determined by gene expression interrogation." *Oncology letters* vol. 10,3 (2015): 1350-1370. doi:10.3892/ol.2015.3462

Li, X., Yin, F., Yuan, F., Cheng, Y., Dong, Y., Zhou, J., Wang, Z., & Wang, J. (2020). Establishment and validation of a prognostic nomogram based on a novel five‐DNA methylation signature for survival in endometrial cancer patients. *Cancer Medicine*, *10*(2), 693–708. https://doi.org/10.1002/cam4.3576

Lin, X., Li, Y., Wang, J., Han, F., Lu, S., Wang, Y., Luo, W., & Zhang, M. (2017). LHX3 is an early stage and radiosensitivity prognostic biomarker in lung adenocarcinoma. *Oncology Reports*, *38*(3), 1482–1490. https://doi.org/10.3892/or.2017.5833

Lung Adenocarcinoma Study - NCI. (2018, August 30). Www.cancer.gov. https://www.cancer.gov/ccg/research/genome-sequencing/tcga/studied-cancers/lung-adenocarcinoma-study

Ma, Sung Jun et al. "Association of Pack-Years of Cigarette Smoking With Survival and Tumor Progression Among Patients Treated With Chemoradiation for Head and Neck Cancer." *JAMA network open* vol. 5,12 e2245818. 1 Dec. 2022, doi:10.1001/jamanetworkopen.2022.45818

Sui, Q., Liang, J., Hu, Z., Chen, Z., Bi, G., Huang, Y., Li, M., Zhan, C., Lin, Z., & Wang, Q. (2020). Genetic and microenvironmental differences in non-smoking lung adenocarcinoma patients compared with smoking patients. *Translational Lung Cancer Research*, *9*(4), 1407–1421. https://doi.org/10.21037/tlcr-20-276

The Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, *511*(7511), 543–550. https://doi.org/10.1038/nature13385

US Preventive Services Task Force et al. "Screening for Lung Cancer: US Preventive Services Task Force Recommendation Statement." *JAMA* vol. 325,10 (2021): 962-970. doi:10.1001/jama.2021.1117

Yang, P et al. "Adenocarcinoma of the lung is strongly associated with cigarette smoking: further evidence from a prospective study of women." *American journal of epidemiology* vol. 156,12 (2002): 1114-22. doi:10.1093/aje/kwf153

Yang, W., Lin, A., Zhu, W., Wei, T., Luo, P., Guo, L., & Zhang, J. (2021). Catenin Alpha-2 Mutation Changes the Immune Microenvironment in Lung Adenocarcinoma Patients Receiving Immune Checkpoint Inhibitors. *Frontiers in Pharmacology*, *12*. https://doi.org/10.3389/fphar.2021.645862