A Multi-omic Analysis of Lung Adenocarcinoma: An Examination of Differential Smoking Status and Sex-related Survival Outcomes to Inform Clinical Decision-making

Avinash Chauhan, Kenneth Nguyen, Brandon Ye

QBIO 490 Final Project

December 5, 2022

INTRODUCTION

Broadly defined as malignancies affecting the mucus-secreting cells of the lung periphery, lung adenocarcinoma (LA) is the leading cause of cancer mortality in the United States and comprises nearly 40% of all non small-cell lung cancers (NSCLCs) as the most commonly diagnosed primary subtype (Myers, DJ. et al., 2022). Chemotherapy or radiation therapy, often coupled with lobectomy or segmentectomy, is the current standard for treatment of early-stage, non-metastatic LA and provides the best chance of cure. More advanced, metastatic presentations of LA, however, are typically deemed incurable given currently available interventions (Harbeck, N. et al., 2019). There are a multitude of well-evidenced risk factors for increased LA incidence, including demographic, occupational, and lifestyle-related factors, with an important indicator being smoking. As both LA incidence and fatality have witnessed a steady increase over the past decades, the indication has become increasingly associated with a high cost at both patient and societal sublevels, specifically with regard to disease surveillance and treatment (Nagy-Mignotte, H. et al., 2011). Therefore, there exists an evident need to improve upon the current clinical paradigm for LA diagnostics and treatment to promote earlier detection of the disease prior to metastatic progression, improve survival outcomes, and ultimately reduce mortality.

Of the extensively-studied risk factors associated with LA, the association between smoking and LA incidence seems to be widely understood. As a complex mixture of carcinogenic chemicals, cigarette smoke conveys a high potential to metabolically activate DNA adducts, leading to miscoding and mutations in *KRAS*, *TP53*, and other genes important for growth control mechanisms and preventing tumorigenesis (Hecht, S. S., 2012). However, despite this knowledge, the connection between smoking and LA mortality remains unclear; conclusions that smoking and LA mortality follow the association between smoking and LA incidence (Ferketich, A. K. et al., 2012) are confounded by others that report no significant difference in long-term survival outcomes between smoking and non-smoking subgroups (Meguid, R. A. et al., 2010). Therefore, there exists potential value in investigating the relationship between smoking and LA mortality utilizing a large-scale dataset, especially in the context of influential marker genes in the LA genomic mutation profile.

Given that the current large-scale study of LA is hindered by a lack of high-quality data, standardized study design, and agreeable parameters, leading to evidently confounding results. Such shortcomings are largely addressed by The Cancer Genome Atlas (TCGA), an NIH-funded public-data project which aims to create a comprehensive collection of multi-omic cancer data via large-scale genome sequencing and various other forms of integrated multi-dimensional analyses (Tomczak, K. et al., 2015). For this study of LA, the TCGA Lung Carcinoma (TCGA-LUAD) database, in combination with proteomic data from the Clinical Proteomic Tumor Analysis Consortium (CPTAC) program, provides valuable clinical and demographic as well as multi-omic data for requisite analyses. Utilizing the R and Python programming languages, we synthesize this multimodal data, analyze it through various data visualization techniques, and abstract relevant conclusions.

Herein, we discuss differential LA survival outcomes by smoker status with special consideration for the unique genetic and proteomic profile of each subgroup. Secondarily, we analyze sex-related differences in survival within the smoking subgroup, concluding that while the differential survival between smoking and non-smoking subgroups is statistically insignificant in the context of this dataset, the statistically significant difference in the survivorship of male versus female smokers presents an avenue for new clinical insights. The purpose of this study, therefore, is to (a) confirm what is already understood about the relationship between smoking status and LA incidence, (b) generate novel insights regarding LA mortality informed by a holistic analysis of multi-omic data to ultimately (c) arrive at clinically-relevant conclusions with the potential to guide the future of LA treatment and personalized therapy.

METHODS

The multi-omic analysis was conducted by querying lung adenocarcinoma clinical and genetic data from the NCI TCGA database (accession code TCGA-LUAD) as well as CPTAC proteomic data for the lung adenocarcinoma dataset.

The non-proteomic analyses were conducted on R using packages BiocManager, maftools, TCGAbiolinks, ggplot2, SummarizedExperiment, survival, and survminer. Clinical data was processed into a clinical dataframe and then used to produce a grouped boxplot and histogram. The grouped boxplot was created by using the is.na function to filter out invalid values and label smokers/non-smokers as well as the boxplot function. The histogram was created by subtracting the year of tobacco smoking onset from the year of initial pathological analysis for each patient and using the hist function.

Upon completion of the clinical dataframe, boolean masking was employed to create two columns of filtered data for the smoking status within the dataset. The newfound column, named smoking_status, assigned labels of smoking history (smoking_positive and smoking_negative) based on the already existing values of smoking tendencies in the dataframe. If patients had no discernible values within these smoking-related columns, they were assigned to the smoking_negative cohort, and vice versa.

Kaplan-Meier curves were drawn via the survival and survminer packages to depict the differential effects of smoking status and sex on survival, ultimately yielding a measure of statistical significance. The differential effects of smoking status were appraised by the aforementioned smoking status, while the differential effects of sex on survival were conducted solely among patients with a smoking history. Survival time was dependent on available data within the dataframe; if the “days to death” variable was not marked as NA, then the value was used to indicate survival time, and if it was marked as NA, then the “days to last followup” variable was used in its place.

The updates to the clinical dataframe were processed via the maftools package. An initial oncoplot was developed with the oncoplot function and processed maf data, involving both the subsetted columns as clinical features, to plot the most commonly mutated genes within the

smoking subpopulation. The dataframe was then subsetted into two separate data frames for both smokers and nonsmokers, and boolean masking was used to store and subset the patient barcodes. The co-oncoplot function was then used to draw two oncoplots side by side, providing a differential analysis of gene expression between the two smoking status cohorts. The gene with the most disparate mutational frequency within the co-oncoplot was highlighted by co-lollipop plots, which also utilized the subsetted binary objects for both cohorts within the lollipopPlot2() function. A similar process was performed for the stratification of sex, where an oncoplot was developed for male and female cohorts within the smoking subpopulation; upon subsetting and boolean masking to store the sex-stratified barcodes, the co-oncoplot compared the mutational differential rate and the co-lolliplot plot highlighted the gene locus of the most differentially expressed mutation elucidated by the co-oncoplot.

The proteomic analysis was conducted using Python, with packages cptac, numpy, pandas, seaborn, matplotlib.pyplot, and scipy.stats. After the packages were installed and the LUAD data was obtained with the cptac.download function, proteomic and transcriptomic data was extracted with get_omic functions. Filtered datasets for this RNA and protein data were created to only include genes in both the RNA and protein datasets using the np.intersect1d function. Then, correlations for each gene combination were calculated with np.ndarray and stats.spearmanr, and these correlations were plotted on a correlation heatmap using the sns.heatmap function. To obtain the protein/RNA, protein/protein, and RNA/RNA heatmaps, the corresponding datasets were provided as arguments for the stats.spearmanr function. This analysis yielded the most prominent genes for further analysis of co-occurrence and potential mutual exclusivity. These genes were explored within a Draftsman plot, which incorporated RNA Summarized Experiment data within R. Masks for each of the most prominent gene mutations were created, while smoking status and survival columns were developed that were commensurate with those within the clinical dataframe. The clustering color scheme was set by smoking status, and the resulting draftsman plot plotted correlations between each of the genes included in this analysis.

RESULTS

Upon analyzing the ages of lung adenocarcinoma diagnosis based on smoking status (Figure 1), there is no major difference between the two groups. While the median age of diagnosis for lung adenocarcinoma is slightly higher for non-smokers (≈68 years) than smokers (≈66 years), the spread is wide and similar for both groups, suggesting that based on the TCGA clinical data, smoking status has a minimal, if any, effect on how early patients develop lung adenocarcinoma.

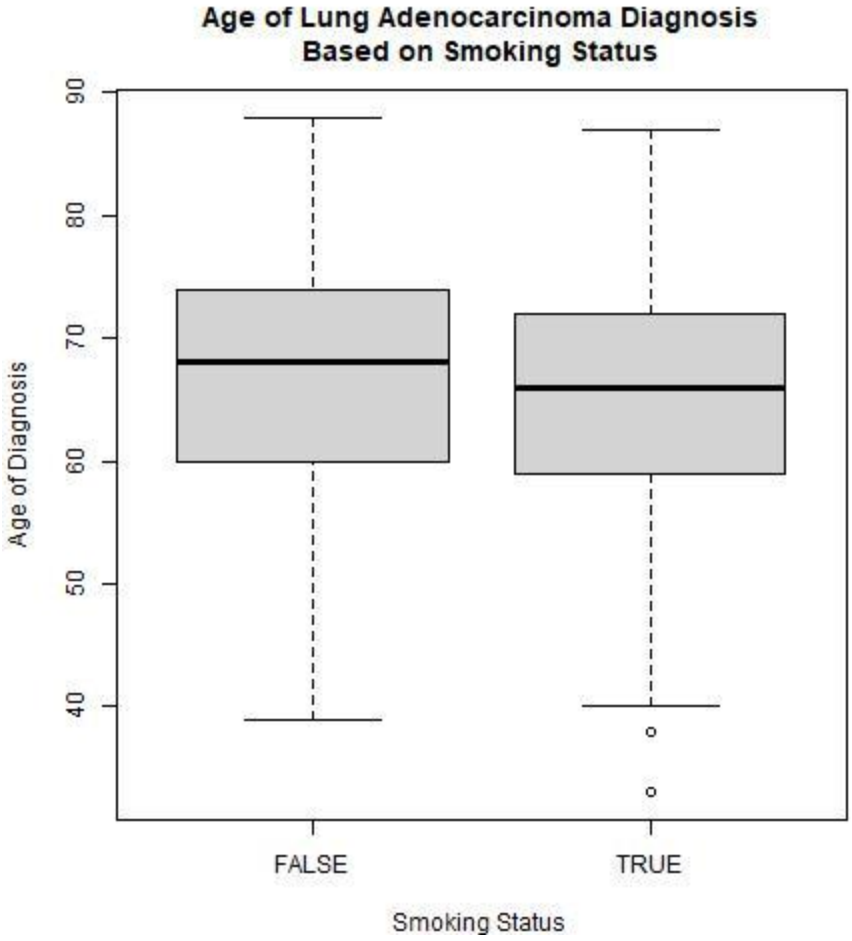

Figure 1. A paired boxplot showing that lung adenocarcinoma patients tend to get diagnosed slightly earlier if they smoke (median≈68 years) vs. if they don't smoke (median≈66 years). However, due to the large spread for both groups, no significant conclusions can be drawn from the difference in diagnosis age between the two groups.

Interestingly, even confirmed smoking statuses are poor predictors of when patients are diagnosed with lung adenocarcinoma. Upon analyzing a histogram of the difference in years between the first time a lung adenocarcinoma patient smoked and the year the patient was diagnosed with lung adenocarcinoma (Figure 2), there is a large spread in the distribution of data. The median is contained in the 40-45 year bar, although this time difference comfortably ranges from around 10 years to around 75 years. This suggests that a patient's smoking status does not have a clear effect on the timeframe of clinical diagnosis.

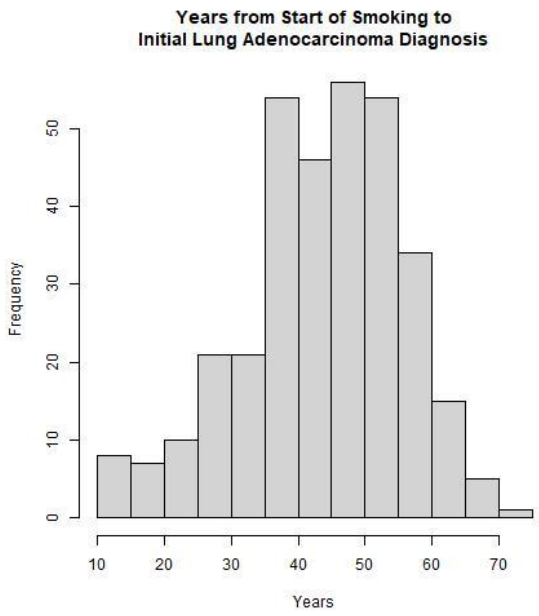

Figure 2. A histogram of years from a patient's start year of smoking to LA diagnosis year, indicating a median 40-45 year difference. There is also a wide and normally-distributed spread, indicating that the year difference between these two events can vary widely.

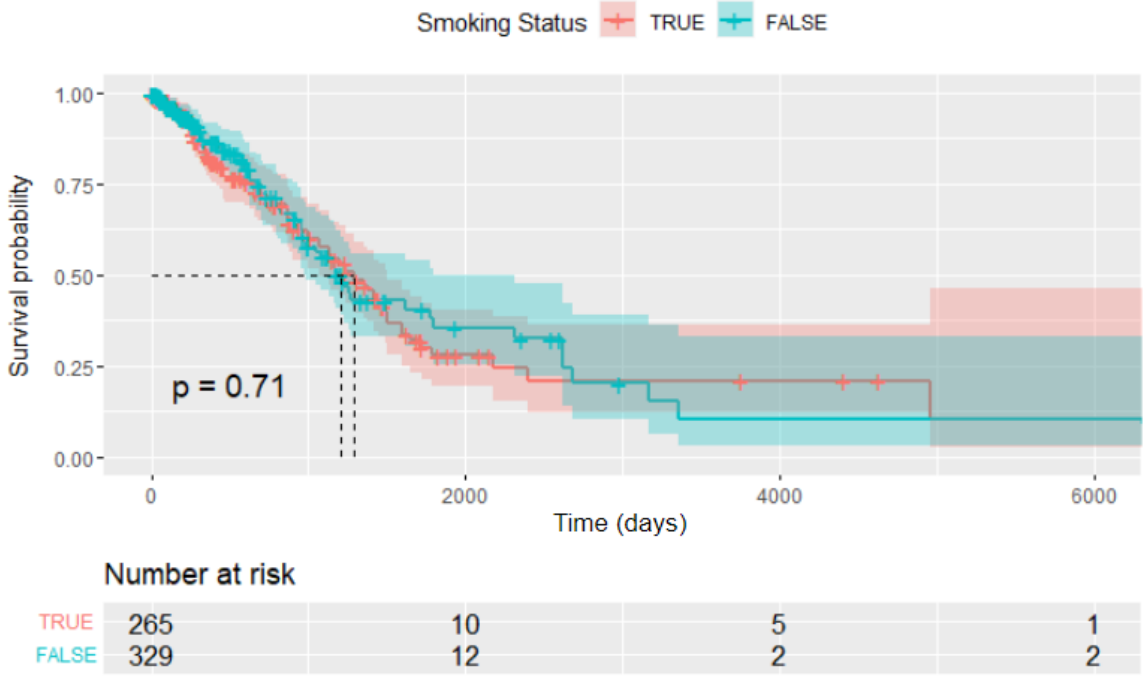

Figure 3. Kaplan-Meier analysis shows that LA patients who smoke do not exhibit a significantly lower survival rate (p = 0.71) than LA patients that do not smoke. Plots were made using the survplot function on R to plot the survival probabilities of various anatomical regions, using breast cancer clinical data retrieved from TCGA-LUAD. Time is calculated here as the time from diagnosis to the death or final follow-up, depending on data availability.

The first Kaplan-Meier survival analysis of smoking status found a relatively insignificant difference ($p = 0.71$) between the survival probabilities of smoking and non-smoking cohorts (Figure 3). At 50% survival, there appears to be a largely negligible difference in the time interval between the two cohorts. The confidence intervals appear to be heavily intersected throughout the plot, and these confidence intervals grow at higher extremes of the time axis by virtue of the increasingly limited patient counts, indicating the limited statistical power of insights made at this range.

A secondary Kaplan-Meier survival analysis, which took the smoking cohort and stratified it by sex, found a relatively more significant difference ($p = 0.062$) between the survival probabilities of males and females (Figure 4). Interestingly, despite this p-value holding far more significance than the prior survival analysis ($p = 0.71$), there still appears to be a largely negligible difference in the time interval between the two cohorts at 50% survival. The confidence intervals appear to be heavily intersected throughout the plot, and these confidence intervals grow at even higher extremes of the time axis than the prior survival plot, as there are no surviving male patients beyond approximately 3000 days, while there are 4 such female patients.

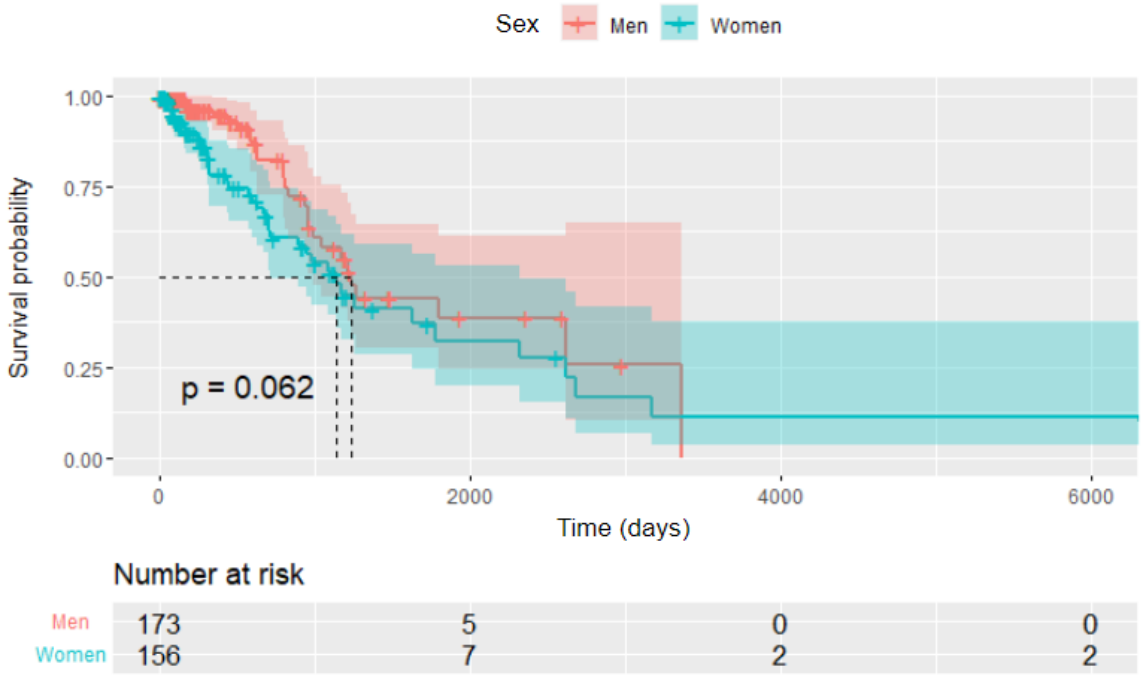

Figure 4. Kaplan-Meier analysis shows that male smoking patients have somewhat significant differential survival probability ($p = 0.062$) compared to female smoking patients. Time is calculated here as the time from diagnosis to the death or final follow up, depending on data availability.

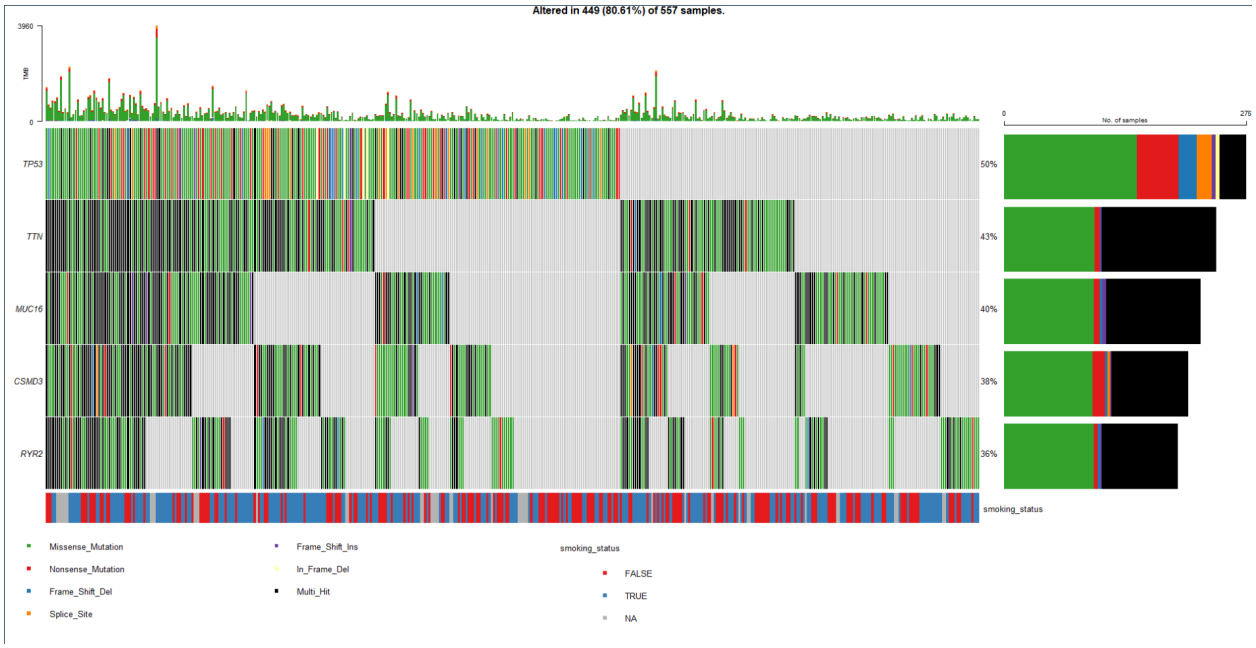

Figure 5. Oncoplot utilzing mutation annotation format (MAF) data highlights the top 5 most frequently mutated genes within LA patients who smoked. *TP53* demonstrated the highest diversity of mutations relative to the mutational distribution of the next 4 more frequently mutated genes.

Oncoplot analysis was focused on the smoking cohort, and the five most commonly mutated genes were determined: *TP53*, *TTN*, *MUC16*, *CSMD3*, and *RYR2* (Figure 5). The right-handed distribution of mutation type across each gene was largely homogeneous for most of the genes, as four of the genes were largely composed of missense and multi-hit mutations. However, *TP53* displayed a conspicuous diversity of mutation, with a relatively high rate of nonsense, frameshift, and splice site mutations. There appeared to be little mutual exclusivity between each of the mutational distributions.

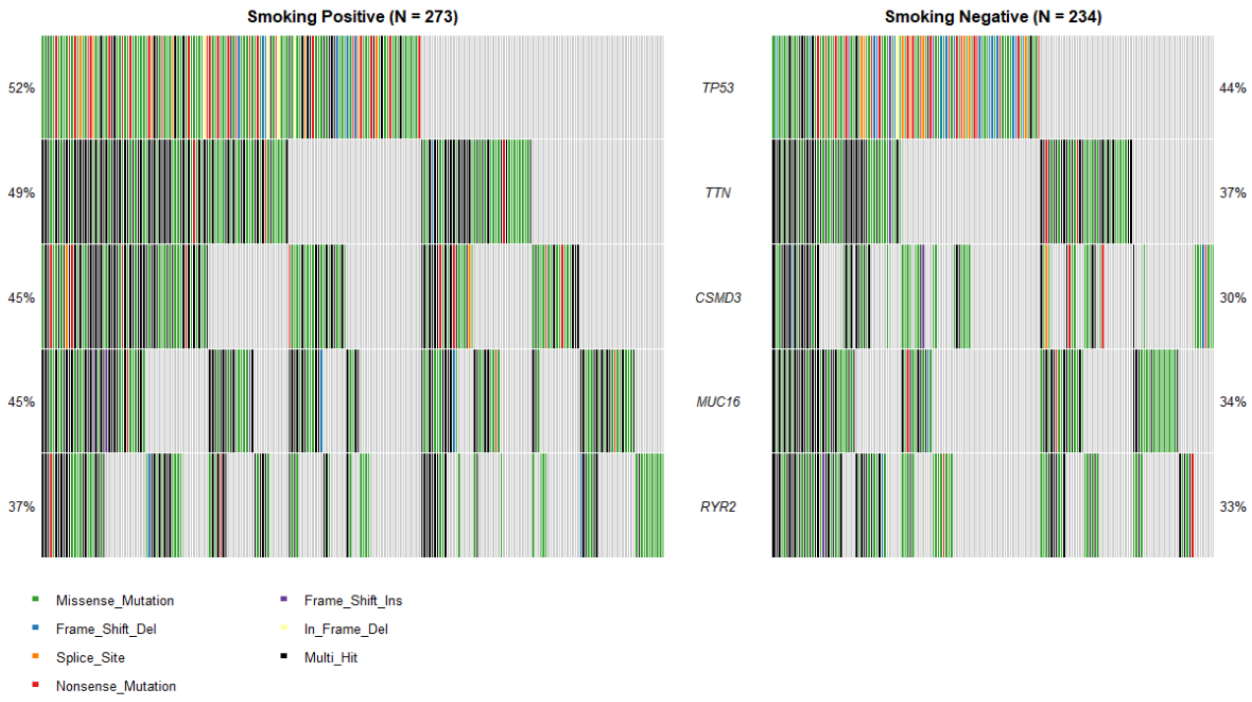

Figure 6. Co-oncoplot analysis identifies *TP53*, *TTN*, *CSMD3*, *MUC16*, and *RYR2* as the top 5 genes with the highest rate of somatic mutations in the overall patient dataset. Rates of mutation were higher for the smoking cohort across each of the 5 genes.

Co-oncoplot analysis utilized the five most commonly mutated genes appraised in Figure 5, mapping the differential mutation rate between the smoking and non-smoking cohorts (Figure 6). For each of the five genes, the smoking cohort exhibited higher rates of mutation compared to its nonsmoking counterpart. The *TTN* gene displayed a particularly significant mutational differential, with a 12% higher mutation rate within the smoking cohort. Within the individual cohorts, however, the order of the genes with the highest rates of mutation remained the same.

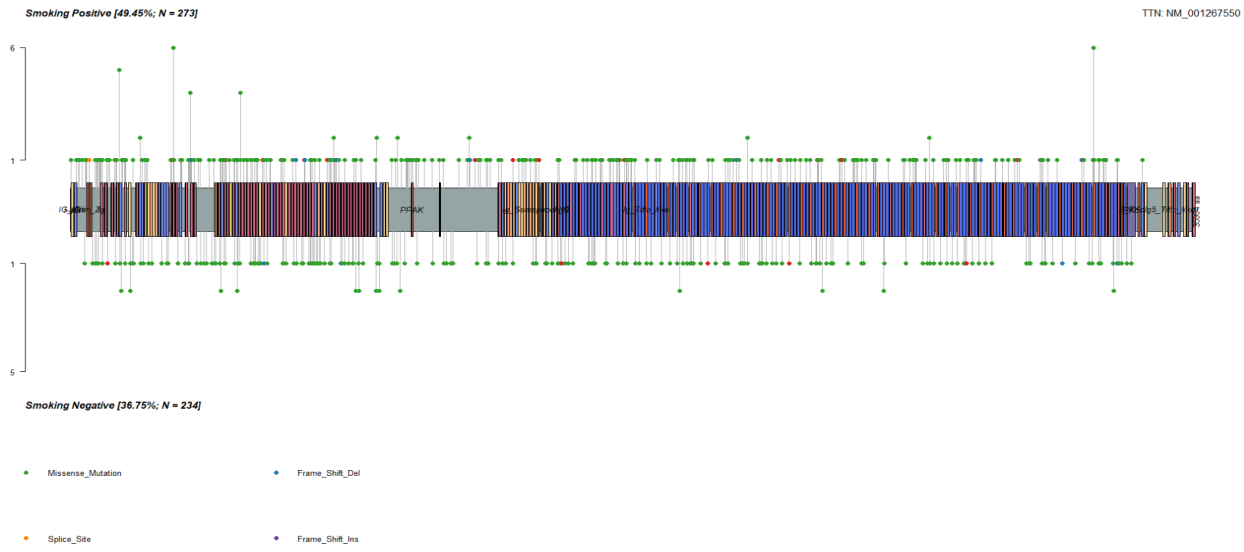

Figure 7. Co-lollipop analysis of *TTN* in smoking and nonsmoking patients reflects differential mutation frequency in space. The leftward domain of the *TTN* gene locus exhibited higher rates of mutation within smokers than non-smokers.

The gene *TTN*, which demonstrated a significant mutational rate differential between the smoking and nonsmoking cohorts (Figure 6), reflected distribution variations across the gene locus (Figure 7). The spatial distribution was largely similar for both cohorts in the central regions of the gene locus. However, a spatial region on the gene locus consisting of a number of left-handed domains demonstrated significantly higher rates of mutation within the smoking cohort than the non-smoking cohort. *TTN* mutation expression was spatially skewed right for smoking patients, while it was much more evenly distributed for non-smoking patients. However, there was generally a similar diversity and frequency of mutation types for both cohorts of patients.

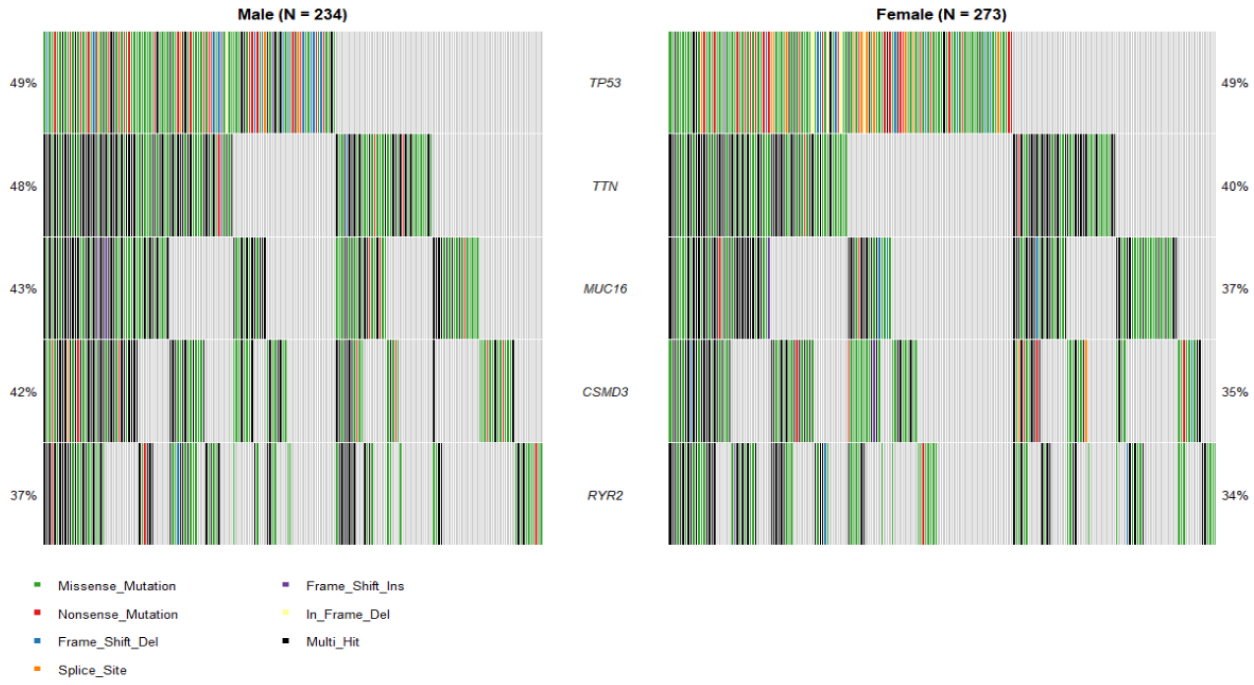

Figure 8. Co-oncoplot analysis shows the top 5 genes with the highest rate of somatic mutations within the smoking patients, stratified by sex. Rates of mutation were higher for the smoking cohort for 4 of the 5 genes.

Co-oncoplot analysis utilized the five most commonly mutated genes appraised in Figure 5, mapping the differential mutation rate between the male and female cohorts within the smoking population (Figure 8). For four of the five genes, the male cohort exhibited higher rates of mutation compared to its nonsmoking counterpart. The *CSMD3* gene displayed a particularly significant mutational differential, with a 7% higher mutation rate within the smoking cohort. Within the individual cohorts, however, the order of the genes with the highest rates of mutation remained the same.

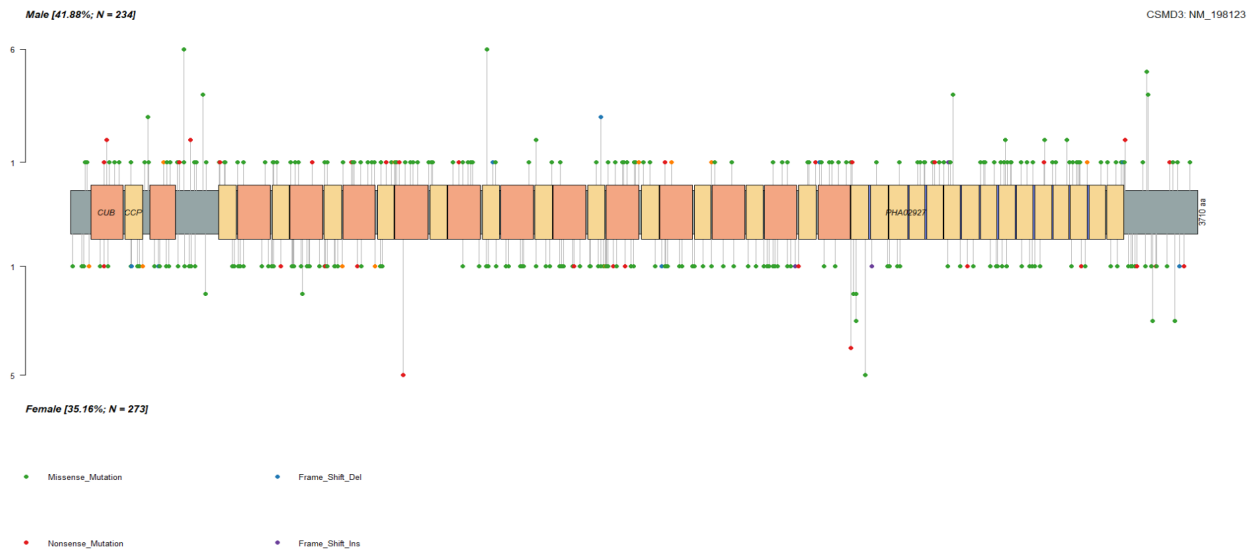

Figure 9. Co-lollipop analysis of *CSMD3* between smoking and nonsmoking patients reflects differential spatial mutation frequency. The leftward and rightward domains of the *CSMD3* gene locus exhibited higher rates of mutation within males than females.

The gene *CSMD3*, which demonstrated a significant mutational rate differential between the male and female cohorts of smokers (Figure 8), reflected distribution variations across the gene locus (Figure 9). The spatial distribution was somewhat heterogeneous throughout the gene locus. A number of hotspots on the gene locus reflected significantly higher rates of mutation. For instance, the regions between domains on the left and right extremes of the gene locus displayed much higher rates of mutation among males, while a central domain displayed a significantly higher rate of mutation within the female cohort. However, there was generally a similar diversity and frequency of mutation types for both cohorts of patients.

Given all of these findings for individual genes, we then transitioned to an analysis of pairs of genes to identify any correlations between different pairings. Upon creating a heatmap comparing protein expression with RNA expression for different pairs of genes (Figure 10), it became clear that three genes (*LRP1B*, *TTN*, and *RYR2*) stood out due to their relatively high correlations. While most pairings yielded poor correlations, many of the pairings between these three genes had a correlation coefficient value of around 0.5-0.6, indicating a moderate positive linear correlation between these pairings. Interestingly, the correlation values remained similar when comparing the RNA expression of the three genes of interest (*LRP1B*, *TTN*, and *RYR2*) to the same protein's expression, suggesting a potential correlation in RNA expressions between these three genes.

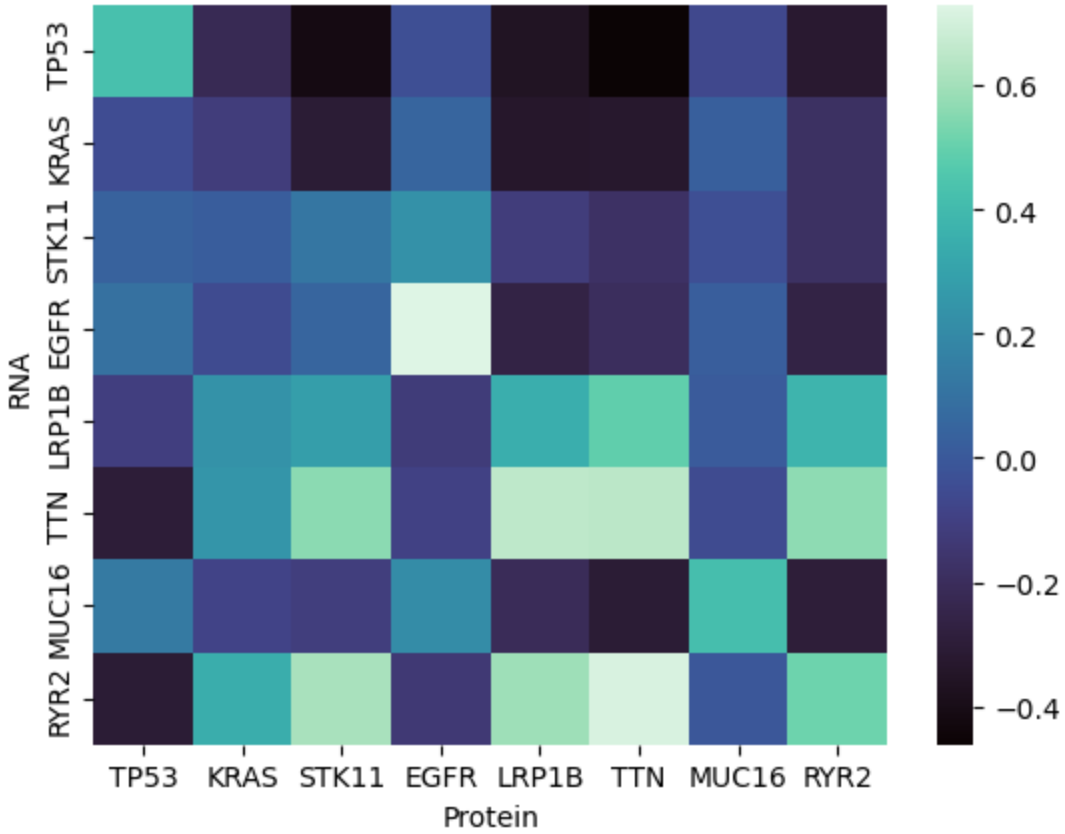

Figure 10. A heatmap showing strengths of correlations between RNA expression and protein expression for different combinations of commonly-mutated LA genes, indicating stronger correlations for RNA-protein pairs including *LRP1B*, *TTN*, and *RYR2*. Additionally, for each protein, pairings with RNA from each of the three genes yield similar correlations.

Upon creating a new heatmap comparing only protein expression for different pairs of genes (Figure 11), *LRP1B*, *TTN*, and *RYR2* still maintained high correlation values relative to other gene pairings. The Spearman correlation (r) values for these three pairings were around 0.6-0.7, suggesting slightly stronger linear positive correlations for protein-protein pairings than protein-RNA pairings.

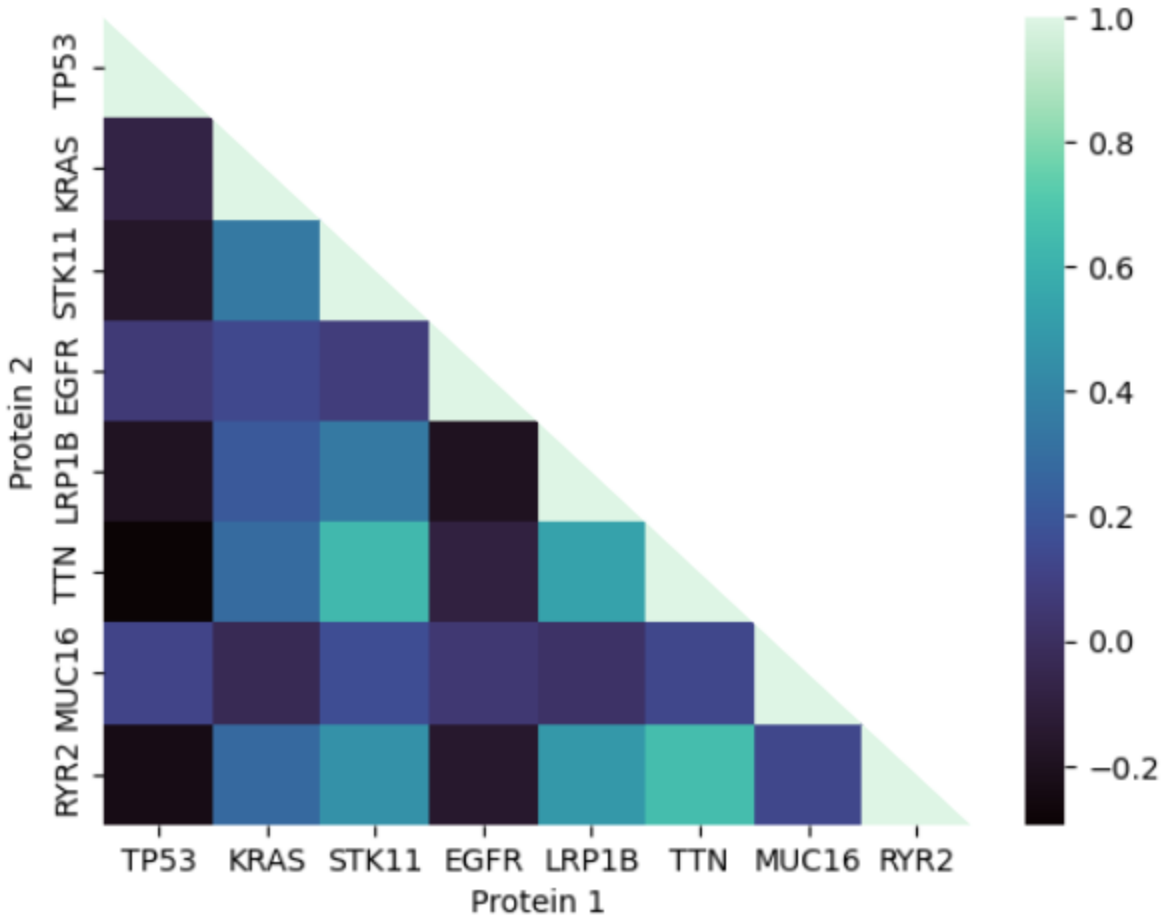

Figure 11. A heatmap showing strengths of correlations between protein expression for different combinations of commonly-mutated LA genes, indicating stronger correlations for protein-protein pairs including *LRP1B*, *TTN*, and *RYR2*. These correlation values are similar to the correlation values in the RNA/protein heatmap (Figure 10).

Upon returning to the RNA-RNA analysis and creating a new heatmap comparing only RNA expression for different pairs of genes (Figure 12), genes *LRP1B*, *TTN*, and *RYR2* once again maintained high correlation values relative to other gene pairings. The r-values for these three RNA-RNA pairings were around 0.4-0.6, which is lower than the r-values for the protein-RNA and protein-protein correlations. However, these correlation values still suggest a moderate linear positive correlation for RNA-RNA pairings for these three genes.

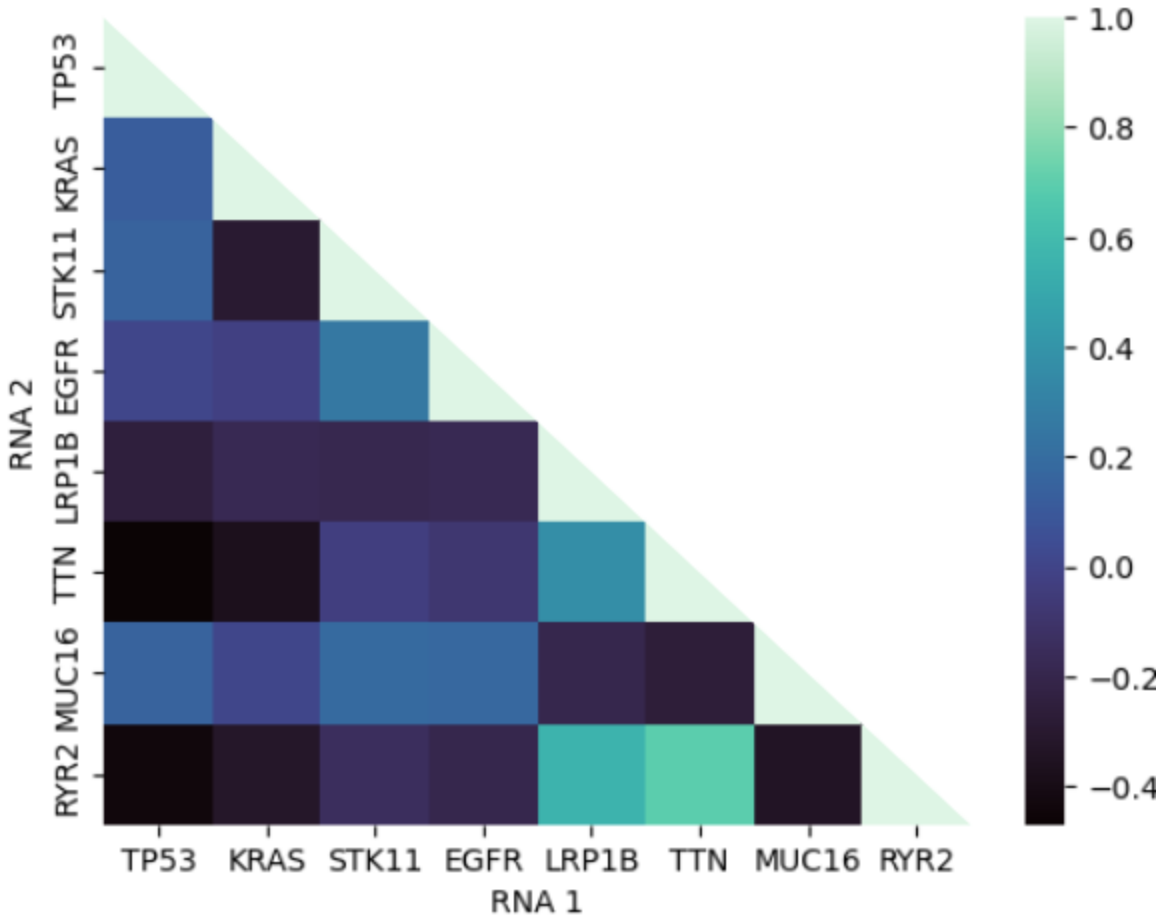

Figure 12. A heatmap showing strengths of correlations between RNA expression for different combinations of commonly-mutated LA genes, indicating stronger correlations for protein-protein pairs including *LRP1B*, *TTN*, and *RYR2*. These correlation values are a bit weaker than the correlation values in the RNA/protein heatmap and the protein/protein heatmap (Figures 10-11).

The RNA-RNA correlations displayed in the heatmap can be corroborated with a Draftsman plot of RNA counts for genes of interest (now including gene *CSMD3*, which was not available with CPTAC-downloaded data). Upon analyzing the plots for pairings with genes *LRP1B*, *TTN*, and *RYR2* (Figure 13), a faint positive trend can be made out. Although high RNA count outlier values appear for both axes on these plots, a general positive diagonal trend can be made out for these three plots, especially for genes *TTN* and *RYR2*. The RNA expression for these genes also have the highest correlation value in Figure 12. While it is challenging to discern whether there are major differences in transcriptomic expression for smokers and non-smokers based on the Draftsman plot, it is clear that these three genes have positive transcriptomic correlations.

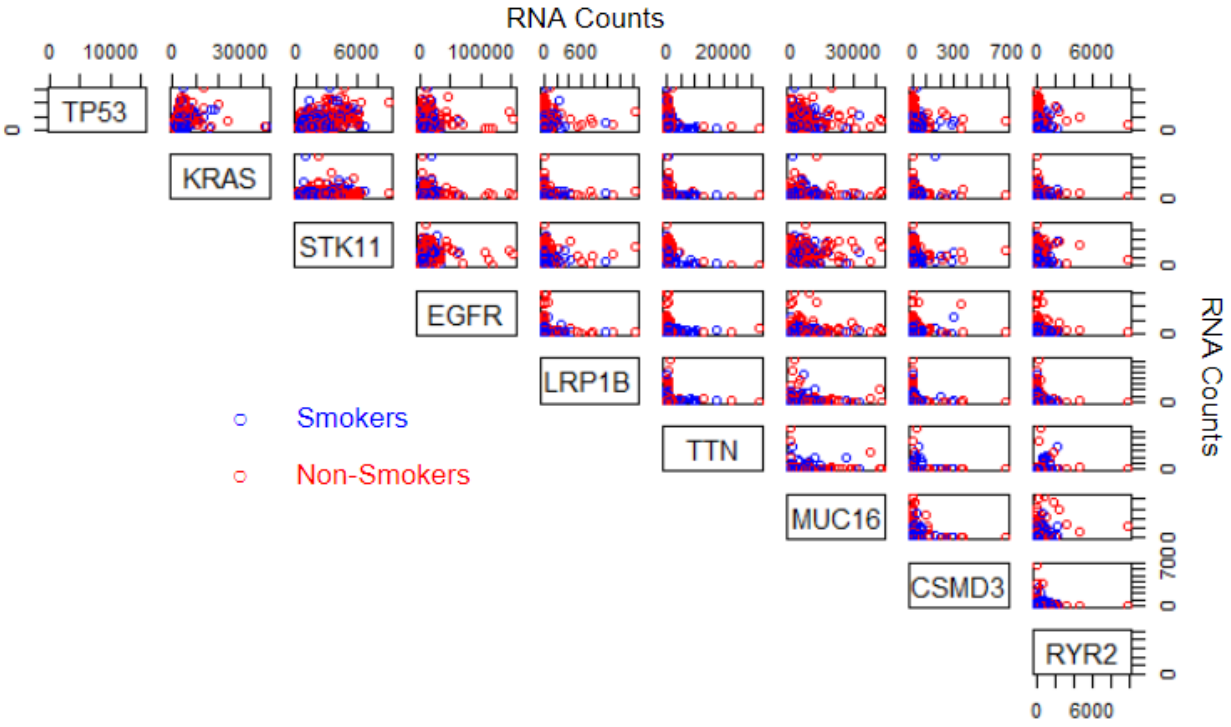

Figure 13. A Draftsman plot showing moderately positive linear correlations for RNA transcriptomic expression of pairings including genes *LRP1B*, *TTN*, and *RYR2*. While outliers on both axes worsen the correlation for the gene pair comparisons, positive linear trends can still be identified.

DISCUSSION

Preliminary analyses of clinical and demographic data suggest that much of what is understood about the LA disease state is unable to be confirmed by the available dataset. For instance, the knowledge that LA incidence is strongly positively associated with smoking status–with smoking being linked to a significantly earlier onset of the disease–is unable to be confirmed convincingly (Figure 1). This observation may be supported by the general distribution of the available dataset (Figure 2), which suggests that it follows a relatively uniform, normalized distribution with a slight left skew. As a consequence of low distributional diversity in the patient dataset, results concerning differentials between subpopulations (i.e. in disease incidence or mortality) will have inherently decreased resolution. Therefore, it may follow naturally that there is no differential survivorship observable between smoking and non-smoking patients (Figure 3), although this conclusion is in fact supported by existing literature, as previously discussed.

Broadly speaking, for NSCLCs in general, it is agreed that the survival of smoking patients is adversely affected in comparison to non-smoking counterparts. In one study, the median overall survival time for smokers was found to be lower than those of non-smokers (21.1 months vs. 41.9 months, respectively) to a statistically significant extent (p=0.027) (Lee, S. J. et al., 2014).

However, for LA indications specifically, as previously stated, there exists no widely-accepted relationship between the survivorship of smokers and non-smokers. In one study, the difference in survivorship between the two groups was found to be statistically significant (p=0.004), with Kaplan-Meier survival estimates of 16% for smokers and 23% for non-smokers at 5 years after diagnosis (Nordquist, L. T. et al. 2004). In such cases, smoking is understood to be an independent negative prognostic factor for LA survival. However, a contrasting study finds that smoking status at the time of diagnosis has little impact on the survival rate for patients with all NSCLCs, especially after surgery with curative intent. While there existed a minor trend toward an elevated risk of death for current versus never-smokers (hazard ratio, 1.20; 95% CI, 0.98-1.46; p = .07), this trend was eliminated upon adjustment for covariates (p=0.97) (Meguid, R. A. et al., 2010).

Whereas no significant difference was observed in the mortality of smokers versus non-smokers in the available dataset, secondary analysis of survivorship among male and female smokers indicates statistically significant results: within the smoking subpopulation, women face significantly more adverse survival outcomes (p = 0.062). Explanations for this conclusion are generally well-cited in literature, from both mechanistic and public health perspectives. Biologically, various studies attribute worsened survival in women to their unique pathways for metabolizing the carcinogens present in cigarette smoke (Mederos, N. et al., 2020). Genomically, female smokers have a higher expression of *CYP1A1* genes in the lungs than males, resulting in greater carcinogen activation, which may be induced in certain hormonal pathways (Uppstad, H. et al., 2011). With respect to DNA adducts and capabilities for repair, studies have shown that women have higher levels of DNA adducts than men, with increased levels of stable adducts believed to play a role in the initiation of carcinogenesis. Furthermore, preclinical data suggests that women have lower DNA repair capacity than men, with reduced capacity for repair being associated with an increased risk for lung cancer (Garm, C. et al. 2012). With the understanding

that LA incidence may be mechanistically different between female and male smokers, there is tremendous potential to devise novel therapeutics that target the aforementioned pathways in female carcinogenesis.

In the investigation of underlying mechanisms influencing the observed associations between smoking status, sex, and survival outcomes, we find that genomic analyses via oncoplots, co-oncoplots, and co-lollipop plots provide novel insights. First, we observe a unique mutational profile in the patients of our dataset in comparison to other large-scale studies. For example, genes we identified to have the highest mutational frequency in LA patients (*TP53*, *TTN*, *MUC16*, *CSMD3*, and *RYR2*, according to Figure 5) overlap only marginally with those identified by the Lung Adenocarcinoma Tumor Sequencing Project, which cites *TP53, KRAS, STK11, EGFR,* and *LRP1B* as the five most mutated genes (Greulich, H. 2010). Differential analysis of mutation frequency of the genes presented in Figure 5 reveals that, with the exception of the *RYR2* gene, mutational frequency is increased in the smoking subpopulation across all genes of interest (Figure 6). Within these, the association between smoking and increased *TP53* and *TTN* mutational frequency is well-documented, with *TP53* mutations serving as primary events for other microsatellite-stable metastatic tumors and *TTN* mutations disrupting spectrin α erythrocytic 1 (a key player in cell adhesion, cell-cell contact, and thus tumorigenesis) function through calmodulin 2 and troponin C1 (Yu, X. J. et al., 2019). Elevated mutation frequency of *TTN* in smoking versus non-smoking populations, especially in certain upstream and downstream loci (i.e. Proximal and Distal Ig-regions) (Figure 7) conveys an increase in the tumor mutational burden (TMB) and objective response to the immune checkpoint blockade, which if associated with response to chemotherapy, has potential to guide clinical decision-making (Xue, D. et al., 2021).

Regarding secondary genomic analyses involving sex, potentially clinical valuable insights can be made as well. Figure 8 demonstrates significantly higher mutation rates for *TTN*, *MUC16*, *CSMD3* in males compared to females. Increased mutational frequency in these genes is mentioned in reviews of other cancer types but is not well-studied in the context of LA. However, the existence of genes with differential mutation frequency, in combination with the unique mutational patterns in space observed across sexes (Figure 9), demonstrates that similar to our analysis of smoking status, there lies inherent value in sex-based therapies for LA given the disease's unique incidence and mortality across sexes.

Finally, synthesizing genomic and proteomic data, we consider RNA/RNA, RNA/Protein, and Protein/Protein correlations associated with genes of interest to identify relationships between potentially clinically relevant genes. Figures 10, 11, 12, and 13 illustrate moderately strong correlations between genes *LRP1B*, *TTN*, and *RYR2* in all omic levels, especially RNA/Protein, also confirmed by Figure 14. In a biological context, these strong correlations are largely corroborated by existing literature, which classifies genes *LRP1B*, *TTN*, and *RYR2* as those associated mechanistically with TMB. Studies in a closely related NSCLC, lung squamous cell carcinoma (LUSC), regard TMB to be a valuable independent indicator of favorable response to immunotherapy, giving genes associated with it important clinical implications (Xie, X. et al., 2021). While similar studies are not readily available for LA, the biological similarity between LA and LUSC suggests that each of these genes, either independently or in association with

another, may not only serve as important biomarkers for the disease, but also be valuable prognostic indicators with the potential to justify the use of various immune therapies.

Ultimately, we show that while the given dataset does not indicate that smoking and non-smoking LA patients present with statistically significant differential survival outcomes, other analyses, primarily genomic but also transcriptomic and proteomic, provide insights of potential clinical value. Namely, we find that the unique mutational profiles of men versus women and smokers versus nonsmokers offer opportunities for the development of more effective targeted therapies. Future directions for this project include contextualizing RNA and protein up/downregulation with a biological lens, conducting pair-wise correlation analysis to identify highly related genes in LA incidence, progression, and mortality, and utilizing machine learning techniques to ultimately quantify the importance of certain genomic and proteomic alterations in disease outcomes such as recurrence and survival.

REFERENCES

1. Ferketich, A. K., Niland, J. C., Mamet, R., Zornosa, C., D'Amico, T. A., Ettinger, D. S., Kalemkerian, G. P., Pisters, K. M., Reid, M. E., & Otterson, G. A. (2012). Smoking status and survival in the National Comprehensive Cancer Network Non-small cell lung cancer cohort. *Cancer*, *119*(4), 847–853. https://doi.org/10.1002/cncr.27824

2. Garm, C., Moreno-Villanueva, M., Bürkle, A., Petersen, I., Bohr, V. A., Christensen, K., & Stevnsner, T. (2012). Age and gender effects on DNA strand break repair in peripheral blood mononuclear cells. *Aging Cell*, *12*(1), 58–66. https://doi.org/10.1111/acel.12019

3. Greulich, H. (2010). The Genomics of lung adenocarcinoma: Opportunities for targeted therapies. *Genes & Cancer*, *1*(12), 1200–1210. https://doi.org/10.1177/1947601911407324

4. Harbeck, N., Penault-Llorca, F., Cortes, J., Gnant, M., Houssami, N., Poortmans, P., Ruddy, K., Tsang, J., & Cardoso, F. (2019). Breast cancer. *Nature Reviews Disease Primers*, *5*(1). https://doi.org/10.1038/s41572-019-0111-2

5. Hecht, S. S. (2012). Lung carcinogenesis by tobacco smoke. *International Journal of Cancer*, *131*(12), 2724–2732. https://doi.org/10.1002/ijc.27816

6. Lee, S. J., Lee, J., Park, Y. S., Lee, C.-H., Lee, S.-M., Yim, J.-J., Yoo, C.-G., Han, S. K., & Kim, Y. W. (2014). Impact of smoking on mortality of patients with non-small cell lung cancer. *Thoracic Cancer*, *5*(1), 43–49. https://doi.org/10.1111/1759-7714.12051

7. Mederos, N., Friedlaender, A., Peters, S., & Addeo, A. (2020). Gender-specific aspects of epidemiology, molecular genetics and outcome: Lung Cancer. *ESMO Open*, *5*. https://doi.org/10.1136/esmoopen-2020-000796

8. Meguid, R. A., Hooker, C. M., Harris, J., Xu, L., Westra, W. H., Sherwood, J. T., Sussman, M., Cattaneo, S. M., Shin, J., Cox, S., Christensen, J., Prints, Y., Yuan, N., Zhang, J., Yang, S. C., & Brock, M. V. (2010). Long-term survival outcomes by smoking status in surgical and nonsurgical patients with non-small cell lung cancer. *Chest*, *138*(3), 500–509. https://doi.org/10.1378/chest.08-2991

9. Myers, D. J., & Wallen, J. M. (2022). Lung Adenocarcinoma. In *StatPearls*. StatPearls Publishing.

10. Nagy-Mignotte, H., Guillem, P., Vesin, A., Toffart, A. C., Colonna, M., Bonneterre, V., Brichon, P. Y., Brambilla, C., Brambilla, E., Lantuejoul, S., Timsit, J. F., & Moro-Sibilot, D. (2011). Primary lung adenocarcinoma: Characteristics by smoking habit and sex. *European Respiratory Journal*, *38*(6), 1412–1419. https://doi.org/10.1183/09031936.00191710

11. Nordquist, L. T., Simon, G. R., Cantor, A., Alberts, W. M., & Bepler, G. (2004). Improved survival in never-smokers vs current smokers with primary adenocarcinoma of

the lung. *Chest*, *126*(2), 347–351. https://doi.org/10.1378/chest.126.2.347

12. Tomczak, K., Czerwińska, P., & Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, 19(1A), A68–A77. https://doi.org/10.5114/wo.2014.47136

13. Uppstad, H., Osnes, G. H., Cole, K. J., Phillips, D. H., Haugen, A., & Mollerup, S. (2011). Sex differences in susceptibility to pahs is an intrinsic property of human lung adenocarcinoma cells. *Lung Cancer*, *71*(3), 264–270. https://doi.org/10.1016/j.lungcan.2010.09.006

14. Xie, X., Tang, Y., Sheng, J., Shu, P., Zhu, X., Cai, X., Zhao, C., Wang, L., & Huang, X. (2021). Titin mutation is associated with tumor mutation burden and promotes antitumor immunity in lung squamous cell carcinoma. *Frontiers in Cell and Developmental Biology*, *9*. https://doi.org/10.3389/fcell.2021.761758

15. Xue, D., Lin, H., Lin, L., Wei, Q., Yang, S., & Chen, X. (2021). TTN/TP53 mutation might act as the predictor for chemotherapy response in lung adenocarcinoma and lung squamous carcinoma patients. *Translational Cancer Research*, *10*(3), 1284–1294. https://doi.org/10.21037/tcr-20-2568

16. Yu, X. J., Chen, G., Yang, J., Yu, G. C., Zhu, P. F., Jiang, Z. K., Feng, K., Lu, Y., Bao, B., & Zhong, F. M. (2019). Smoking alters the evolutionary trajectory of non‐small cell lung cancer. *Experimental and Therapeutic Medicine*. https://doi.org/10.3892/etm.2019.7958