The Effect of Race on Transcriptomic, Proteomic, and Clinical Data of Lung Cancer Patients

Peter Chong, Tessa Ferrari, Miracle Ini-Abasi, Brian Tinsley

Introduction

The Cancer Genome Atlas (TCGA) is a cancer genomics program founded by the National Cancer Institute and the National Human Genome Research Institute. It includes data collected from thousands of samples, spanning genomic, epigenomic, transcriptomic, and proteomic data. Similar to TCGA is the National Cancer Institute's Clinical Proteomic Tumor Analysis Consortium (CPTAC). CPTAC makes use of proteomics and genomics to give researchers a better understanding of cancer diagnosis, treatment, and prevention. We used data from both TCGA and CPTAC in our analysis of lung adenocarcinoma.

Lung cancer is not only the most prevalent cancer, but it is also the leading cause of death in the world (de Alencar, 2020). Lung cancer is an extremely dangerous cancer that starts in the lungs and often occurs among avid smokers, although it can also take effect in non-smokers (Dela Cruz, 2011). In 2022 alone, the American Cancer Society (AMS) predicts that there will be around 236,740 new cases of lung cancer, with almost half resulting in deaths.

Since lung adenocarcinoma is such a prevalent issue, it is important to understand different biomarkers which can give information about the genetic, transcriptomic, proteomic, and clinical factors that can increase the chances of developing this cancer. Our research question looks at the relationship between lung adenocarcinoma and race; more specifically, we examined the differences in patient outcomes of lung adenocarcinoma based on race. In addition, we looked at whether this is genetically predisposed based on race or whether there are different patient treatments in terms of time and diagnosis that influence patient outcomes. For research in

the biomedical field, race is classified into the following categories: non-Hispanic White, non-Hispanic Black, Hispanic, Asian/Pacific Islander, American Indian/Alaskan Native (Schabath, 2016). This naming convention slightly differs from that of TCGA. TCGA groups race as Asian, Black or African American, White and native. However, in this experiment we chose to look specifically at the Asian, Black and White categories as there was not sufficient data for the native category.

Studies have shown that lung adenocarcinoma incidence differs based on race, with Black patients having the highest incidence of lung adenocarcinoma as well as the lowest survival rate. On the other hand, Hispanics have the lowest incidence rate (Patel et al, 2016). Research has linked this to the striking contrast in treatment, as White patients are over 50% more likely to receive prompt surgery, chemotherapy, and radiation than Black patients (Schabath, 2016).

Multi-omic data analysis is interesting for this experiment as it allows lung adenocarcinoma to be studied on different levels. This new data approach makes use of genomic, transcriptomic, and proteomic methods to provide extensive information on the changes seen in different races. For instance, multi-omics can shed light on why lung adenocarcinoma is more prevalent in Black patients and why they have a lower survival rate through the use of statistical analysis. First, we examined the genetic differences between patients of the different races and based on these differences examined the survivability using clinical data plots. Finally, we were able to see how these results compared to the research found in previous studies.

Methods

In order to determine if there were overall differences in the diagnostic timeline for patients of different races, we compared their age of initial diagnosis using a boxplot. To do this,

we collected clinical data from TCGA for lung adenocarcinoma (LUAD) patients and eliminated any patients with an unlisted race from our data set. Of the 617 patients in the clinical data, 82 had no reported race, 465 White patients, 59 Black patients, 10 Asian patients, and one American Indian patient who was left out of this analysis due to lack of data. Then we used R to create a boxplot to compare the initial age at diagnosis based on patient race, which shows the minimum, maximum, median, lower quartile, and upper quartile ages from the data.

Next, we created a Kaplan-Meier survival plot (KM plot) in order to see how the estimates for survival compare based on race. We again used TCGA clinical data for lung adenocarcinoma patients. To create the 'days to death' dataset for the KM plot, if any patient was missing data, we replaced it with known 'days to last known alive' data or 'days to last follow up' data. Then we stratified the chart by race to see their survival probabilities. Patients who identified as Native American were excluded from the chart due to a lack of patient data.

We used the same TCGA clinical data and used the stage column to create a bar graph showing the proportion of White and Black patients diagnosed at each stage of cancer, ranging from I to IV.

Next, we created a mutation allele frequency (MAF) object which contains mutation (and clinical) data for the patients in the TCGA data. Of the 507 patients contained in the MAF object, 442 had race data. Of the 442 patients with data, there are 380 White and 52 Black (or African American). There were 8 Asian patients and 1 American Indian or Alaska Native patient, but they were left out of this analysis due to lack of data. We then created an Oncoplot which showed the 10 most highly mutated genes in all 507 patients; this plot was used as a reference when determining genes of interest for the next steps in the analysis. Using the race data, we were able to split the MAF object into separate objects for White and Black patients. Using the newly

created White and Black MAF objects, we then generated a cooncoplot showing the differences in frequencies and types of mutations for specific genes.

Once the cooncoplot was created, we used the R function “lollipop2” to create lollipop plots for White and Black patients for 4 of the genes in the generated oncoplot. Lollipop plots reveal the mutation types and frequencies at selected genes (*TP53, TTN, MUC16*, and *CSMD3)* between White and Black patients.

After creating the lollipop plots, we were interested in RNASeq data, which examines the number of counts each gene is expressed. After installing the package “SummarizedExperiment” and completing the proper query and data download, we were left with a transcriptomics data frame. In this data structure, we can access clinical data through the “colData” of the transcriptomic summarized experiment dataframe, as well as specific gene information through the “rowData” on the same object. Before completing the analysis, we had to first remove the patients who did not report their race. Of the 598 patients in the data set, 67 did not have race reported. Similarly to the MAF clinical data, the overwhelming majority of patients identified as Black and White, so Asian and American Indian patients were removed from the transcriptomic analysis. Now that the data had been cleaned, we accessed the counts via the “unstranded” column of the transcriptomic summarized experiment dataframe. Again using boolean indexing, we were able to filter out the counts for the *TP53*, *TTN*, *MUC16*, and *CSMD3* genes. Once we split up the counts by gene, we created boxplots to show the minimum, maximum, median, lower quartile, and upper quartile of RNASeq counts for each gene between White and Black patients.

In order to compare how protein and gene expression data correlate between different races, we created three heatmaps: one for White patients, one for Black patients, and one for Asian patients. To do this we collected clinical, transcriptomic, and proteomic data from CPTAC

and created subsets of the transcriptomic and proteomic data based on the race found in the clinical data set. Then we created heat maps comparing the protein and gene expression of the top 10 most mutated genes in the overall dataset (which were identified by an oncoplot from TCGA data). However, one of the genes (*CSMD3*) lacked information in the CPTAC dataset, so it was not included in the final heatmaps.

Results

Figures 1 through 3 analyze clinical data seeking to examine differences in populations at time of presentation and diagnosis between races, specifically through examining the time at which populations obtain their initial diagnosis, survival curves and stage at initial diagnosis.

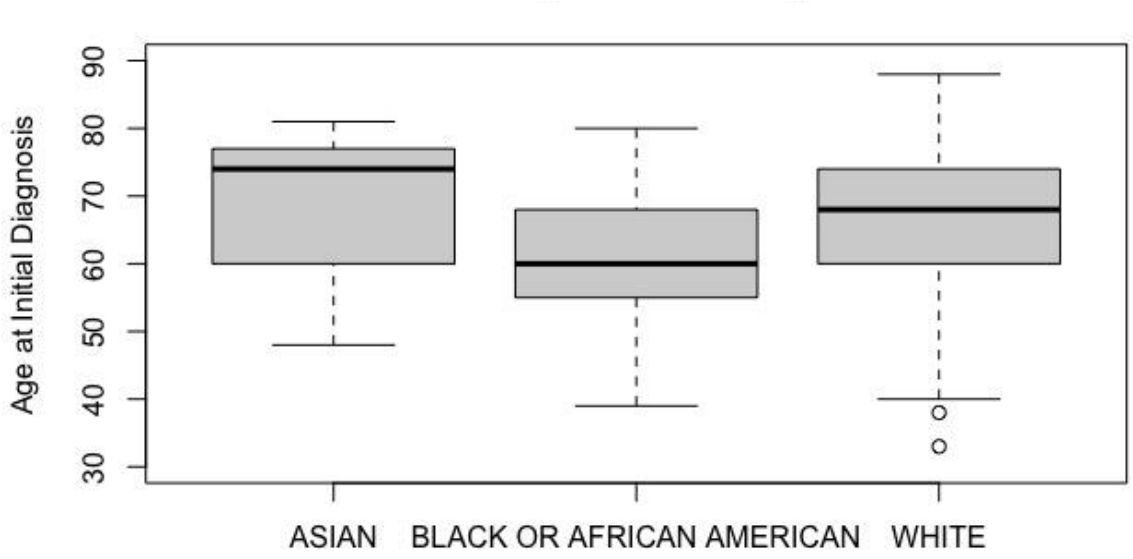

Figure 1. Differences in Age at Initial Diagnosis between Asian, Black, and White Patients

Black populations have the youngest median, followed by White populations and then Asian populations. The Asian sample had n of 10, the Black sample had n of 59, the White sample had n of 465.

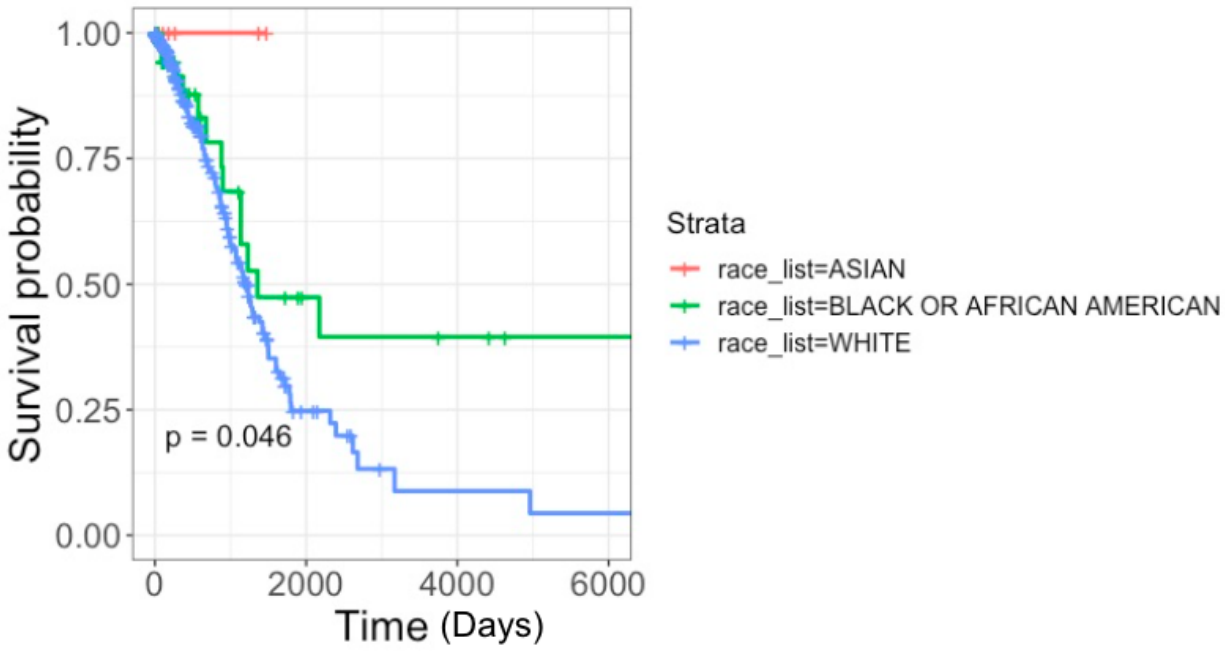

Figure 2. A Kaplan-Meier Survival Curve between Asian, Black, and White populations.

Asian populations appear to have insufficient data for any conclusions. Shortly after initial diagnosis, the White patients have lower survival rates than Black patients.

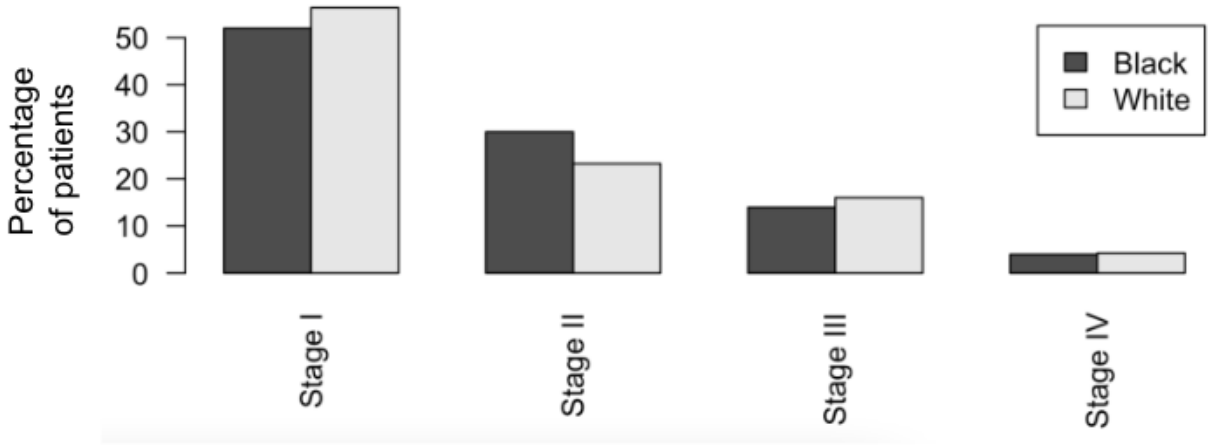

Figure 3. a) Cancer Stage at Diagnosis between Black and White Patients

A comparative bar graph showing the proportions of the stage of lung adenocarcinoma at which each population was diagnosed. White patients had an n of 374 and Black patients had an n of 50.

Interestingly, the survivability of White and Black patients is similar in early stages. However, the overall survivability of White patients is much lower than Black and Asian individuals after about 1000 days (Fig. 2). We also see that Black patients are diagnosed at an earlier age than others (Fig. 1). In Fig. 3, we can see that a greater proportion of Black patients are diagnosed with stage II cancer than White patients. One thing to note is that starting with Fig. 3, many of the figures lack sufficient Asian data, which resulted in the unfortunate need to drop their data from the analysis.

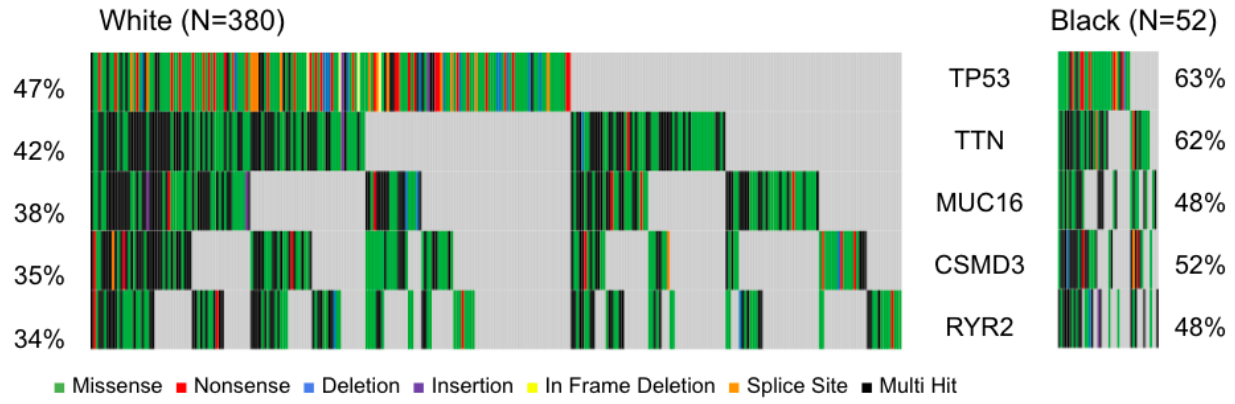

Figure 4. Cooncoplot between Black and White Patients

The figure demonstrates that a higher mutation frequency is prevalent in Black samples across all of the top 5 genes, *TP53*, *TTN*, *MUC16*, *CSMD3*, and *RYR2* by a considerable amount. White patients

Fig. 4 reveals the mutation types and frequencies in the top five most mutated genes for White and Black patients. Genes of note are *TP53*, which has a 16% difference in mutation frequency between groups, *TTN* with a 20% difference, *MUC16* with 10%, and *CSMD3* with a 17% difference. Interestingly, the mutation rate is significantly higher in Black patients for all genes.

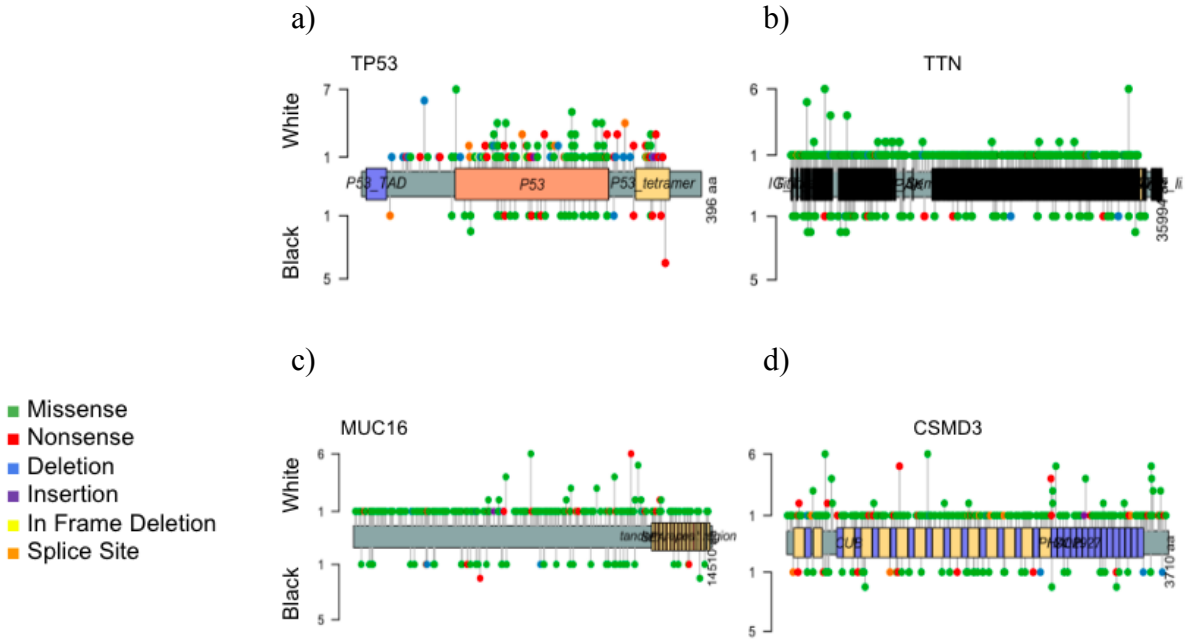

Figure 5. Lollipop Plots of the Top Four Genes

Four lollipop plots showing the mutation type and frequency for a) *TP53*, b) *TTN*, c) *MUC16*, and d) *CSMD3*. White patients had an n of 380 and Black patients had an n of 52.

In Fig. 5, the specific locations and types of mutations on each of the selected genes' proteins are shown. a) White patients have many more mutations at the front end of the TP53 protein. Additionally, here are more frame-shift mutations in White patients. b) Despite having a lower overall mutation frequency, the mutations for White patients are much more spread out across the entire *TTN* gene, as shown by the mutation locations on the TTN protein. In Black patients, we can see a higher number of nonsense mutations. c) The lollipop plot for MUC16 shows a similar pattern, with White patients having a broader range of mutation locations, but an overall lower mutation rate. d) In the CSMD3 lollipop plot, there are a handful of frame-shift mutations toward the end of the protein in Black patients. Frame-shift mutations are of particular interest because they can be very damaging to the encoded protein structure.

Now that we have seen the specific gene mutation differences, we are interested in the expression of these genes. Fig. 6 shows the RNASeq counts for each selected gene by race.

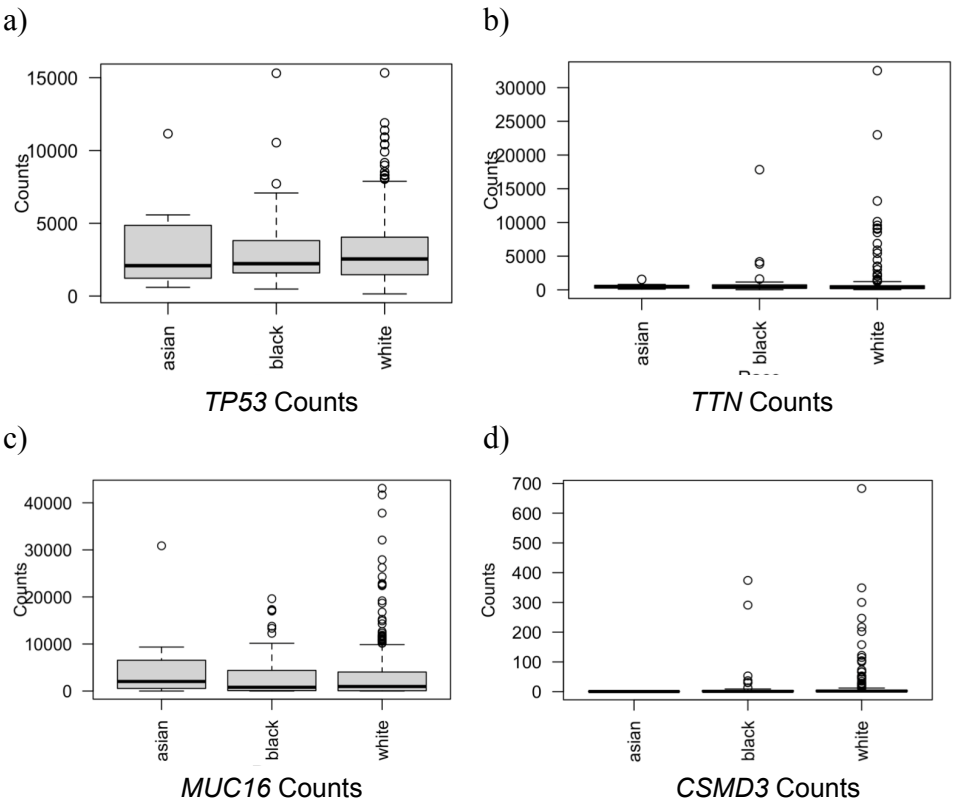

Figure 6. Boxplots for *TP53*, *TTN*, *MUC16*, and *CSMD3* Counts by Race

Four boxplots showing the transcriptomic counts for a) *TP53,* b) *TTN,* c) *MUC16,* and d) *CSMD3* genes.

White patients have slightly higher expression of *TP53*, as well as a large number of upward outliers. Three Asian patients also had outliers. For *TTN*, expression rates were lower overall, but expression medians were similar across all groups and again White patients had many outliers. For *MUC16*, Asian patients had the highest median expression level, but white patients had many upward outliers that outweigh the total of eight patients in the Asian group.

Expression for *CSMD3* is lower, only reaching a maximum of about 700 counts. The next step in the analysis is looking at heatmaps which connect RNA expression versus protein abundance.

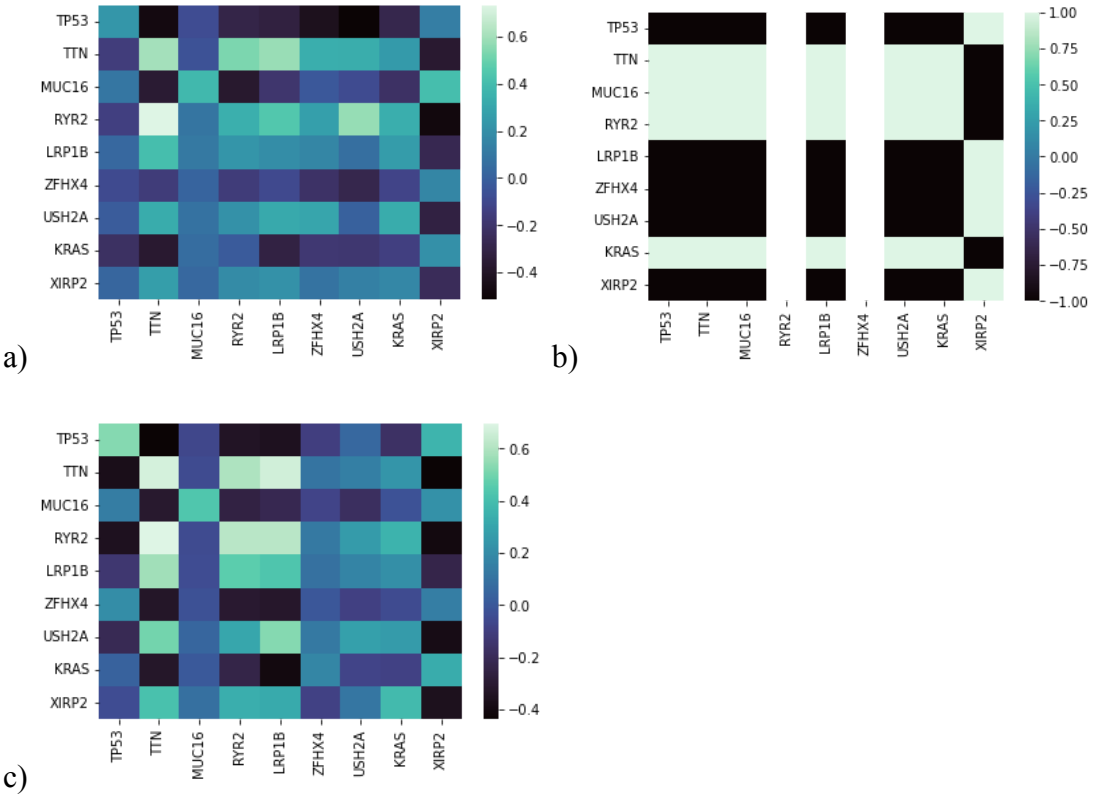

Figure 7. Heatmaps of RNA Expression between Races

Heatmaps demonstrating RNA expression versus protein correlation in a) White, b) Black, and c) Asian populations. The top 10 most mutated genes in the lung adenocarcinoma dataset were chosen for comparison (excluding *CSMD3* due to lack of data). The b) heatmap has a lack of patients, therefore resulting in a limited heatmap that shows little gradation.

The heatmaps connect RNA expression versus protein abundance. In doing so, after looking at genes and proteins separately, the connection between them through RNA expression is ensured. Between the White and Asian heatmaps, the general pattern is similar, although there are some differences—notably the second, third, and fourth column being lighter in Fig. 5c than in Fig. 5a. It is important to note that although the rest of this paper has focused more on

comparing Black and White populations, namely due to a lack of sample size in Asian populations for sufficient analysis, there were not enough Black patients for an adequate heatmap (Fig. 7b).

Discussion

The purpose of this study is to determine differences in patient outcomes based on race. These differences could be genetic or from a general standard of care in medical institutions. The goal is that findings from these research methods might illuminate genetic differences between races that have yet to be noted, or that we can shed light on differences in the medical care provided to different races. Either way, we hope to help guide cancer research in a direction to increase survivability as well as equality for all people, regardless of their background.

Within clinical data, Black populations are diagnosed at an earlier median age (Fig. 1). Coupled with the MAF analysis in Fig. 5, where it is shown that Black samples experience a higher number of mutations across the top mutated genes (and therefore most likely across mutated genes in general), we concluded that black populations likely experience more severe lung cancer. It would therefore follow that Black populations therefore probably experience an earlier onset of lung cancer. This would explain the earlier diagnosis rate of Fig. 1. Of course, this connection between these two distinct data observations should be researched further to determine whether it is correct, and if so, to detail the causal link.

The survival curves demonstrate an initial decline in survival rate but a higher eventual survival in Black individuals (Fig. 3). In other words, White populations, shortly after the initial diagnosis, show decreased survivability. This is likely due to the overall higher age of diagnosis shown in Fig. 1. Considering that the White patients are typically older, they have a considerably

lower five year survivability already. This is especially in light of other sources that say that Black individuals have a lower overall survival rate across 1-year, 3-year, and 5-year periods (Shi et al.) To further explore this possible explanation, survivability curves between ages, as well as patients that died of cancer-related causes versus non-cancer-related causes, should be created. Stratifying this data against race would further help connect those curves back to lung cancer between races.

Furthermore, a higher proportion of White individuals are diagnosed at Stage I than Black individuals, indicating that White patients are diagnosed earlier in the progression of their cancer (Fig. 2). Of course, while data involving diagnosis does not directly translate to medical care, it does imply that White patients have better access to medical care, such as better doctors, better equipment, higher frequency of appointments, or subconsciously more detailed care being put into their treatment. Further studies could be conducted to more directly link these social factors to lung adenocarcinoma screening rates.

The lollipop plots in Fig. 5 show us specific mutation locations, but also, generally, that Black patients have a less broad range of locations for mutations. This could be attributed to the smaller population size. In future studies, we could look further into these differences in mutation frequency by looking at more clinical variables other than just race; it is possible there is some other clinical variable that is affecting these mutation frequencies. Furthermore, social factors could be investigated for their impact on mutation location and frequencies.

The boxplots in Fig. 6 showed very limited observations about RNASeq expression. The most common pattern between the four boxplots is that White patients have many outliers in the upward direction, which are most likely attributed to the greater population size. It would be interesting to look at only the outliers for White, Black, and Asian patients and determine what

overlapping factors contribute to their expression levels being so high. Specifically, patients' similar (or dissimilar) mutation locations and frequencies could be explored. Aside from noting the higher expression of *MUC16* in Asian patients, it is hard to gain much insight about Asian patients due to the very small population size. In general, it might be interesting to look at the overall average expression rates and each of these genes specifically. For example, the highest count for *CSMD3* was at about 700, whereas for *MUC16*, the maximum count was closer to 45,000. Therefore, the differences in counts between genes should also affect how each of these genes are explored in the future.

The heatmaps between races have a lack of data in Black patients, while Asian patients have a relatively large sample size. Since this paper mainly focuses on the comparison between Black and White patients, the heatmaps unfortunately do not lead to meaningful relevant information.

This study mostly highlights areas for further research, especially when it comes to getting more data and equal sample sizes across different groups, but also in examining causal links between social factors and the observed discrepancies between racial groups. Perhaps there is something going on at a more basic level that we are overlooking, such as people of different racial groups not wanting to share their information as frequently as others. Either way, finding a way to get more data is imperative in the study of the effects race has on outcome and survival, as well as on genetic factors.

References

de Alencar, V., Formiga, M. N., & de Lima, V. (2020). Inherited lung cancer: a review. *Ecancermedicalscience*, *14*, 1008. https://doi.org/10.3332/ecancer.2020.1008

Dela Cruz, C. S., Tanoue, L. T., & Matthay, R. A. (2011). Lung cancer: epidemiology, etiology, and prevention. *Clinics in chest medicine*, *32*(4), 605–644. https://doi.org/10.1016/j.ccm.2011.09.001

Lung cancer statistics: How common is lung cancer? (2022, February 14). Retrieved April 26, 2022, from https://www.cancer.org/cancer/lung-cancer/about/key-statistics.html

Patel, M., Wang, A., & Kapphan, K. (2016). Racial and Ethnic Variations in Lung Cancer Incidence and Mortality: Results From the Women's Health Initiative. Retrieved April 24, 2022, from https://ascopubs.org/doi/10.1200/JCO.2015.63.5789

Pennycuick, A. (2019, December 4). On the Origins of Lung Cancer. Retrieved April 24, 2022, from https://www.atsjournals.org/doi/full/10.1164/rccm.201911-2176ED

Schabath, M. B., Cress, D., & Munoz-Antonia, T. (2016). Racial and Ethnic Differences in the Epidemiology and Genomics of Lung Cancer. *Cancer control : journal of the Moffitt Cancer Center*, *23*(4), 338–346. https://doi.org/10.1177/107327481602300405

Shi, H., Zhou, K., Cochuyt, J., Hodge, D., Qin, H., Manochakian, R., Zhao, Y., Ailawadhi, S., Adjei, A. A., & Lou, Y. (2021, December 10). Survival of black and white patients with stage IV small cell lung cancer. Frontiers. Retrieved May 2, 2022, from https://www.frontiersin.org/articles/10.3389/fonc.2021.773958/full

The cancer genome atlas (TCGA). (2020, December). Retrieved April 24, 2022, from https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas

Centers for Disease Control and Prevention. (2016). Current cigarette smoking among US adults aged 18 years and older. Retrieved by https://www. cdc. gov/tobacco/campaign/tips/resources/data/cigarette-smoking-inunited-states. Html.

Acknowledgments

Ellison (Dr. Jerry Lee), QBIO Department (Dr. Rohs, Dr. Calabrese, and Katie), and Nicole,

David, and Kate.