Examining the Survival Rates and Gene Mutations of Lung Cancer Patients Based on Sex and Smoking History

By: Vamsi Chavali, Sabrina Zhong, Mengdi Chai

Introduction:

Lung cancer is the most commonly diagnosed cancer and has the highest death rate worldwide, with an estimated 2 million diagnoses and 1.8 million deaths per year (Thandra et al.). It is characterized by uncontrollable cell growth caused by mutations in specific genes that lead to the development of a tumor cell (Thandra et al., 2021). Because of the high prevalence and mortality rate of lung cancer, it is essential to discover the pathway and mechanism of lung cancer-related genes. Expression levels of cytochrome P4501A1 (CYP1A1) and Glutathione-S-transferase M1 (GSTM1) have been shown to be a risk factor for the development of lung cancer, especially in women (Zhang & Ye, 2020). Higher levels of CYP1A1 and mutations in GSTM1 result in the detoxifying process of the body being defective (Zhang & Ye, 2020). This, in turn, leads to a build-up of the carcinogenic substance from cigarettes, which seems to be more enhanced in women than in men (Ragavan & Patel, 2020). Additionally, there are also mutations in specific genes that have been linked to smoking. According to Al-Obaide et al., KLF6, TERT, MSH5, and GATA3 are some of the genes affected by DNA methylation caused by smoking that can lead to the development of lung cancer (Al-Obaide et al., 2018).

The Cancer Genome Atlas Program (TCGA) is a cancer genomics program that collects over 20,000 cancer tissue samples and matches them with normal samples spanning over 33 cancer types (National Institute of Health). The Clinical Proteomic Tumor Analysis Consortium (CPTAC) is a database that provides over 1100 patients' proteogenomic data to analyze cancer

(National Cancer Institute). These two publicly available databases are used for multi-omic data analysis, which integrates data sets from many omic groupings.

In this study, we explored the expression levels of GSTM1 and CYP1A1 in males and females and explored survival rates of lung cancer in different biological sex to identify whether biological sex does play a role in patient outcomes. To further the investigation of biological sex and its role in lung cancer development, we also analyzed the 4 genes linked to smoking (KLF6, TERT, MSH5, GATA3) between men and women and compared the survival rates of men and women with different smoking backgrounds. We hypothesized that biological sex and smoking history will affect the expression levels of sex-linked, mutation levels of smoking-linked genes, and patient outcomes.

Methods:

The analysis was mainly conducted in R and Python using lung cancer clinical and RNAseq data. The data was sourced from TCGA with the TCGAbiolinks library using accession code "LUAD." In the dataset, there were a total of 585 patients. Of the patients that had sex listed in their clinical data, 242 of the patients were male and 280 were female. Using the "Tobacco Smoking History" column in the clinical data, we compared patients who were listed as lifelong nonsmokers (category 1) with patients who have a history of smoking (category 2-4). Of the female patients, there were 51 nonsmokers and 215 smokers. Of the males, there were 18 smokers and 209 nonsmokers. The visualizations and analyses were created using an assortment of libraries within R and Python. In R, TCGABioLinks was used to access the data, DESeq2 was used to perform differential expression analysis, Maftools was used for mutation analysis, Survival and Survminer were used to create Kaplan-Meier Curves, and ggplot2 was used to create boxplots and heatmaps. In Python, CPTAC was used to access proteomic data, matplotlib

was used to develop scatterplots and seaborn was used to construct the heatmap to analyze gene expression relative to other genes.

Results:

The oncoplot in Figure 1 depicts the top 20 genes with the most mutations from the patients in the dataset. The TP53 gene is at the top of this list as 276 of the patients have a mutation on this gene. This is approximately half of the patients. The most common mutation for these genes is missense mutation shown as green. The genes that are focused on in this paper are not included in the top 20 genes (Fig. 1).

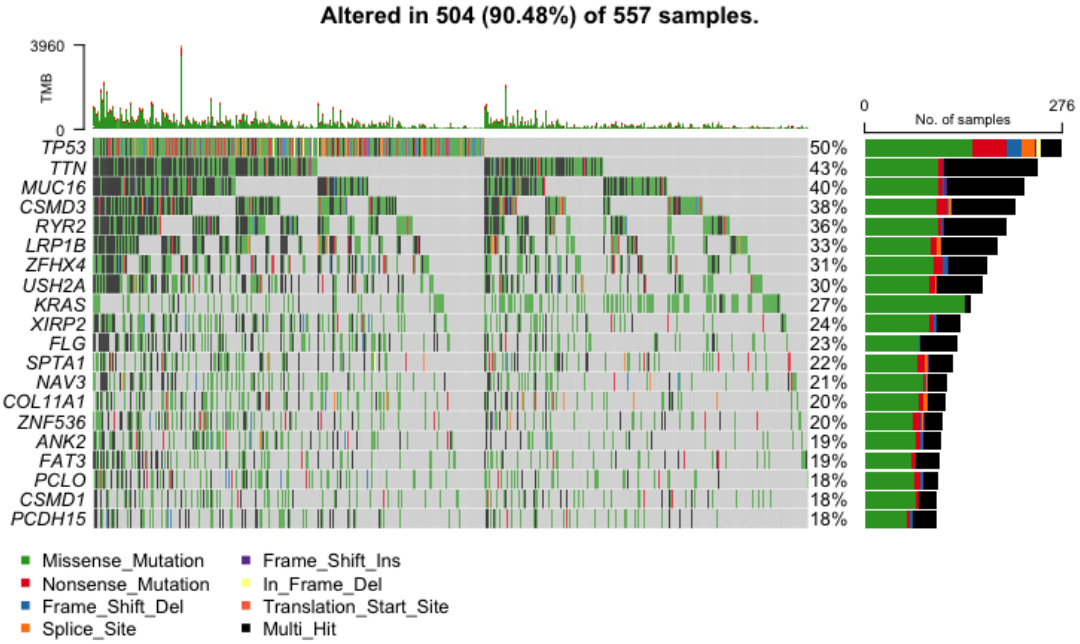


Figure 1. Oncoplot of the top 20 most mutated genes among the lung cancer patients in the TCGA and MAF dataset. The most common mutation in most genes is missense mutation. The most mutated gene is TP53 with 279, or 50%, of the lung cancer patients with a mutation in this gene. The genes examined in this research paper are not included in the top 20 genes.

Comparing the survival time of men versus women using a boxplot, we found that there is not a noticeable difference between the survival time between different sex. From figure 2 and excluding the outliers, the female patients lived until 40.38 to 88.39 years old, and the males lived until 42.38 to 88.98 years old. The median survival time for females is 68.28 years. The median survival time for males is 69.43 years. The median survival time for men and women is very similar. There is one outlier in the female patients and 2 for the male patients (Fig. 2).

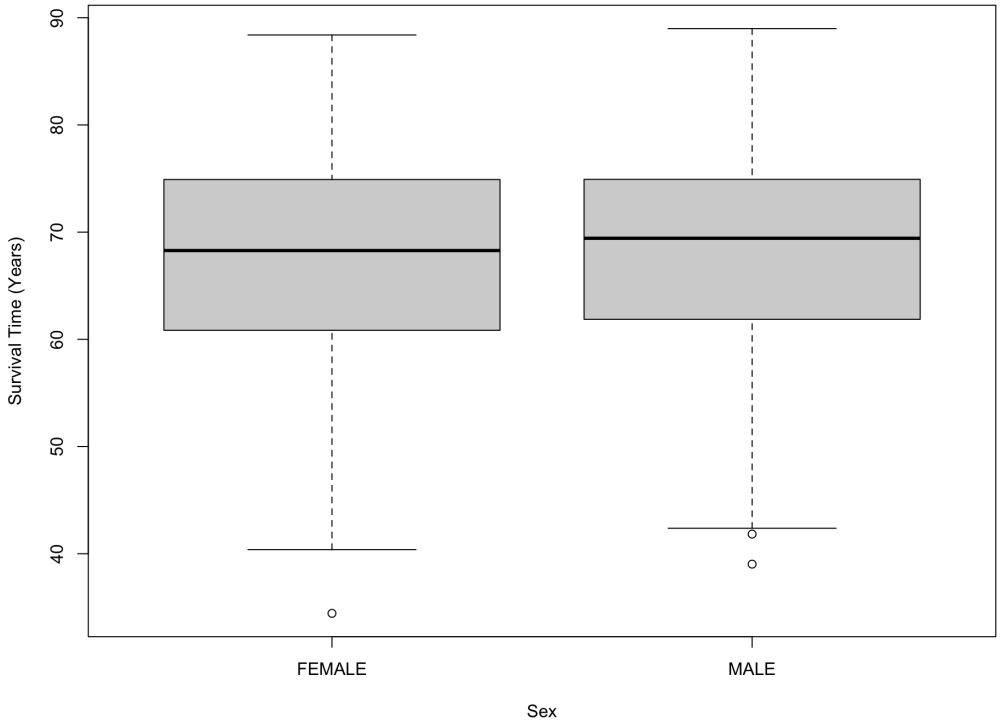


Figure 2. Boxplot of female and male lung cancer patients and their survival time in days. The survival time was calculated as the difference between the patient's date of birth and days to death. There does not seem to be a significant difference between the survival time of males and females as the median survival time for females is 68.28 years whereas for males it is 69.43 years.

We also observed the difference in GSTM1 gene counts and CYP1A1 gene counts among male and female patients as these genes have been researched to have a correlation with sex in lung cancer patients. The quartiles of the boxplot were hard to decipher without removing the outliers in the dataset. After removing the 29 outliers, the men and women patients both had a median of 0 counts of GSTM1. The women had a GSTM1 count range from 0 to 830 whereas the men had a range from 0 to 802. There does not seem to be a difference between the counts of GSTM1 between biological sex (Fig. 3). The same could be said about the CYP1A1 gene counts. There were 97 outliers in the CYP1A1 gene count dataset that were removed to produce the boxplot (Fig. 4). Again, the median count for both men and women was 0. The range from both men and women was 0 to 10. The counts for this gene are much lower than the counts for GSTM1 (Fig. 3, Fig. 4).

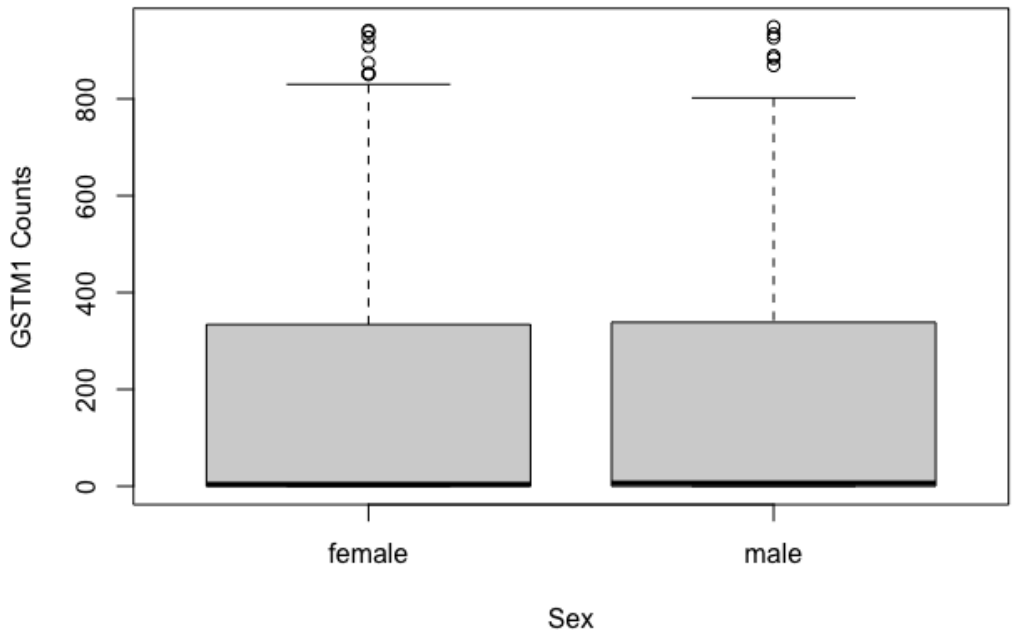


Figure 3. Boxplot of the GSTM1 gene counts between female and male lung cancer patients. The median for both boxplots is 0. The range for females is 0 to 830 counts. For males, it is 0 to 802 counts. There are still a few outliers with larger counts even after removing 29 outliers. There does not appear to be a difference between gene counts that correlates to biological sex.

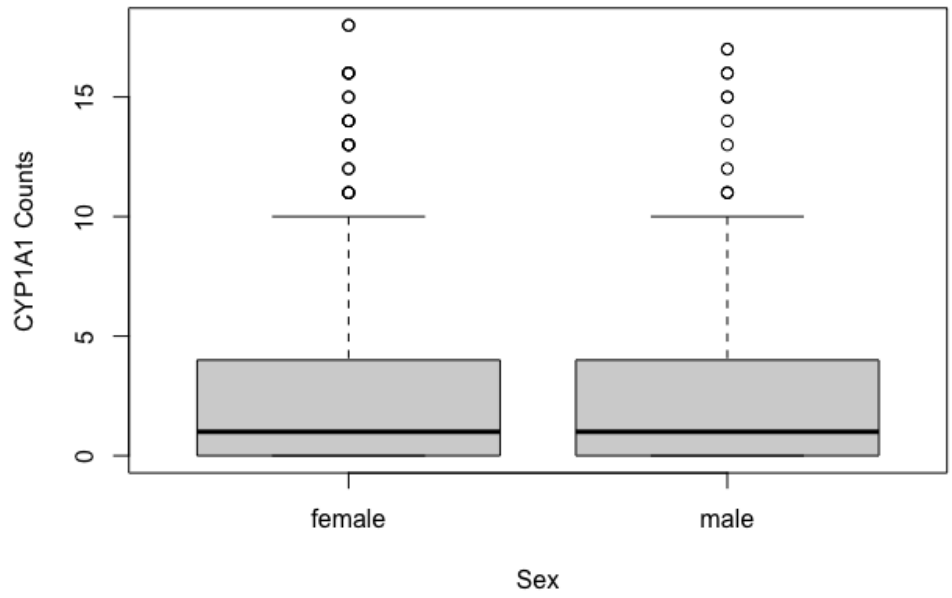


Figure 4. Boxplot of CYP1A1 gene counts between female and male lung cancer patients. 97 outliers were removed from the dataset. There are still a few outliers remaining after the initial clean-up. The median for both sexes is 0, and the range for both is 0 to 10. There does not appear to be a difference in gene counts between sex.

When observing the relationship of the GSTM1 gene with RNA and protein data from CPTAC, there does seem to be a correlation between the expression of RNA transcripts and proteins (p=3.84E-51). The correlation coefficient of the linear regression model is 0.863. Additionally, there seems to be two different groups in the scatter plot. There is one cluster of patients near the bottom left corner that shows a negative correlation of GSTM1 with RNA

transcripts and protein. The other group of patients is clustered around the top right corner of the figure. These patients show a positive correlation of GSTM1 with the RNA transcripts and proteins (Fig. 5).

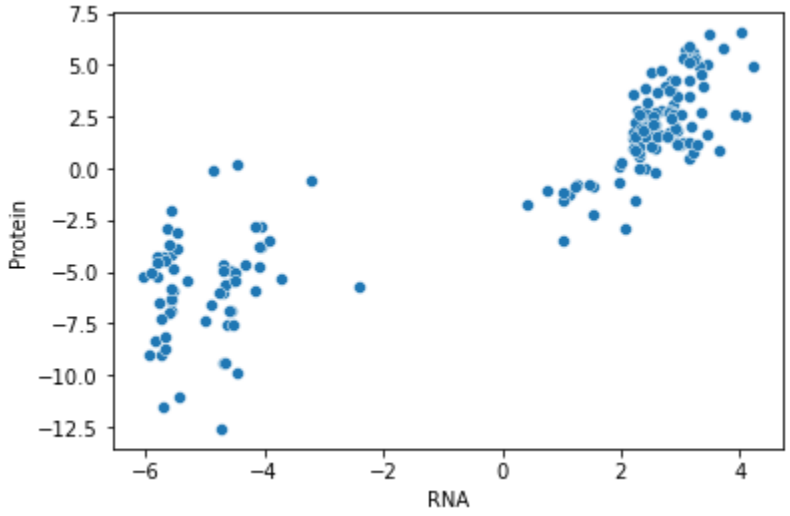


Figure 5. Scatterplot of the correlation between the RNA and protein data from CPTAC of the GSTM1 gene. The linear correlation coefficient is around 0.863, and there seems to be a significant correlation between the expression of the RNA transcript and protein (p=3.84E-51). There is one cluster of patients in the bottom left corner who seem to have low expression whereas the cluster on the top right does have a high expression.

Kaplain-Meier plots also show that there is not a significant difference between men's and women's survival rates. Overall, there is a drastic decrease in the survival probability for both men and women as time goes on. After around 2000 days, only about 0.25 females were alive and only about 0.275 males were alive. However, the two lines seem to be at around the same probability for each time frame (Fig. 6).

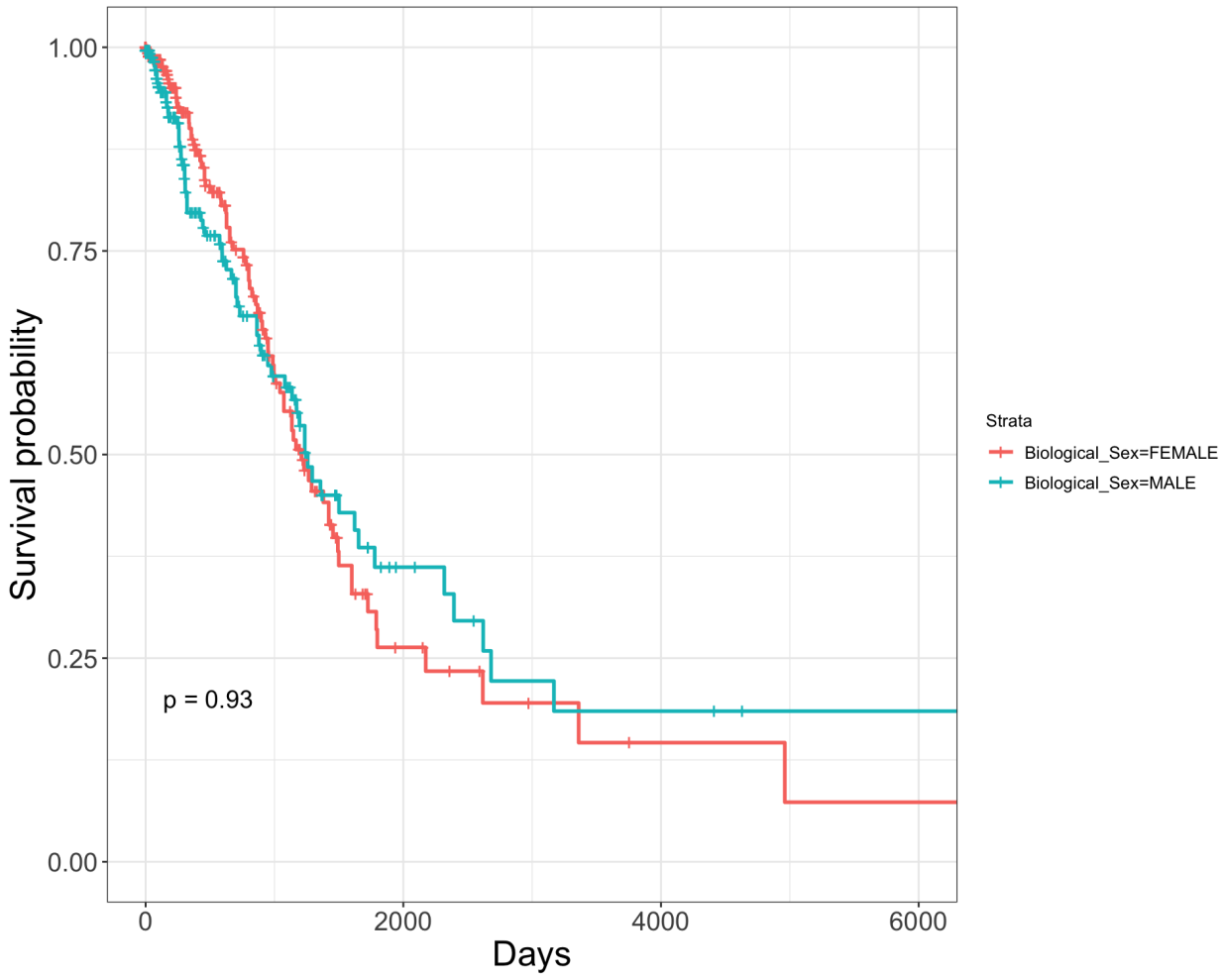


Figure 6. Kaplan-Meier plot of the survival probability between the 242 men and 280 women over time. The data for this graph was taken from the patient's clinical data in the TCGA dataset. The survival probability was calculated by the vital status of the patients based on the days until death as the time. From day 1000 onwards, females have a lower probability of survival. There is not a significant difference between the survival probability of the two groups (p=0.93).

When comparing the nonsmoker and various smoker groups of the patient sample, the most significant difference in survival rates between men and women comes from the nonsmoker group whereas there does not seem to be a significant difference in the smokers' groups. For

patients who do not have a history of smoking, the survival probability seems to be lower for men than women until around day 300. At this time, the survival probability for men is 0.5, and for women, it is about 0.85. The line depicting the survival probability for men is cut off after day 300 which may be due to the limited sample size of nonsmoker males (Fig. a). For the plot representing the patients who are listed as current tobacco smokers, the survival 7probability seems to be higher than the ones seen in figure 2. However, there is still no significant difference when comparing the survival rate of men and women in this category (Fig. 7b). For patients who have stopped smoking for at most 15 years, the Kaplain-Meier plot has the lowest p-value ($p = 0.12$). Females have a lower survival probability over time than males do. Even though this pattern occurs in the graph, the p-value is still too high to conclude the significance of the difference between survival probabilities. The survival probability of this group seems to be the lowest as the survival probability for women is 0 right before 200 days and for men is around 0.12 from day 2300 onwards (Fig 7c). The survival probability of patients who have stopped smoking for more than 15 years is the lowest. Females have a survival probability of 0 at close to 2000 days. For males, their survival probability flatlines at about 0.125 from about 2300 days onwards. This also shows that females who have smoked for a long time have a lower survival probability than men, but the difference does not seem to be significant ($p=0.38$) (Fig. 7d).

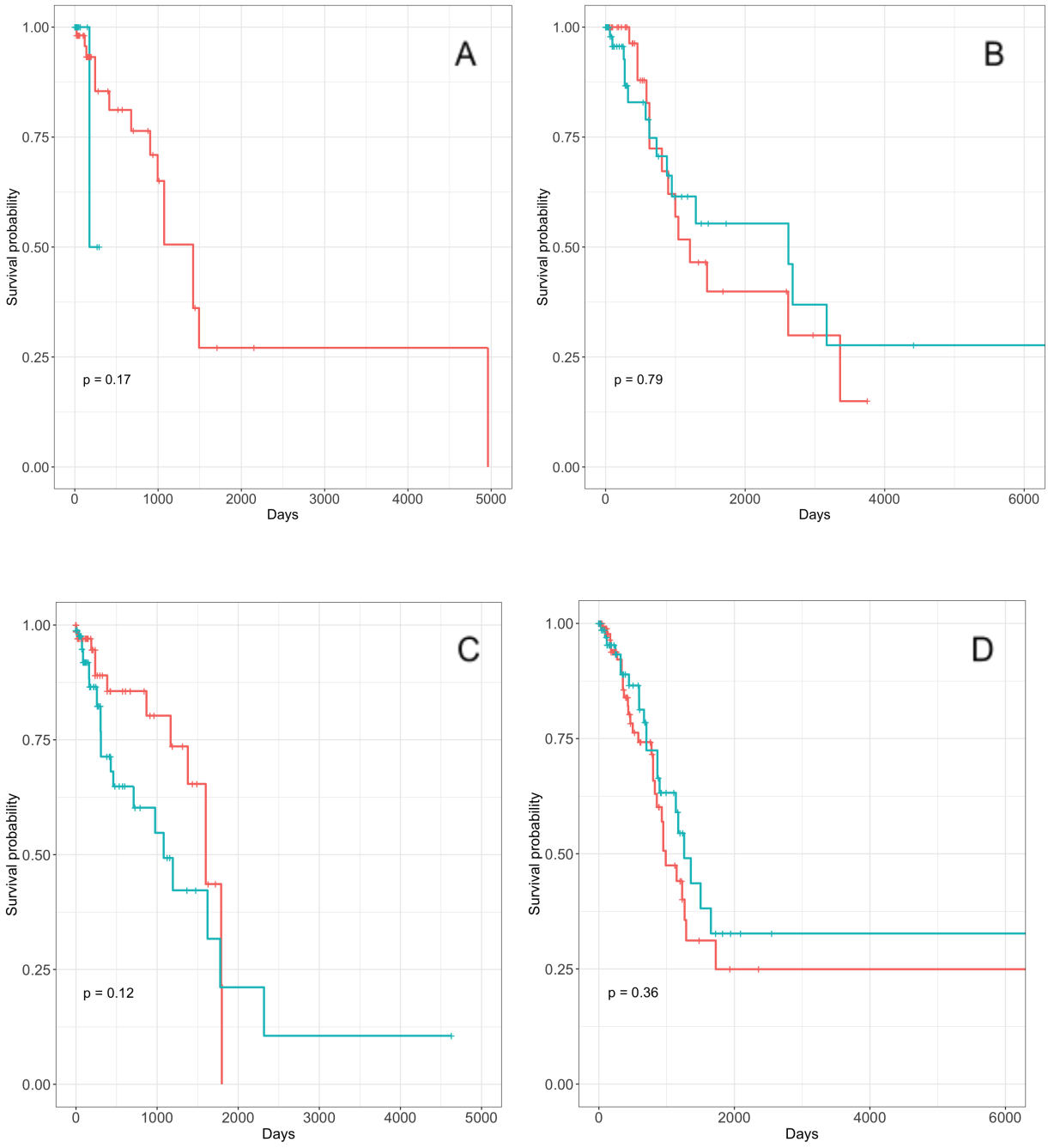

A
B
C
D
Survival probability
Days
p = 0.17
p = 0.79
p = 0.12
p = 0.36
1.00
0.75
0.50
0.25
0.00
0
1000
2000
3000
4000
5000
6000
Strata
Biological_Sex=FEMALE
Biological_Sex=MALE

Figure 7. Kaplan-Meier plot comparing the survival probability over time between men and women amongst different tobacco smoking backgrounds. The patients were subset into categories using the tobacco smoking indicator listed in the clinical data. The days until death were used as the time and the vital status was used as the event to calculate the survival rates of the patients. **A)** This plot compares patients who do not have a history of smoking, categorized as a 1 in the tobacco smoking history indicator. There does not seem to be a significant difference between the survival probability (p=0.17). There are only 18 males that fit in this category. **B)** This plot compares the patients who are current smokers. Their tobacco smoking indicator is listed as 2. There does not seem to be a significant difference in the survival probabilities (p=0.79). **C)** The two lines compare the males and females who have a tobacco smoking history indicator as 3. They are a reformed smoker and have not smoked for more than 15 years. There does not seem to be a significant difference between the survival probabilities between sex (p=0.12). This p-value is the lowest among all the plots. **D)** These patients had a tobacco smoking history indicator as a 4 in the clinical data which means they were not smoking at the time of the interview and have not been smoking for less than or equal to 15 years. There does not seem to be a significant difference between the survival probabilities between sex (p=0.36).

The lollipop plots compare the type and frequency of mutations in genes related to smoking between patients who have been lifelong non-smokers and patients who have a history of smoking. We further split the lollipop plots to depict the differences in sex as well. Overall, there seem to be more mutations in the genes related to smoking for both men and women patients who have a history of smoking. There does not seem to be any mutations in these genes for patients who do not have a history of smoking. All of the genes have missense mutations (Fig. 8-11). Other mutations include splice sites, frameshift deletions, and nonsense mutations.

The gene with the most amount of mutations is GATA3 for both men and women compared to the other genes (Fig. 8a, Fig. 8b).

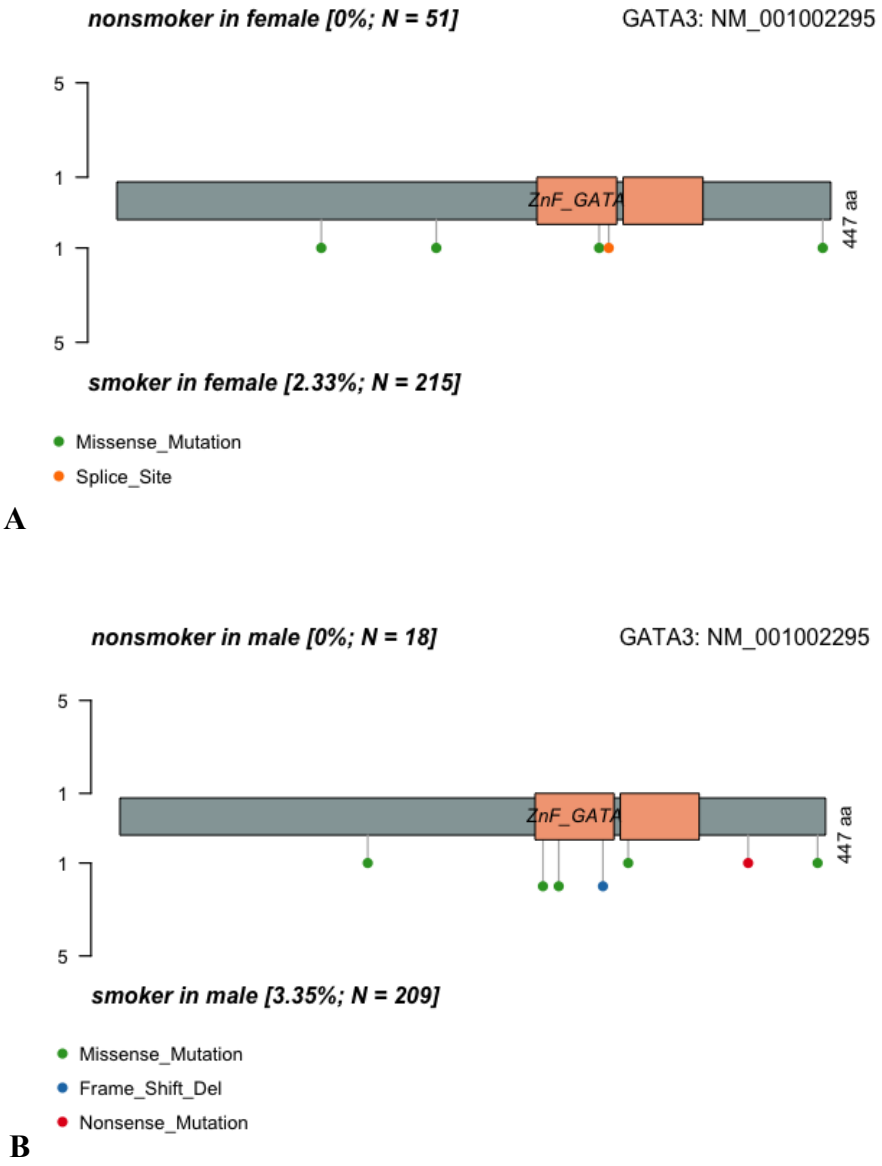
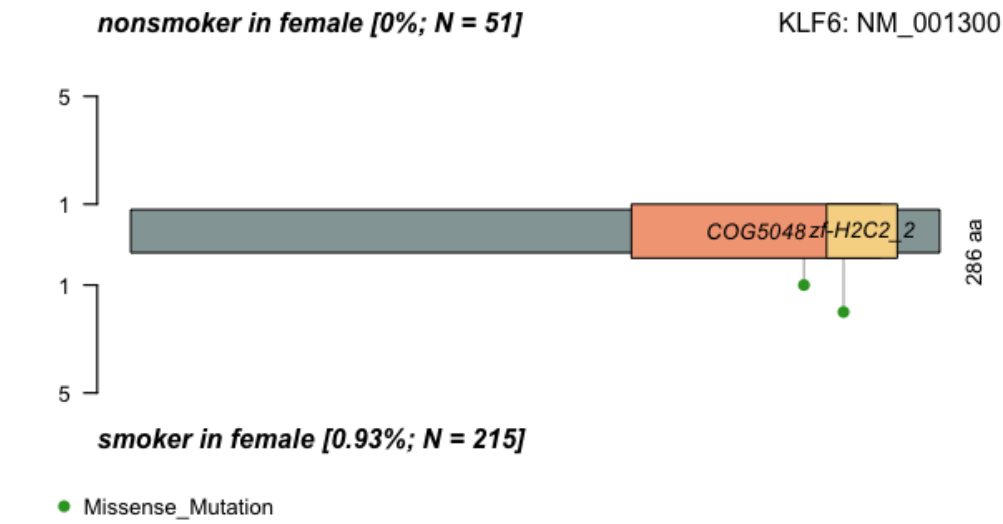
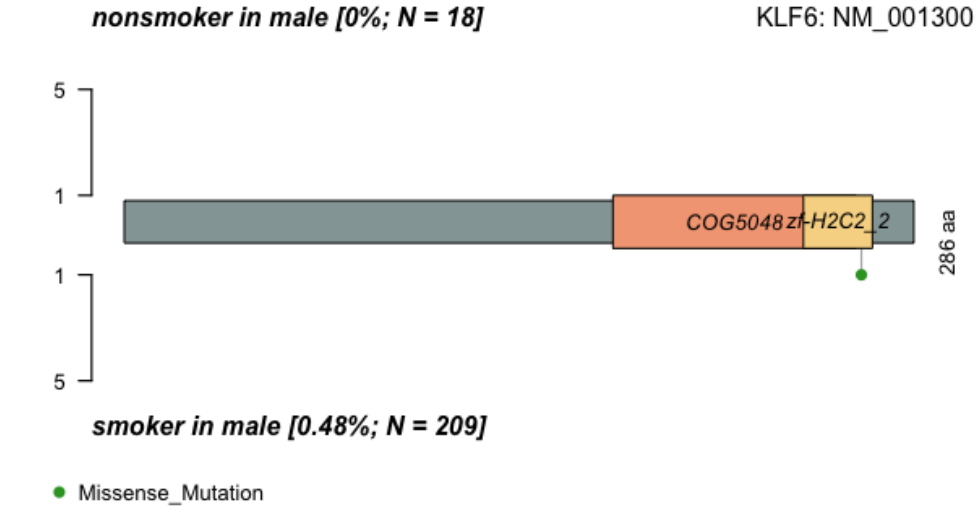


Figure 8. Lollipop plot of the mutation type and frequency in GATA3 gene from the MAF dataset between smoker and nonsmoker patients. Znf_GATA = GATA-type, zinc finger binding DNA protein. **A)** A plot for the female patients in the dataset. 4 out of the 215 female smoker patients have missense mutations and one has a splice site mutation on this gene. Two mutations

occur on the zinc finger protein. **B)** Lollipop plot for the male patients. 5 of the smoker patients have a missense mutation, one has a frameshift mutation, and one has a nonsense mutation.



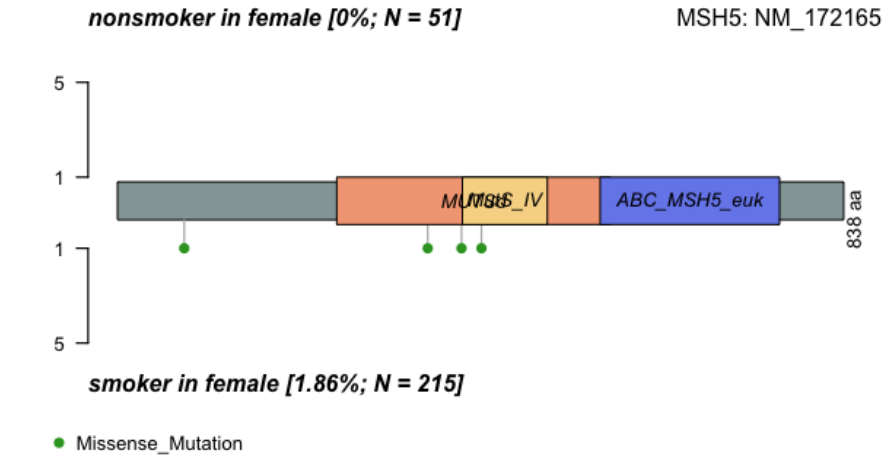
A



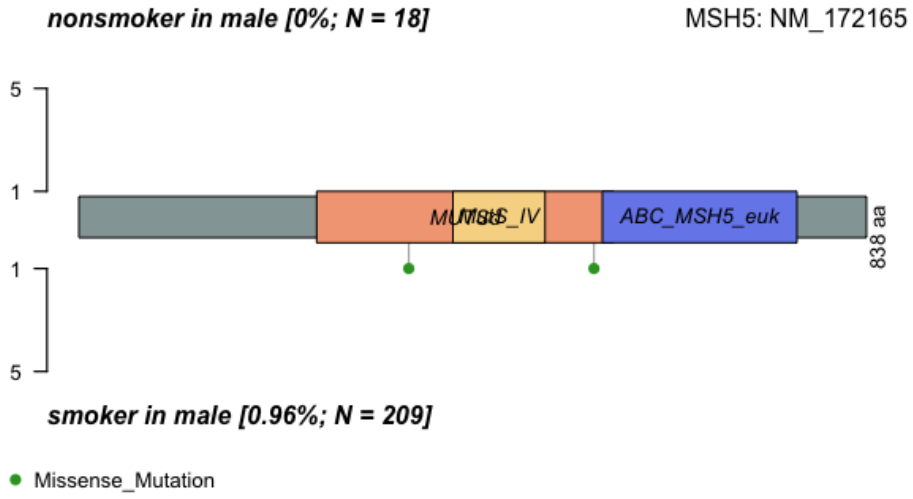
B

Figure 9. Lollipop plot of the mutation type and frequency in KLF6 gene from the MAF dataset between smoker and nonsmoker patients. COG5048, zf-H2C2_2 = zinc finger domain. **A)** This plot shows the mutations on KLF6 in female patients. Of the female patients, two have a history of smoking, and two of them have a missense mutation on the zinc finger domain. There are no

mutations for this gene in nonsmoker patients. **B)** This is the lollipop plot for the subset of patients who are male. Of the male smoker patients, only one has a missense mutation on the zinc finger domain. The nonsmoker male patients do not have any mutations on KLF6.



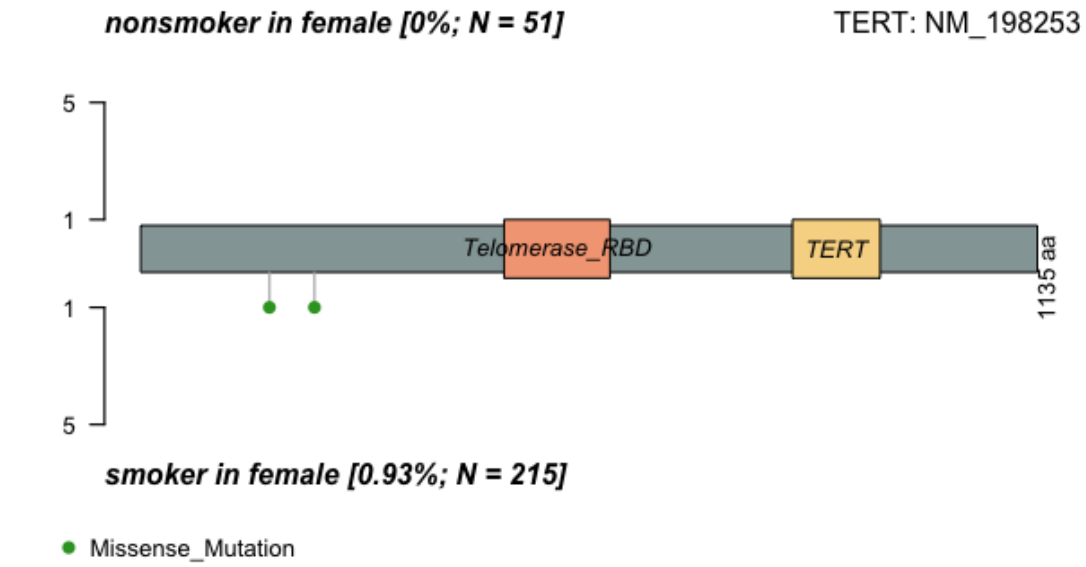
A



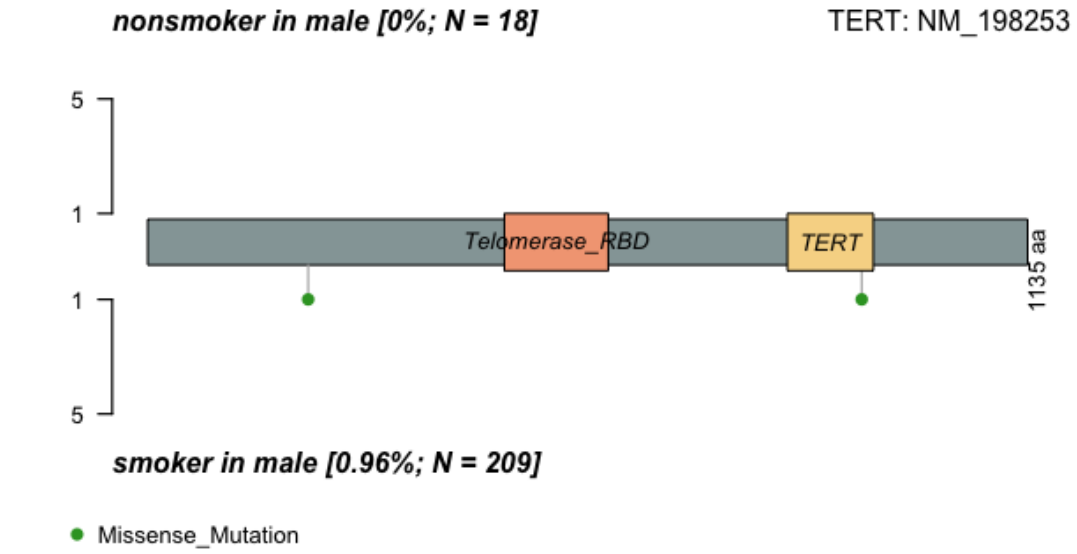
B

Figure 10. Lollipop plot of the mutation type and frequency in MSH5 gene between smoker and nonsmoker patients. ABC_MSH_euk = ATP-binding cassette domain of eukaryotic MutS3 homolog. **A)** The plot for female smokers and nonsmoker patients. Out of the 215 female smoker

patients, 4 of them have a missense mutation on this gene. **B)** Plot of the male smoker and nonsmoker patient mutations. Of the 209 smoker patients, 2 of them have a missense mutation on this gene.



A



B

Figure 11. Lollipop plot of the mutation type and frequency in TERT gene between smoker and nonsmoker patients. Telomerase_RBD = Telomerase ribonucleoprotein complex, an RNA binding domain. **A)** A plot of the nonsmoker and smoker female patients. Only two female smoker patients have a missense mutation on this gene. **B)** Lollipop plot of smoker and nonsmoker male patients. There are two missense mutations on this gene in the male smoker patients.

Discussion:

Lung cancer has myriad risk factors that can disproportionately affect individuals. In this study, we examined two of those factors: sex and smoking history. The general consensus assumes the existence of a relationship between biological sex and lung cancer survival. However, results show that female patients with lung cancer have no comparative advantage in survival compared to their male counterparts (Fig. 2 & Fig. 6). This contradicts the findings presented at the World Conference on Lung Cancer by the International Association for the Study of Lung Cancer: women with newly diagnosed stage I to stage III non-small cell lung cancer had a longer OS than their male counterparts. The reason for the contradiction may vary, but can likely be attributed to the small size of patients and missing data in our study. Additionally, when analyzing the sex genes *GSTM1* and *CYP1A1*, we found that there is no significant correlation between sex gene expression in male and female patients. Thus we can conclude, that sex genes do not have an explicit connection with lung cancer (Fig. 3 & Fig. 4) which is in agreement with Zhang et al.. In the study, they concluded that the combined impact of GSTM1 present/null and CYP1A1 MspI polymorphisms do not result in a substantially higher LC risk (Zhang et al., 2020). In this study, two statistically insignificant major clusters were found when analyzing RNA transcript and protein expression levels in GSTM1. As a potential

avenue for future research, the relationship shown in Fig. 5 that shows patients having high expression of RNA transcript and protein related to GSTM1 should be further explored.

When examining the smoking history of lung cancer patients, it was found that there exists a difference between life-long non-smoker patients and their survival probabilities between the two sexes, (Fig. 7A) which is corroborated by the article from Park et al., that nonsmokers had a higher sex difference in lung cancer incidence (Park et al., 2020). Evaluating tobacco smokers, there was not an obvious difference between sex survival time and smoking status (refers to current smokers, stop smoking less than 15 years smokers, and stop smoking more than 15 years smokers) (Fig. 7B & 7C & 7D) which is verified by Park et al. in the article states that there was no statistically significant interaction between sex and smoking status (Park et al., 2020). However, it is observed that there is no significant decrease in the survival probability with the smoking-year increases, regardless of sex, and this finding is contrary to popular belief and the paper from Flanders et al., the fatality rate from lung cancer climbed far faster with each extra year of smoking than with each additional cigarette per day (Flanders et al., 2003). The problem occurs in the small population size. In our research, there are only 585 patients, but in the study from Flanders et al., there are more than 6,569,144 participants included. The small population size increases the probability of inaccurate results. In future research, we can look at datasets with a more balanced sample size between smokers and nonsmokers to further examine the difference in survival rates.

This study posits that the presence of smoking-related genes can affect lung cancer incidence rates. The genes KLF6, TERT, MSH5, and GATA3 were found to only have a mutation in smokers (Fig. 8 & Fig. 9 & Fig. 10 & Fig. 11). A recent study found that the association between lung cancer-related genes and tobacco smoking was validated by the discovery of 13

unique CpGs in 8 genes with lower DNA methylation than never-smoking sites: KLF6, STK32A, TERT, MSH5, ACTA2, GATA3, VTI1A, and CHRNA5 (49) (Al-Obaide et al., 0001), thus agreeing with our findings. Although there is limited sample size for nonsmoking patients, those smoking-related genes must have some effects on the risk of lung cancer and the survival probabilities which are yet to be discovered for our future research. In addition, the oncoplot does not include any of these genes as the top 20 most mutated genes. Future research efforts can be focused more on genes on the oncoplot, such as TP53, and its relationship to patient survival probability or association with smoking-related gene expression and mutation.

Previous research by Al-Obaide et al. and Ragavan and Patel has illustrated that lung cancer has a significant correlation with sex and smoking. If we separate those factors by only examining the sex effect on lung cancer or smoking effect on lung cancer, we could easily find the relationship between one single factor and lung cancer. However, when we try to combine those factors together and find a combination of factors that plays a role in lung cancer, it becomes complicated. Until now, we still do not have a clear conclusion about the combination of sex and smoking history on the risk and survival outcome of lung cancers.

References

Albain KS, et al. Abstract OA06.01. Presented at: *International Association for the Study of Lung Cancer's World Conference on Lung Cancer*; Sept. 23-26, 2018; Toronto, Canada.

Al-Obaide, Ibrahim, B. A., Al-Humaish, S., & Abdel-Salam, A.-S. G. (2018). Genomic and Bioinformatics Approaches for Analysis of Genes Associated With Cancer Risks Following Exposure to Tobacco Smoking. *Frontiers in Public Health*, *6*, 84–84.

Alvarado, Muñoz, A. M., Bartra, M. S., Valderrama-Wong, M., González, D., Quiñones, L. A., Varela, N., Bendezú, M. R., García, J. A., & Loja-Herrera, B. (2021). Frequency of CYP1A12A polymorphisms and deletion of the GSMT1 gene in a Peruvian mestizo population. *Pharmacia*, *68*(4), 747–754.

Flanders, W. D., Lally, C. A., Zhu, B.-P., Henley, S. J., & Thun, M. J. (2003, October 14). *Lung cancer mortality in relation to age, duration of smoking, and daily cigarette consumption: Results from Cancer Prevention Study II1*. American Association for Cancer Research. Retrieved April 22, 2022, from https://aacrjournals.org/cancerres/article/63/19/6556/510417/Lung-Cancer-Mortality-in-Relation-to-Age-Duration

National Cancer Institute. (n.d.). *CPTAC*. Office of Cancer Clinical Proteomics Research. Retrieved April 24, 2022, from https://proteomics.cancer.gov/programs/cptac

National Institute of Health. (n.d.). *The Cancer Genome Atlas Program*. National Cancer Institute. Retrieved April 21, 2022, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Park, B., Kim, Y., Lee, J., Lee, N., & Jang, S. H. (2020, December 31). *Sex difference and smoking effect of lung cancer incidence in Asian population*. Cancers. Retrieved April 22, 2022, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7794680/#:~:text=There%20was%20no%20statistically%20significant,P%2Dinteraction%3A%200.261).

Ragavan, M. V., & Patel, M. I. (2020). Understanding sex disparities in lung cancer incidence: are women more at risk?. *Lung cancer management*, *9*(3), LMT34.

Thandra, Krishna Chaitanya, et al. "Epidemiology of Lung Cancer." *Contemporary Oncology (Poznan, Poland)*, Termedia Publishing House, 2021, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8063897/.

Zhang, He, X.-F., & Ye, X.-H. (2020). Association between the combined effects of GSTM1 present/null and CYP1A1 MspI polymorphisms with lung cancer risk: An updated meta-analysis. *Bioscience Reports*, *40*(9).