2021

# The Impacts of External Patient Variables on Gene Expression Profiles of Lung Adenocarcinomas

Michael Madsen

# The Impacts of External Patient Variables on Gene Expression Profiles of Lung Adenocarcinomas

A Thesis Presented

by

Michael Madsen

To the Keck Science Department

Of Claremont McKenna, Pitzer, and Scripps Colleges

In partial fulfillment of

The degree of Bachelor of Arts

Senior Thesis in Biology

May 3, 2021

# Table of Contents

## Abstract

The search for improvements to detection and treatment of cancers is a paramount goal for all of medicine. The most important step for oncological research is to expand the knowledge base of the genetic characteristics and abnormalities that give rise to cancer. In our present day, one of the most pressing and deadly forms of cancer is that of the lung, with lung adenocarcinomas being the most prevalent variation of the disease. Improving our cancer genomic insight can provide the seedings for improved cancer detection, novel cancer treatment, and serve as a guide for avenues to explore with future oncological research. Using The Cancer Genome Atlas (TCGA), a cancer genomics program from the National Cancer Institute, a data set of 512 lung adenocarcinoma samples was compiled for analysis into the interconnectivity of patient characteristics and their impact on gene expression profiles. For this sample population, patient sex, age, race, smoking history, and tumor stage were analyzed using a permutational multivariate analysis of variance (perMANOVA) modeled with non-metric multidimensional scaling (NMDS) to determine significant interactions between the variables on tumor genetic differentiation. Patient smoking history and tumor stage, sex and tumor stage, and sex and age were all found to have statistically significant impacts on gene expression, while race on its own was also found to have a significant impact. This analysis highlights that male and female cancers might differentiate quite differently and illuminates a need to explore sex related differences in cancer progression. Additionally, the research emphasizes the continued buildup of mutations throughout a cancer's proliferation, showing the need for investigation into cancer development after it has been initially discovered. Lastly, this research demonstrates that patient variables interact significantly to impact gene expression profiles and accentuates the need for research about how external factors combine and interconnect to make each cancer unique.

## Introduction

Curing cancer has been an ultimate goal of medical research, attracting ideas crossing many disciplines and systems (Nakamura et al., 2016). However, a cure for cancer is a challenging target to reach due to the difficulties of detection and treatment along with the immeasurable complexity of cancer itself (Auyang, 2006). Cancer at its core is chaotic, developing from the accumulation of random mutations that eventually override and overwhelm the cell's anti-cancer defense mechanisms, leading to uncontrolled cell growth and division (Weinberg, 1996). While we do have an understanding of the core mechanisms that work to prevent cancer and the elements of normal cell functioning that go awry when cancer arises, the extraordinary complexity of life inherently means that every cancer is unique (Hartmaier et al., 2017). Although cancers have many common similarities which have formed the basis for many contemporary cancer treatments, gaining a complete genetic understanding of all cancers is the most essential step in the development of novel cancer therapies (Heim et al., 2014). Beyond simply enhancing our knowledge base of cancer genetics, increased genomic analyses could help us recognize the warning signs of cancer earlier, better prepare for new cancer patterns that arise, and most importantly, generate more personalized cancer therapies to better fight back against a patient's unique cancer variant (Chin et al., 2011).

Safer and more effective treatments are a primary objective for biomedical research, especially in oncology, so improvements to our current treatment methods are paramount. Generalized cancer therapies, such as radiation therapy or chemotherapy, impose a great toll on patients (Aslam et al., 2014). These treatments work to destroy cancer cells, but collaterally damage healthy cells (Baskar et al., 2014). Targeted therapies allow for a massive step forward in tumor therapy, as specific cancer cell biological processes are targeted by the drug. However, many cellular processes are so complex our still limited understanding and biotechnological capabilities mean that we can only target specific cell mechanisms (Suda & Mitsudomi, 2014). Moreover, cancer cells might become resistant to these therapies over time and cancers can develop with multiple cellular abnormalities, meaning targeted therapies cannot be the sole treatment in most cases (Groenendijk & Bernards, 2014).

This is where the importance of personalized therapies reveals itself. Only around 5 to 10 percent of cancers come from inherited genetic mutations (Riley et al., 2012). For these patients, genetic testing and early targeted therapies can be a crucial tool for attacking their cancers. However, this still leaves out the large majority of patients whose cancers developed from random mutations over their lifetime. As our genomic knowledge base increases, we will begin to discover more and more abnormalities in each cancer variant and slowly develop treatments for these abnormalities (Moreira & Eng, 2014). With more treatment options available, doctors can prescribe personalized treatment plans that attack the exact mutinous pathways in a specific patient's cancer, increasing treatment efficacy, improving quality of life, and prolonging patient lifespan (Friedman et al., 2015). Unfortunately, we are still a long way away from having a sufficient stockpile of treatments that can be personalized to every patient, but the first step in getting there is analyzing the genomics of as many cancer variants as possible and discovering patterns for the foundation of future therapies.

This paper's analysis attempts to explore a tiny fraction of the sprawling macrocosm that is cancer research by investigating the genomics of lung cancer. When choosing a primary focus for cancer investigation, no origin is more prevalent than cancer in the lung.

Based on statistics compiled by the SEER Cancer Statistics Review, there are estimated to be over 235,000 new cases of various forms of lung cancer every year in the United States alone. It is the second most prevalent form of cancer for both men and women and is primarily diagnosed in patients over the age of 65. Furthermore, it is also one of the deadliest forms of cancer, with a 5-year survival rate of only 18.1 percent (Howlader et al., 2020). This means lung cancer is the leading cause of death amongst all cancers, with over 130,000 patients succumbing to the disease every year in the US (World Health Organization, 2021). All of this signifies that lung cancer is a top priority for cancer researchers, doctors, and patients alike.

In addition to being categorized by the organ of origin, cancer is also classified by the type of cell it originates from (Travis et al., 2013). For lung cancer, this subsequent breakdown begins by classifying the cancer as small cell, non-small cell, or carcinoid tumors. Carcinoid tumors are quite rare and originate from the lung's neuroendocrine cells. Small cell lung cancers are more common, making up a bit less than 20% of all total lung cancers, and typically originate from bronchi cells. However, far and away the most common type of lung cancer is non-small cell lung cancer (NSCLC), which accounts for over 80% of all lung cancers. While not as fast growing as its small cell counterparts, NSCLCs often go undetected and don't show symptoms until it has reached an advanced stage (Markman, 2021).

Being such a large section of lung cancers, NSCLC is typically broken down further by the tissue it arises from (Travis et al., 2013). Undifferentiated carcinomas account for 10 to 15 percent of NSCLC while squamous cell cancer accounts for about 30 percent of NSCLC. However, the most prevalent form of NSCLC, and the most common type of lung cancer overall, are lung adenocarcinomas. These cancers arise in the mucus secreting glands on the outside of the lung and account for over 40 percent of all NSCLC cases. As such, increasing our knowledge about the fundamental mechanisms that lead to the emergence of lung adenocarcinomas is imperative (Markman, 2021).

When attempting to understand the fundamental aspects of cancers, the most important investigative goal is finding the genetic differences that make each cancer unique (Hartmaier et al., 2017). As mentioned previously, only 5 to 10 percent of cancers are caused by uncontrollable genetic mutations, so pinpointing the external factors that promote cancer formation can not only help us generate new cancer therapies but is also key to improvements in early cancer detection and the encouragement of cancer-preventative lifestyle changes. Proactivity towards attacking cancer before it becomes insurmountable is paramount and goes hand-in-hand with treatment therapies (Valle et al., 2015). If our ultimate oncological goal is cancer treatment personalized to the patient, we need to have a comprehensive understanding of how external and internal factors manifest individual-specific tumor genomics.

For lung cancer, the most obvious of these external factors is smoking. The impacts of smoking are well documented and today it is considered a widely accepted medical truth that smoking directly and severely increases your chances of developing lung cancer, among other potential diseases (Walser et al., 2008). Based on a study by the World Health Organization, cigarette smoking is expected to cause 10 million fatalities per year, and most of those are related to lung cancer (Proctor, 2001).

Like other lung cancers, lung adenocarcinomas are too significantly correlated to previous smoking (Brownson et al., 1987). However, lung adenocarcinomas are also the most common type of lung cancer found in people who have never smoked, bringing into question the other risk factors that lead to such a prevalence among the non-smoking populous (Myers & Wallen, 2021). While asbestos, radon, heavy metals, and diesel have also been linked as risk factors, these risk factors are rare and are often difficult to study and measure throughout the average cancer patient's life.

For many patients, oftentimes there is no clear external catalyst for their cancers. This is especially true for the lung adenocarcinoma patients who develop the cancer without previous smoking, highlighting the importance of the rest of the genome for cancer susceptibility. Consequently, we are forced into examining the patient themselves to attempt to understand their cancer. While any two humans are 99.9 percent genetically similar this still leaves approximately 3 million base pair differences throughout the entirety of the human genome (Hernandez et al., 2006). It is highly possible that many of the differences in tumors stem from these 3 million base pair differences. Furthermore, while each patient's encoded genetic differences might only be catalysts for a small portion of cancers, the genomic differences from a variety of patient factors likely interact differently to generate unique cancer variants within each patient (Campa et al., 2011).

Research has been done into the interconnectivity between specific genes and their interactions leading to cancer development (Y. Li et al., 2019). However, these studies only look into the interaction of pinpointed genes. In these narrow lensed genetic interaction analyses, misidentification of certain genetic ties and the overlooking of others are not uncommon (Wu & Ma, 2019). To better comprehend intra-genomic influences, we need to broaden our view into robust wholescale genetic interaction analyses and the relation to discernable external factors that create these genetic individualities.

Any particular risk factor or patient feature on their own can tell us a lot about differences in tumor genetics and development, but these impacts are superficial and don’t serve to help the ultimate goal which is personalized medicine (Yoshino & Maehara, 2007). For any individual patient, they cannot be quantified by their age, smoking status, or any other characteristic. The patient and their cancer are a sum of hundreds of thousands of different factors, and to fully grasp how cancer arises and how it might be treated and prevented we need to understand how these factors interact to see beyond the impacts of any one particular trait.

For this preliminary investigation into these complex interplays, we must start by looking at the relationships of the clearest factors available to us. Patient demographics offer perfect variables for analysis as they are not only recorded for nearly every sample taken from the hospital, but the findings from investigating these characteristics can be applicable to every future cancer patient (Dick et al., 1997). Not only do simple demographics such as sex or age provide a clear stratifying attribute for analysis, there are clear links to genetic differentiations in tumors for each of these characteristics (Lopes-Ramos et al., 2020). The next step is to widen our view and connect these genetic differentiators to each other.

For the purpose of this analysis, patient sex, age, race, smoking status, and the stage of the patient’s tumor were selected for investigation. Although there are many other factors that also contribute to the impacts found from any one demographic, these characteristics offer the clearest and most well-defined variable to be used in the designation of subgroups for research, prognosis, and treatment (Zahm & Fraumeni, 1995). They all also have clear ties to differential tumor genetic expression (Patel et al., 2010).

Based on what we know about the interconnectivity of the human genome and the impact of external factors on tumor genetics, we hypothesize that there are some factor interactions that drive tumor genomic differentiation. The interdependence of the human genome inherently dictates there to be multiple causative elements to any singular genetic change (Barillot et al., 1999). This, coupled with the well-studied influence of outside elements on tumor gene profiles means it is most likely that these elements interconnect to alter gene expression (Colak et al., 2013).

## Methods

The data used in this analysis comes from The Cancer Genome Atlas (TCGA). An endeavor from the National Cancer Institute, TCGA is a cornerstone program for the consolidation of cancer genomic and other genetic information. TCGA has sequenced over 20,000 tumor and normal tissue samples from 33 different types of cancer and has generated over 2.5 petabytes of genetic data. The publicly available data has been analyzed countless times across the medical sphere to restructure the way we classify cancers, bolster researcher's understandings of the biochemical underpinnings of cancer, and begin to identify those vital genomic patterns that targeted, personalized therapies will be directed at (Weinstein et al., 2013).

The processes involved in obtaining, sequencing, and processing tumor data differs tremendously across all different institutions and for all different cancers, but the general structure for all cancer analysis is somewhat uniform. The first layer of tumor data is collected when clinical information about the patient is documented. Biopsy samples of the patients' tumors along with samples of the normal tissue are then extracted and stored for further analysis. These samples are eventually processed into molecular analytes that can be used for gene sequencing, protein expression profiles, and other analytic processes.

While each lab has different analysis procedures and uses different analysis products and software, most tend to follow a similar outline. For example, Kratz Lab in the University of California, San Francisco begins by storing samples taken from the operating room in liquid nitrogen before transferring into longer term cold storage in the lab. RNA extraction is performed using Qiagen miRNeasy Mini kit and RNA quality is assessed using Agilent RNA 6000 Nano kit on Bioanalyzer. Subsequently, library prep is performed with Illumina Stranded total RNA method with Ribo-Zero Plus and sequenced on Illumina Nextseq 500. Finally, this sequencing data is sent out to a database like TCGA to be compiled with similar sequencing data from around the country (Mendez et al., 2017). All metadata collected through TCGA is compiled in the Genomic Data Commons Data Portal, a comprehensive data platform that consolidates cancer data from TCGA and other research programs.

RNA sequenced gene expression data for the tumors of 513 lung adenocarcinoma patients was compiled from TCGA on the Genomic Data Commons. One of the samples was removed as the patient's race was listed as the sole American Indian/Alaskan Native, a sample size too small to draw confident interpretations from, bringing the total samples for the data set to 512. For these 512 samples, the gene expression data for 19,340 different genes were recorded through RNA sequencing. To simplify the data, the raw counts of the reverse transcribed mRNA sequences at each gene loci were reduced to counts per million.

For each sample, diversity patient information was recorded. There exists a variation in what medical professionals do and do not record and the accuracy of that recorded data

(Brown et al., 2018). Due to this, the patient biographical information had to be focused upon five main biographical variables: sex, age, race, smoking history, and tumor stage (Table 1).

Biological sex serves as one of the largest differentiators between any two humans. Aside from the hormonal and physiological differences, there are clear genetic differences between the sexes (Short & Balban, 1994). Moreover, there has been a well-studied impact of sex on the development of certain cancers. Most importantly, this includes some forms of lung cancer, making sex an appropriate variable for this analysis (Dorak & Karpuzoglu, 2012).

Age is the single largest risk factor in cancer development (Armitage & Doll, 1954). Additionally, age is known to cause significant genetic changes over time that result in the process of human aging (Wheeler & Kim, 2011). It is theorized that many of the genetic changes in cancer and aging might be intertwined, so the need for further investigation dictates the inclusion of patient age data in this study (Aunan et al., 2017).

Race is also a well-studied genetic differentiator between individuals. While socioeconomic, environmental, and other factors have significant influences on racial impacts, there is clear evidence of varying disease susceptibility between racial groups (Anderson et al., 2004). This susceptibility disparity also applies to many forms of cancer, making race a valuable variable for this study (Özdemir & Dotto, 2017).

Smoking is the most infamous risk factor associated with lung cancer, being linked to 30 to 40% of all lung cancer deaths (L. A. Loeb et al., 1984). However, while some analysis has been conducted exploring the impacts on genetic expression of cancers due to smoking, it is certainly understudied for being the most significant cause of lung cancer (Woenckhaus et al., 2006). Conversely, deeper investigation into the genetics of lung cancer for patients who have no history of smoking is vital in illuminating the cancer-causing mechanisms for these cases without clear causative factors (Subramanian & Govindan, 2008).

Mutations that lead to the evolution of cancer do not stop once the cancer has developed, and many cancers increasingly develop subsequent mutations due to impaired DNA repair mechanisms (K. R. Loeb & Loeb, 2000). Moreover, research has shown significant differences in the profiles of various stages of esophageal cancer, indicating the potential for similar genetic differences to be found across other variations of cancer (Zhou et al., 2003). While tumor staging is based on the tumor's physical and metastatic characteristics, its vital to better understand the genetic alterations that occur as cancer progresses.

Sex data for the patients was the only variable recorded for all samples. The breakdown between males and females was fairly even, with 237 samples coming from males and 275 samples coming from females.

Age data was mostly complete throughout the entire data set, with only 19 samples recorded without an age. Ages were recorded at the time of the sample collection and ranged from 33 to 88 years of age, with a mean age of 65.30 and median age of 66. While some ages were recorded to the exact age in years and days, the data was simplified to simply the year of age. From here, ages were broken down into subgroups of 20 years to allow for group wide analysis of the data. These specific groupings were selected as they allowed for sufficient samples within each grouping and are aligned with groupings of early adulthood (31-50), late adulthood and early old age (51 to 70), and geriatric stage (71 to 90) patients (Vobr, 2013).

Comparatively, race data was quite poorly recorded, with 66 samples recorded without race. In addition to the aforementioned American Indian/Alaska Native patient whose sample was removed, there were 7 Asian, 52 African American, and 387 White samples. The patients classified as White included both non-Hispanic or Latino and Hispanic and Latino patients, as breakdown by Hispanic/Latino status would lead to another layer of incomplete recorded data and further reduce the value of the data.

Smoking history was recorded in two separate ways for the initial data set. Both ‘years smoked’ and ‘packs per year smoked’ were recorded at varying degrees, with samples being recorded with both, one, or neither of the two variables. ‘Packs per year’ data was selected as it was more completely recorded than ‘years smoked’, allowed for stratification based on the severity of the patient’s smoking, and had been used previously by researchers in the Kratz Lab when doing similar analyses (Mendez et al., 2017). While using how many cigarettes smoked is not a perfect measure of smoking consumption and risk, it is still one of the clearest indicators and has a strong correlation to smoking related conditions and addiction (Schane et al., 2010).

These patients were then parsed into non-smoking, light smoking, and heavy smoking groups. Non-smoking includes both patients recorded as smoking zero packs per year and patients whose packs per year were not recorded. For patient’s whose packs per year smoked were recorded, the breakdown between heavy and light smokers was based on Corrine Husten’s data analyzing how to best quantify heavy and light smokers (Husten, 2009). The most common figure cited in her paper seemed to be 5 cigarettes per day, which assuming 20 cigarettes per pack and 365 days in a year, equates to 91.25 packs per year. This resulted in a final breakdown of 162 non-smokers, 318 light smokers, and 32 heavy smokers.

Tumor-stage was the final factor, with 8 samples listed without a stage. In recent years, tumor staging has been standardized around the world to ensure better prognosis, diagnostic, and research can be done according to tumor stages. For lung cancers specifically, the American Joint Committee on Cancer completed their worldwide initiative for uniform staging standards (Chheang & Brown, 2013). Within the TCGA data, tumor stage was recorded with the main stages I, II, III, and IV, and with substages a and b for stage I, II, and III tumors. To simplify the data, the substages were removed bringing the total samples to 274 stage I, 120 stage II, 84 stage III, and 26 stage IV samples.

With the data now segregated into cleaner groups, the wholescale analysis of the dataset could begin. To conduct our test of significance, we utilized a permutational multivariate analysis of variance (perMANOVA). Biological data is rarely normally distributed, and this is especially so when dealing with gene expression data (Troyanskaya et al., 2002). As a non-parametric multivariate statistical test, the perMANOVA allows us to reduce the dimensionality of the intergroup and intragroup interactions of 5 distinct variables on 19,340 expressed genes into one singular test of significance. To do this, the perMANOVA generates its own model for determining significance using random permutations of the data to determine the dissimilarity matrix, rather than using some predetermined mathematical structure (joshuaebner, 2018).

The gene data was first imported into R such that each gene was transposed to be the leading variable for each column. The simplified patient data was then also imported and merged with the gene expression data to create the overall data frame. It is essential to ensure that all of the patient data was imported as categorical data while the gene expression data must be imported as numeric data.

Using the Vegan package for R, the perMANOVA function was run with 999 permutations and using Bray-Curtis as the dissimilarity matrix (Oksanen et al., 2020). Gene expression data served as the Y variable which was ran across each of the 5 patient information characteristics, which served as the X variables. From this, we were able to see the statistical significance of each patient characteristic on the gene expression data, as well as the significance of the interactions. In order to properly understand where the significant interactions amongst the variables occurred, the data was read from the most complex interactions, meaning the 5-way interaction between all the patient characteristics, down towards the effects of each characteristic on their own, with alpha set at 0.05. While all 5 characteristics showed significance in their effect on gene expression, this significance is likely a consequence of the upstream interactions of the characteristics.

Because of this, the interactions that were significant and subsequently chosen for visualization and analysis were the interaction between patient age group and sex, between tumor stage and smoking status, and between patient sex and tumor stage. Race was not found to have a statistically significant effect on gene expression when interacting with any of the other variables. However, it did have a statistically significant interaction on its own and was therefore included for further analysis.

To further assess the variation in gene expression for each variable, the significant results were plotted along non-metric multidimensional scaling (NMDS) planes using metaMDS in Vegan. Ordination analysis is a technique for plotting multivariate data onto a coordinate plane in order to graphically visualize calculated significant differences. While other forms of ordination analysis attempt to highlight variation within the data, NMDS offers the most accurate plotting for the dissimilarity between points since the plane is generated from the dissimilarity of the data itself. Being that NMDS structures ordination on a Cartesian plane, this also offers us the ability to increase or decrease the number of axes to plot data along, allowing for more flexibility in our ability to graph data (Kenkel & Orloci, 1986).

Two-dimensional scaling was first attempted, but after running metaMDS with 999 permutations per model run and 100 different runs per dimension, the model stress was still too large, and another dimension was necessitated. After running the model in 3 dimensions with 20 different runs per dimension, a sufficient model stress of 0.1623233 was produced. Model stress acts as a goodness of fit marker for the observations, and a stress level between 0.5 and 0.15 is considered a fair fit (Oksanen, n.d.). This model generated 3 NMDS values for each patient, collapsing the expression of the 19,340 genes into 3 linear axes.

To complete our analysis, the NMDS scores of the metaMDS model were combined with their respective patient information variables. The 3 NMDS scores served as the x, y, and z coordinates for the plotting of the data into 3-dimensional space. A 3-dimensional graph was first attempted for data visualization but resulted in a convoluted mess of data points. Therefore, 3 separate 2-dimensional graphs were utilized showing the 3 combinations of NMDS axes. These graphs were further simplified by facet wrapping each interaction by the primary variable, allowing for easier differentiation for each variable.

## Results

Permutational analysis of variance revealed three variable interactions that were statistically significant within our lung adenocarcinoma data set. The first of these was the interaction between the patients' smoking statuses and their tumor stages (perMANOVA, $F_{1,7}$

= 1.2177, p-value = 0.038, Figure 1). Plotting via non-metric multidimensional scaling allowed us to visualize the most noticeable differentiations in overall gene expression. For the interactions between stage II tumor patients who were heavy smokers, stage IV tumor patients who were heavy smokers, and unreported tumor stage patients who were light smokers, there were not enough data points to draw 95% confidence ellipses from. Non-smokers in stage IV differed most significantly from heavy and light smokers, while non and light smokers also differed noticeably from heavy smokers in stage I (Appendix 1).

The second significant interaction came with patients' sex and their tumor stages (perMANOVA, $F_{1,4}$ = 1.2476, p-value = 0.010, Figure 2). Plotting via non-metric multidimensional scaling allowed us to visualize the most noticeable differentiations in overall gene expression. For the interactions between female patients who had an unreported tumor stage, there were not enough data points to draw 95% confidence ellipses from. Gene expression data spread tended to vary more and more between males and females as tumor stage progressed, with stage III and IV patients in particular showing noticeable differences in expression profiles between men and women (Appendix 2).

The final significant interaction was between the patients' age group and sex (perMANOVA, $F_{1,3}$ = 1.2876, p-value = 0.036, Figure 3). Plotting via non-metric multidimensional scaling allowed us to visualize the most noticeable differentiations in overall gene expression. This interaction showed a similarly noticeable trend with variation in the data increasing between males and females as the age group of the patients decreased. The clearest differentiation came with our middle age group of 51-to-70-year old's, for which males' data was clearly spread more discernibly than for females (Appendix 3).

Lastly, the only variable that did not significantly interact with any others was race. However, race on its own was found to contribute significantly to difference in tumor gene expression profiles (perMANOVA $F_{1,3} = 2.1867$, p-value < 0.001, Figure 4). Plotting via non-metric multidimensional scaling allowed us to visualize the most noticeable differentiations in overall gene expression. This revealed apparent differentiation in data between the races. White patients and patients whose races were not reported were spread similarly with Black/African American patients and Asian patients each uniquely spreading differently (Appendix 4).

## Discussion

The overall goal of this analysis was to better understand the interactions of multiple external variables for lung adenocarcinoma patients and the effects these interactions have on tumor gene expression profiles. Based on our analysis, we can conclude that the interactions of demographic variables significantly impact these expression profiles. This analysis also brings to light many possibilities for the way lung adenocarcinomas arise and advance.

First and foremost, we must address the main issue with stratifying our data based on interactions between specific variables. Our data set is large enough to draw conclusions from and our analysis revealed statistically significant interactions, but more specific observations drawn from the NMDS visualization of the data spread must be met with the caveat of sample population size, a significant factor for certain groups. While most of the interaction subgroups have more than a sufficient number of individuals, certain groups were populated by only a few individuals, some of which have too few to even generate a 95% confidence interval (i.e., less than 4 per group). Most notably, this occurred with the stage and smoking status interaction, where the high number of categories within each variable led to a larger number of interaction possibilities and therefore some groups with too low a number of patients from which to draw conclusions. While these data points can be looked at with intrigue and used as supplemental support for conclusions drawn from sufficiently populous groups, it should be noted that attempting to make any solid observations would be scientifically improper and inaccurate.

A similar note should be made towards any data interaction of a ‘not reported’ group. As mentioned in the methods to this paper, data collection is often at the whim of who collects it. Even in a database as esteemed and rigorous as TCGA, there are limits on the completeness of any real-life data. Regardless, inclusion of these incomplete samples is still important. Removing any tumor sample due to the lack of a variable of study eliminates the information that the sample might have for our other 4 variables and reduces the validity of any conclusions that come from our data analysis. Furthermore, most of the ‘not reported’ groupings were quite small and can be ignored when looking at the sub-groupings of the significant interactions in a broad sense. The exception to this might be with the individual race isolation. While it would be improper to assume patient characteristics for those with ‘not reported’ for any specific variable, prevalence of White patients throughout this specific data set means it can be reasonably inferred that most of those whose race was ‘not reported’ were white. This is consistent with what we see in the race-based plotting of the data spread, with the not-reported group’s data spreading very similarly to the White race group. However, as with groups with low sample populations, any conclusions made from this should only serve as supplemental information to conclusions about the White group and should never be used for independent conclusions.

With these caveats addressed, we can begin to decipher the meaning of our results. The most glaring conclusion comes from looking at all the graphs together (Figures 1 – 4). Because all of the confidence intervals within every interaction grouping all encompass the same central area in our NMDS Euclidian diagram, we can conclude that none of these factors contribute to tumors with such different gene expressions that they might be considered separate entities. Essentially, this shows us that while tumor gene expression might vary greatly due to any specific variable, these tumors at their core are still fundamentally the same.

This conclusion is reasonable in the context of tumor and human biology. Firstly, these tumors are all lung adenocarcinomas, meaning they originate from similar tissue which would fundamentally have similar natural gene expression profiles (Sonawane et al., 2017). Furthermore, while every cancer is unique in some way and there are many different mechanisms that could go awry leading to the emergence of cancer, there are many functional similarities between cancers and consistencies in the pathways indicative to cancer formation. Studies have even shown similar genetic underpinnings across different types of cancer, so it’s no surprise that multiple cancers of the same type, like we have in our study, are in many ways intrinsically similar (Jiang et al., 2019). The more important conclusion to take away from looking at the graphs holistically is that the spread of the data differs noticeably for each variable breakdown. This shows us tumors are genetically impacted by our observed factors and genetically differentiate distinctly depending on external patient factors.

Let us first look at each of our significant interactions individually, starting with smoking and tumor stage. Unfortunately, due to the small sample sizes for many of the heavy smoking-interaction subgroups, most of our conclusions drawn from this interaction will have to focus on light and non-smokers (Figure 1). However, it is quite interesting to note that for stage I patients, where there was a sufficient population of heavy smokers in addition to light and non-smokers, heavy smokers showed far less differentiation than with the other smoking groups (Appendix 1). One reasoning for this might be that lung cancers detected early for heavy smokers are more likely to be linked directly to their smoking habits, whereas those who do not smoke more than likely have a variety of other underlying factors or random mutations that contributed to the cancer, meaning the genetic expressions for heavy smokers might vary less than those with more varying causes (Powell et al., 2003). However, more data would be needed for heavy smokers with tumors of higher stages to determine if this is indeed the case. Being that there is some level of expressional difference at every stage between light and non-smokers, it is more likely that smoking has a direct effect on the genes which have expressions that are altered during the rise and progression of lung cancer (He et al., 2018).

Patient sex and tumor stage interacting further enlightens on how tumor progression changes with other external factors. Male and female tumors are similarly differentiated at early stages of cancer development but begin to deviate from each other as cancer progresses (Appendix 2). Furthermore, this deviation becomes increasingly apparent at every subsequent increase in tumor stage, rather than simply deviating at the transition between two stages. This seems to indicate that there is some fundamental distinction between genetic males and females that leads to differing progression of cancer.

Significant dichotomy of sex-linked differentiation is also seen across different stages of the patients' life. Genetic variation was greater for males than for females, but there was a slight trend towards genetic similarity as patients aged. One explanation for this might be that for cancer to arise at far younger ages than typically expected, extreme and more genetically unique mutations might arise, leading to expression profiles that are more varied than the more typical elderly lung cancer patient (Berg et al., 2010).

Lastly, we must examine race. While race in itself did not contribute to any significant interactions, investigation into why this disassociation from other factors is the case could tell us just as much as investigation into interactions between two other variables. It is interesting that race did not correlate with any other factors despite a well-established consensus of racial differences in the prevalence of lung cancer (Schabath et al., 2016). While any discussions about the implications of race in regard to human health must be met with the admonition of the extreme cultural, environmental, and socioeconomic circumstances that connect to race, it can enlighten us to interesting hypotheses about race’s impact. Most importantly, in the framework of this analysis, the alienation of race from interacting with other factors might indicate that whatever genetic differences there are between races are completely distinct from the genes that lead to cancer variation.

While looking at each of these interactions on their own is illuminating, it’s important to observe them collectively. It is noteworthy that patient sex and their tumor stage were each a component in two out of those three interactions. This allowed us to observe common themes between each of those variables’ two sets of interaction plots.

For patients’ tumor stages, both its interaction with smoking status and with patient sex revealed that as the tumor of the patients progressed in the stage, the spread of the data increased. Being that cancer stems from mutations which build up over time, it makes sense that mutations would continue to accumulate as tumors advance in stage. Investigation into the driving genetic mutations that lead to metastasis (which is also the designating factor between stage III and stage IV tumors) highlights the accumulation of subsequent mutations being the culprit for this tumor advancement (Yokota, 2000). It can reasonably be assumed that the buildup of mutations occurs similarly between every stage of cancer, even if there is not a clear marker between stages as there is with malignancy. And as mutations in the genome directly tie to differential genetic expression, it is highly likely that this progression of genetic differentiation is tied to secondary mutations within the tumor (Jia & Zhao, 2017).

This enunciates the need to treat tumors at different stages as separate entities and begin to research diagnostic biomarkers and novel treatments on a stage specific basis. With smoking specifically, the genetic differentiation of smokers versus non-smokers is well studied, but most studies are done without regard for tumor staging and how the tumor may have changed as it progressed within the patient (X. Li et al., 2018). While it is obviously difficult to test tumors every step of the way, it is vital to better understand whether these expression differences as the cancer progresses are linked to previous smoking. The relationship discovered between smoking and tumor stage gene expression highlights the importance of investigating this relationship at an isolated level. If it’s widely studied and understood that smoking promotes normal cells mutating to become cancerous, why do we also study how smoking promotes already cancerous cells to mutate further?

Similarly, the two significant interactions with patient sex highlight some important trends for the future of oncological study. Consistently, throughout both its interaction with patient age group and tumor stage, males tended to show far greater data spread when compared with women of the opposing sub-grouping. This is potential due to our model more closely fitting gene expression profiles found in women due to there being more women in the bivariate ‘sex’ factor, meaning men were more likely to be spread farther and more noticeably (Kruskal, 1964). Regardless, the model still shows clear differences in the differentiation of males and females, indicating some underlying biological factor that ties to sex. Deferring genetic biomarkers between sexes have been studied throughout oncology and with non-small cell lung cancer specifically, so it not unreasonable to conclude that sex-linked genetic differentiation might be present at many levels of lung cancer development (Planchard et al., 2009).

Based on our results, the two factors with which sex was found to interact are key. Both age and tumor stage offer us progressive timelines over which to view genetic differentiation, indicating a linkage between time and sex in tumor development. Whether it’s over the course of the patient’s lifetime or the tumor’s, it is clear that cancers in men and women develop differently. Yet, in current oncological research, patient sex is often treated as a control variable rather than an independent variable with impacts to tumor biology (Rubin et al., 2020). The interaction with both of these progressive variables illuminates the potential need to isolate cancer research more harshly by sex.

On the whole, these results show that we may be too liberal in our classification of cancers. Differences in tumor stage are known to contribute to gene expression differences (Ma et al., 2003). So is patient sex (Brannon et al., 2012). Now we need to dive deeper into how segregation by these two easily accessible oncological variables can help us more accurately find genetic targets for diagnosis, prognosis, and novel therapy development.

Despite the success of this analysis, there are many ways the analysis could be improved to provide larger scale results. Most obviously, while 512 patient samples produce a sufficiently large data set, those 512 individuals in no way encompass all the possible variation within the immensely complex universe of lung oncology. In the Genomic Data Commons alone, there are thousands more tumor samples that, coupled with the numerous other cancer data bases throughout the world that are constantly being added to, provide us the opportunity to conduct a similar analysis on a much larger scale. In our analysis, there were many interactions which were nearly statistically significant, and larger-scale analyses could enlighten us onto other significant interactions that we were not able to find in this investigation.

In a similar vein, the five variables we studied were far from representative of all potential external factors that could impact lung cancer development. While these five factors were the most prevalently recorded throughout most data sets, there are many other potential risk factors and patient characteristics that would be worth investigating. External factors such as air pollution, radon, and hazardous chemicals have been known to increase susceptibility to lung cancer (Malhotra et al., 2016). Coupled with the abundance of other physical characteristics that could be recorded for patients, such as body mass index (BMI) or exercise habits, the possibility for other interactions is boundless (Arnold et al., 2016). This analysis serves as the initial step in finding all the ways external factors interact to differentiate lung cancer genomic expression.

Additionally, there are a few aspects of the analysis' methods that could be adjusted to provide better accuracy and insight for future studies. Firstly, due to the categorical nature of the variables required to perform a perMANOVA, there are some potential flaws with the way the data was reduced into specific groupings. The most significant of this comes with the grouping for the factor of smoking status. Patients who had no history of smoking were recorded as 'not reported' for their 'packs per year' values because of inconsistencies in the way data was recorded within the TCGA database. This opens up the possibility for smokers who did not have their ‘packs per year’ recorded being lumped into the same group as those who had never smoked. This was further complicated by the fact that there was an additional smoking status category of ‘years smoked’ which occasionally conflicted with the data in the ‘packs per year’ category. This complication of such a widely supplied data set could be mitigated with a large increase in the dataset size to allow for more flexibility in removing improperly recorded samples.

Similar issues could have arisen in the way the other factors were grouped as well. For example, it might be beneficial to separate Hispanic/Latino patients from the rest of the ‘White’ population within race, especially considering the cancer disparities between Hispanic/Latino and white populations (Zavala et al., 2021). Consolidating stages into just their main stage groups also removes the individual intricacies that make each sub-grouping of stages unique. Expanding the data set would allow both of these problems to be alleviated by removing the need to reduce the groupings down to more general categories. Lastly, grouping patients’ ages into rigid factions takes away the linearity of the age data and prevents the potential observations that could be made with regressions. Even if this isn’t possible, different groupings of ages may provide better results.

Data visualization could also be improved. Visualization of our perMANOVA in Euclidean space is extremely important as the perMANOVA itself only tells us the significance of each interaction. However, while the NMDS plots revealed clear 95% confidence intervals that allowed us to see differences in data spread, there was still convoluted data with each individual plot and increasing the size of the data set would only worsen this issue. To compensate, transforming the data, perhaps by factoring each samples’ coordinates by an exponent, could enunciate the differences in data spread and make visual differences clearer and more apparent.

Furthermore, our analysis only plotted the significant interactions. This not only means that we might be overlooking hidden trends within any of the non-significant interactions, but we also have no baseline for what non-significant interactions look like in Euclidean space. Secondary statistical testing on each sub-group interaction for all variable combinations could reveal these hidden trends within non-significant variable interactions.

Overall, despite these possibilities for improvement, the analysis of this TCGA data set proved ultimately successful and allowed us to see explicit interactions that alter genetic expression profiles of lung adenocarcinomas. Each of the factors we examined here have been extensively studied individually, but it is time we look deeper and try to connect the dots between them. The immense complexity and interconnectedness of cancer should be matched by equally complex and interconnected research methods, and this paper's findings provide a roadmap to guide other researchers on where to begin. As with nearly every experiment in science, more research is needed, but this analysis offers a foundation off of which our understanding of the way cancer forms, develops, and can be treated or prevented can leap off from.

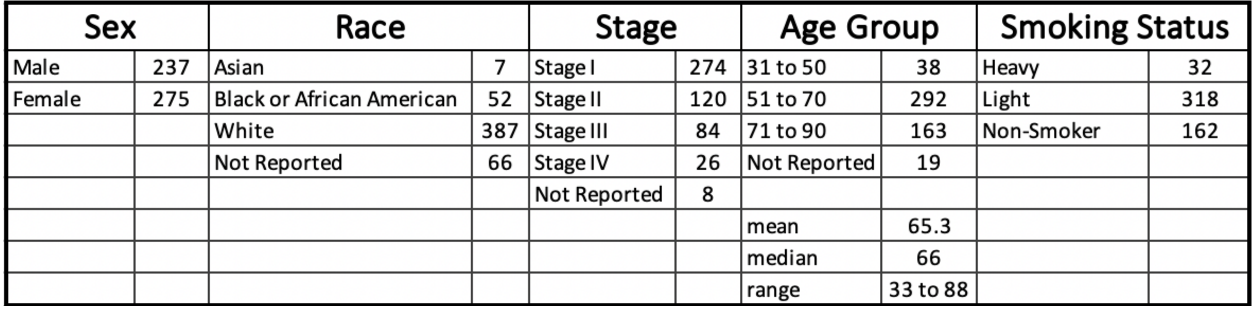
## Acknowledgements

I would like to extend my sincerest gratitude to my thesis advisors Professor Pete Chandrangsu and Professor Matthew Faldyn. As two of my favorite professors during my time at CMC, it was a pleasure having them guide me through this process and their accessibility throughout this process helped alleviate every problem or question I came across.

I would also like to thank Johannes Kratz, MD, Vivianne Ding, and Jack Li at UCSF's Kratz Lab for not only their help in acquiring and deciphering this data set, but also for the extraordinary opportunity to research in their thoracic oncology lab and further my passions for medicine.

# Figures

**Table 1.** Table showing the breakdown of the entire sample population compiled in the TCGA lung adenocarcinoma data set by the 5 demographic variables studied (n = 512). Within race, White refers to both patients who did and did not identify as Hispanic/Latino. Stage reduced substage groupings into only the main stage designation. Mean, median, and range of age group refers to each individual patients' age at the time the tumor sample was taken, after which the patients were segregated into 20-year groups. Heavy smokers were designated as those who smoke more than 5 cigarettes per day.



| Sex | | Race | | Stage | | Age Group | | Smoking Status | |
|---|---|---|---|---|---|---|---|---|---|
| Male | 237 | Asian | 7 | Stage I | 274 | 31 to 50 | 38 | Heavy | 32 |
| Female | 275 | Black or African American | 52 | Stage II | 120 | 51 to 70 | 292 | Light | 318 |
| | | White | 387 | Stage III | 84 | 71 to 90 | 163 | Non-Smoker | 162 |
| | | Not Reported | 66 | Stage IV | 26 | Not Reported | 19 | | |
| | | | | Not Reported | 8 | | | | |
| | | | | | | mean | 65.3 | | |
| | | | | | | median | 66 | | |
| | | | | | | range | 33 to 88 | | |

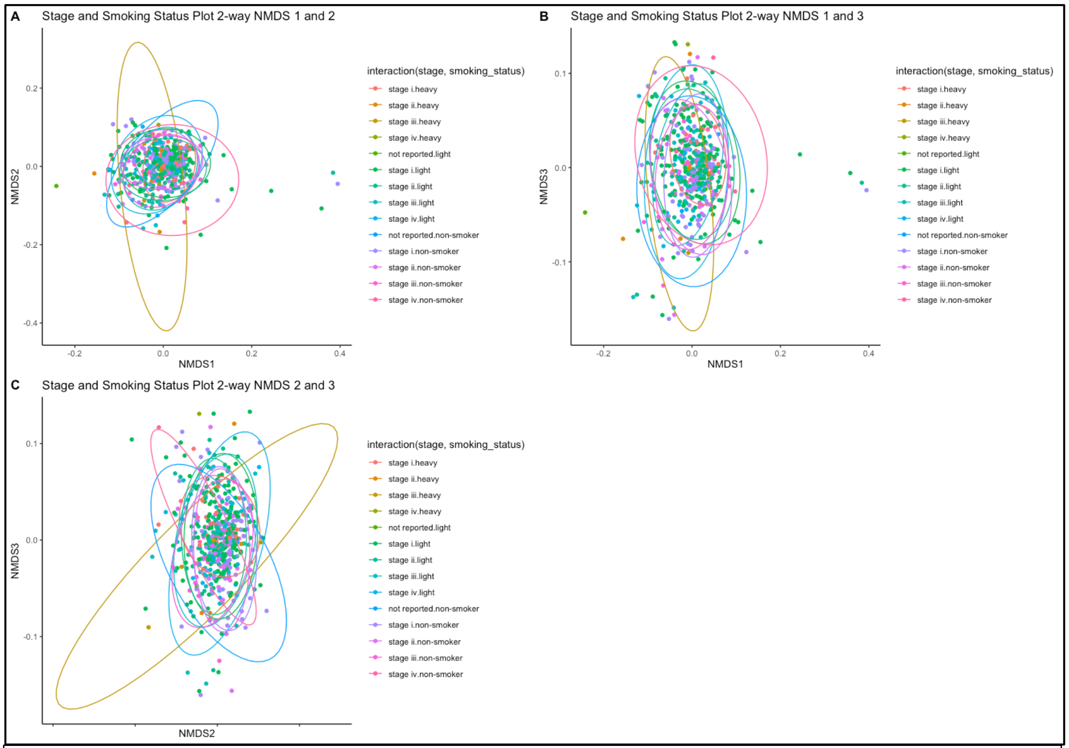


**Figure 1.** Gene expression of lung adenocarcinoma samples plotted in ordination space (n = 512), segregated by the interactions between patient tumor stage and smoking status (perMANOVA, $F_{1,7}$ = 1.2177, p-value = 0.038). Ordination axis are the NMDS values calculated for the data set using metaMDS in Vegan and the 3 NMDS axes generated a model stress of 0.1623233, a fair fit for a data set of this size (Oksanen et al.). Colored ellipses represent the 95% confidence interval for said color's specific stage-smoking interaction.

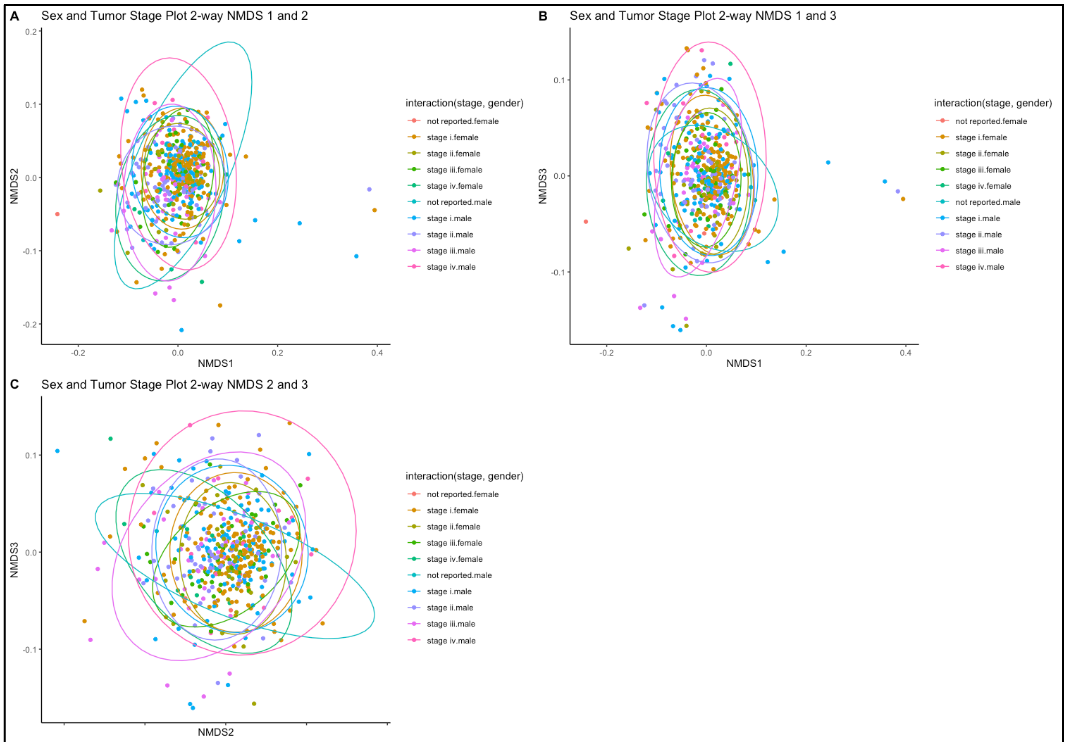


**Figure 2.** Gene expression of lung adenocarcinoma samples plotted in ordination space (n = 512), segregated by the interactions between patient tumor stage and sex (perMANOVA, $F_{1,4}$ = 1.2476, p-value = 0.010). Ordination axis are the NMDS values calculated for the data set using metaMDS in Vegan and the 3 NMDS axes generated a model stress of 0.1623233, a fair fit for a data set of this size (Oksanen et al.). Colored ellipses represent the 95% confidence interval for said color's specific stage-sex interaction.

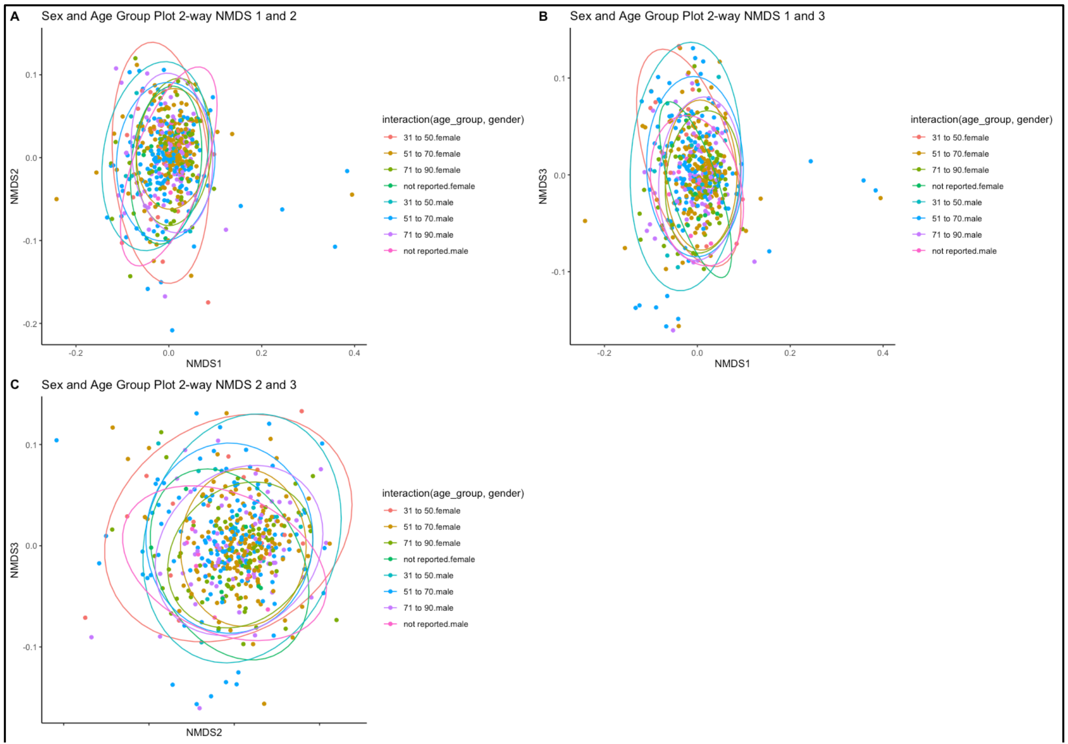


**Figure 3.** Gene expression of lung adenocarcinoma samples plotted in ordination space (n = 512), segregated by the interactions between patient age group and sex (perMANOVA, $F_{1,3}$ = 1.2876, p-value = 0.036). Ordination axis are the NMDS values calculated for the data set using metaMDS in Vegan and the 3 NMDS axes generated a model stress of 0.1623233, a fair fit for a data set of this size (Oksanen et al.). Colored ellipses represent the 95% confidence interval for said color's specific age-sex interaction.

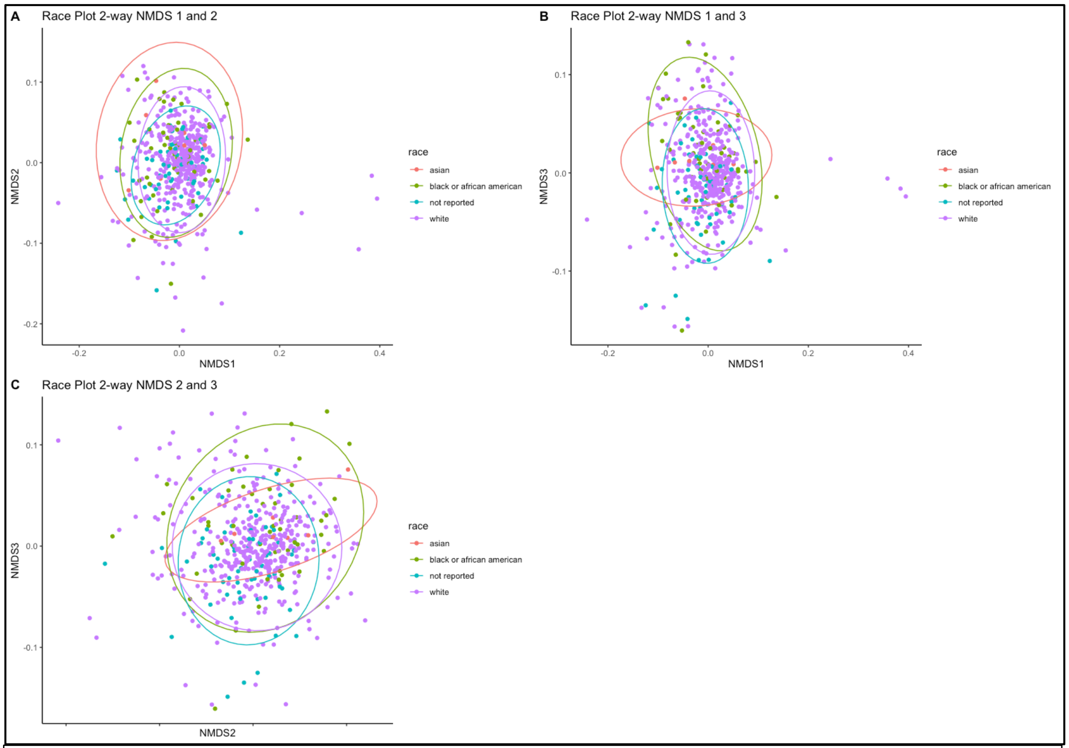


**Figure 4.** Gene expression of lung adenocarcinoma samples plotted in ordination space (n = 512), segregated by the race of each patient (perMANOVA, $F_{1,3} = 2.1867$, p-value < 0.001). Ordination axis are the NMDS values calculated for the data set using metaMDS in Vegan and the 3 NMDS axes generated a model stress of 0.1623233, a fair fit for a data set of this size (Oksanen et al.). Colored ellipses represent the 95% confidence interval for said color's specific race.

## Works Cited

Anderson, N. B., Bulatao, R. A., Cohen, B., & National Research Council (US) Panel on Race, E. (2004). Genetic Factors in Ethnic Disparities in Health. In *Critical Perspectives on Racial and Ethnic Differences in Health in Late Life*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK25517/

Armitage, P., & Doll, R. (1954). The Age Distribution of Cancer and a Multi-stage Theory of Carcinogenesis. *British Journal of Cancer*, *8*(1), 1–12.

Arnold, M., Leitzmann, M., Freisling, H., Bray, F., Romieu, I., Renehan, A., & Soerjomataram, I. (2016). Obesity and cancer: an update of the global impact. *Cancer Epidemiology, 41,* 8-15. https://doi.org/10.1016/j.canep.2016.01.003

Aslam, M. S., Naveed, S., Ahmed, A., Abbas, Z., Gull, I., & Athar, M. A. (2014). Side Effects of Chemotherapy in Cancer Patients and Evaluation of Patients Opinion about Starvation Based Differential Chemotherapy. *Journal of Cancer Therapy*, *2014*. https://doi.org/10.4236/jct.2014.58089

Aunan, J. R., Cho, W. C., & Søreide, K. (2017). The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. *Aging and Disease*, *8*(5), 628–642. https://doi.org/10.14336/AD.2017.0103

Auyang, S. Y. (2006). *Cancer causes and cancer research on many levels of complexity*. 15.

Barillot, E., Pook, S., Guyon, F., Cussat-Blanc, C., Viara, E., & Vaysseix, G. (1999). The HuGeMap Database: Interconnection and visualization of human genome maps. *Nucleic Acids Research*, *27*(1), 119–122.

Baskar, R., Dai, J., Wenlong, N., Yeo, R., & Yeoh, K.-W. (2014). Biological response of cancer cells to radiation treatment. *Frontiers in Molecular Biosciences*, *1*. https://doi.org/10.3389/fmolb.2014.00024

Berg, M., Danielsen, S. A., Ahlquist, T., Merok, M. A., Ågesen, T. H., Vatn, M. H., Mala, T., Sjo, O. H., Bakka, A., Moberg, I., Fetveit, T., Mathisen, Ø., Husby, A., Sandvik, O., Nesbakken, A., Thiis-Evensen, E., & Lothe, R. A. (2010). DNA Sequence Profiles of the Colorectal Cancer Critical Gene Set KRAS-BRAF-PIK3CA-PTEN-TP53 Related to Age at Disease Onset. *PLOS ONE*, *5*(11), e13978. https://doi.org/10.1371/journal.pone.0013978

Brannon, A. R., Haake, S. M., Hacker, K. E., Pruthi, R. S., Wallen, E. M., Nielsen, M. E., & Rathmell, W. K. (2012). Meta-analysis of Clear Cell Renal Cell Carcinoma Gene Expression Defines a Variant Subgroup and Identifies Gender Influences on Tumor Biology. *European Urology*, *61*(2), 258–268. https://doi.org/10.1016/j.eururo.2011.10.007

Brown, A. W., Kaiser, K. A., & Allison, D. B. (2018). Issues with data and analyses: Errors, underlying themes, and potential solutions. *Proceedings of the National Academy of Sciences*, *115*(11), 2563–2570. https://doi.org/10.1073/pnas.1708279115

Brownson, R. C., Reif, J. S., Keefe, T. J., Ferguson, S. W., & Pritzl, J. A. (1987). RISK FACTORS FOR ADENOCARCINOMA OF THE LUNG. *American Journal of Epidemiology*, *125*(1), 25–34. https://doi.org/10.1093/oxfordjournals.aje.a114509

Campa, D., Kaaks, R., Le Marchand, L., Haiman, C. A., Travis, R. C., Berg, C. D., Buring, J. E., Chanock, S. J., Diver, W. R., Dostal, L., Fournier, A., Hankinson, S. E., Henderson, B. E., Hoover, R. N., Isaacs, C., Johansson, M., Kolonel, L. N., Kraft, P., Lee, I.-M., …

Canzian, F. (2011). Interactions Between Genetic Variants and Breast Cancer Risk Factors in the Breast and Prostate Cancer Cohort Consortium. *JNCI: Journal of the National Cancer Institute*, *103*(16), 1252–1263. https://doi.org/10.1093/jnci/djr265

Chheang, S., & Brown, K. (2013). Lung Cancer Staging: Clinical and Radiologic Perspectives. *Seminars in Interventional Radiology*, *30*(2), 99–113. https://doi.org/10.1055/s-0033-1342950

Chin, L., Andersen, J. N., & Futreal, P. A. (2011). Cancer genomics: From discovery science to personalized medicine. *Nature Medicine*, *17*(3), 297–303. https://doi.org/10.1038/nm.2323

Colak, D., Nofal, A., AlBakheet, A., Nirmal, M., Jeprel, H., Eldali, A., AL-Tweigeri, T., Tulbah, A., Ajarim, D., Malik, O. A., Inan, M. S., Kaya, N., Park, B. H., & Amer, S. M. B. (2013). Age-Specific Gene Expression Signatures for Breast Tumors and Cross-Species Conserved Potential Cancer Progression Markers in Young Women. *PLOS ONE*, *8*(5), e63204. https://doi.org/10.1371/journal.pone.0063204

Dick, R. S., Steen, E. B., & Detmer, D. E. (1997). The Computer-Based Patient Record: Revised Edition: An Essential Technology for Health Care. In *The Computer-Based Patient Record: Revised Edition: An Essential Technology for Health Care*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK233055/

Dorak, M. T., & Karpuzoglu, E. (2012). Gender Differences in Cancer Susceptibility: An Inadequately Addressed Issue. *Frontiers in Genetics*, *3*. https://doi.org/10.3389/fgene.2012.00268

Friedman, A. A., Letai, A., Fisher, D. E., & Flaherty, K. T. (2015). Precision medicine for cancer with next-generation functional diagnostics. *Nature Reviews Cancer*, *15*(12), 747–756. https://doi.org/10.1038/nrc4015

Groenendijk, F. H., & Bernards, R. (2014). Drug resistance to targeted therapies: Déjà vu all over again. *Molecular Oncology*, *8*(6), 1067–1083. https://doi.org/10.1016/j.molonc.2014.05.004

Hartmaier, R. J., Charo, J., Fabrizio, D., Goldberg, M. E., Albacker, L. A., Pao, W., & Chmielecki, J. (2017). Genomic analysis of 63,220 tumors reveals insights into tumor uniqueness and targeted cancer immunotherapy strategies. *Genome Medicine*, *9*(1), 16. https://doi.org/10.1186/s13073-017-0408-2

He, X., Zhang, C., Shi, C., & Lu, Q. (2018). Meta-analysis of mRNA expression profiles to identify differentially expressed genes in lung adenocarcinoma tissue from smokers and non-smokers. *Oncology Reports*, *39*(3), 929–938. https://doi.org/10.3892/or.2018.6197

Heim, D., Budczies, J., Stenzinger, A., Treue, D., Hufnagl, P., Denkert, C., Dietel, M., & Klauschen, F. (2014). Cancer beyond organ and tissue specificity: Next-generation-sequencing gene mutation data reveal complex genetic similarities across major cancers. *International Journal of Cancer*, *135*(10), 2362–2369. https://doi.org/10.1002/ijc.28882

Hernandez, L. M., Blazer, D. G., & Institute of Medicine (US) Committee on Assessing Interactions Among Social, B. (2006). Sex/Gender, Race/Ethnicity, and Health. In *Genes, Behavior, and the Social Environment: Moving Beyond the Nature/Nurture Debate*. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK19934/

Howlader, N., Noone, A., Krapcho, M., Miller, D., Brest, A., Yu, M., Ruhl, J., Tatalovich, Z., Mariotto, A., Lewis, D., Chen, H., Feuer, E., & Cronin, K. (2020). *SEER Cancer*

*Statistics Review (CSR) 1975-2017* (SEER Cancer Statistics Review). National Cancer Institite.

Husten, C. G. (2009). How should we define light or intermittent smoking? Does it matter? *Nicotine & Tobacco Research*, *11*(2), 111–121. https://doi.org/10.1093/ntr/ntp010

Jia, P., & Zhao, Z. (2017). Impacts of somatic mutations on gene expression: An association perspective. *Briefings in Bioinformatics*, *18*(3), 413–425. https://doi.org/10.1093/bib/bbw037

Jiang, X., Finucane, H. K., Schumacher, F. R., Schmit, S. L., Tyrer, J. P., Han, Y., Michailidou, K., Lesseur, C., Kuchenbaecker, K. B., Dennis, J., Conti, D. V., Casey, G., Gaudet, M. M., Huyghe, J. R., Albanes, D., Aldrich, M. C., Andrew, A. S., Andrulis, I. L., Anton-Culver, H., … Lindström, S. (2019). Shared heritability and functional enrichment across six solid cancers. *Nature Communications*, *10*(1), 431. https://doi.org/10.1038/s41467-018-08054-4

joshuaebner. (2018, February 21). Permutational Multivariate Analysis of Variance (PERMANOVA) in R. *Archetypal Ecology*. https://archetypalecology.wordpress.com/2018/02/21/permutational-multivariate-analysis-of-variance-permanova-in-r-preliminary/

Kenkel, N. C., & Orloci, L. (1986). Applying Metric and Nonmetric Multidimensional Scaling to Ecological Studies: Some New Results. *Ecology*, *67*(4), 919–928. https://doi.org/10.2307/1939814

Kruskal, J. B. (1964). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, *29*(2), 115–129. https://doi.org/10.1007/BF02289694

Li, X., Li, J., Wu, P., Zhou, L., Lu, B., Ying, K., Chen, E., Lu, Y., & Liu, P. (2018). Smoker and non-smoker lung adenocarcinoma is characterized by distinct tumor immune microenvironments. *Oncoimmunology*, *7*(10). https://doi.org/10.1080/2162402X.2018.1494677

Li, Y., Xiao, X., Bossé, Y., Gorlova, O., Gorlov, I., Han, Y., Byun, J., Leighl, N., Johansen, J. S., Barnett, M., Chen, C., Goodman, G., Cox, A., Taylor, F., Woll, P., Wichmann, H. E., Manz, J., Muley, T., Risch, A., … Amos, C. I. (2019). Genetic interaction analysis among oncogenesis-related genes revealed novel genes and networks in lung cancer development. *Oncotarget*, *10*(19), 1760–1774. https://doi.org/10.18632/oncotarget.26678

Loeb, K. R., & Loeb, L. A. (2000). Significance of multiple mutations in cancer. *Carcinogenesis*, *21*(3), 379–385. https://doi.org/10.1093/carcin/21.3.379

Loeb, L. A., Emster, V. L., Warner, K. E., Abbotts, J., & Laszlo, J. (1984). Smoking and Lung Cancer: An Overview. *Cancer Research*, *44*(12 Part 1), 5940–5958.

Lopes-Ramos, C. M., Quackenbush, J., & DeMeo, D. L. (2020). Genome-Wide Sex and Gender Differences in Cancer. *Frontiers in Oncology*, *10*. https://doi.org/10.3389/fonc.2020.597788

Ma, X.-J., Salunga, R., Tuggle, J. T., Gaudet, J., Enright, E., McQuary, P., Payette, T., Pistone, M., Stecker, K., Zhang, B. M., Zhou, Y.-X., Varnholt, H., Smith, B., Gadd, M., Chatfield, E., Kessler, J., Baer, T. M., Erlander, M. G., & Sgroi, D. C. (2003). Gene expression profiles of human breast cancer progression. *Proceedings of the National Academy of Sciences*, *100*(10), 5974–5979. https://doi.org/10.1073/pnas.0931261100

Malhotra, J., Malvezzi, M., Negri, E., Vecchia, C. L., & Boffetta, P. (2016). Risk factors for lung cancer worldwide. *European Respiratory Journal*, *48*(3), 889–902. https://doi.org/10.1183/13993003.00359-2016

Markman, M. (2021, August 9). *Types of Lung Cancer: Common, Rare and More Varieties*. Cancer Treatment Centers of America. https://www.cancercenter.com/cancer-types/lung-cancer/types

Mendez, P., Fang, L. T., Jablons, D. M., & Kim, I.-J. (2017). Systematic comparison of two whole-genome amplification methods for targeted next-generation sequencing using frozen and FFPE normal and cancer tissues. *Scientific Reports*, *7*(1), 4055. https://doi.org/10.1038/s41598-017-04419-9

Moreira, A. L., & Eng, J. (2014). Personalized Therapy for Lung Cancer. *Chest*, *146*(6), 1649–1657. https://doi.org/10.1378/chest.14-0713

Myers, D. J., & Wallen, J. M. (2021). Lung Adenocarcinoma. In *StatPearls*. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK519578/

Nakamura, Y., Mochiada, A., Choyke, P., & Kobayashi, H. (2016, August 22). *Nanodrug Delivery: Is the Enhanced Permeability and Retention Effect Sufficient for Curing Cancer?* https://pubs.acs.org/doi/abs/10.1021/acs.bioconjchem.6b00437

Oksanen, J. (n.d.). *Vegan: An introduction to ordination*. 12.

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P. R., O’Hara, R. B., Simpson, G. L., Solymos, P., Stevens, M. H. H., Szoecs, E., & Wagner, H. (2020). *vegan: Community Ecology Package* (2.5-7) [Computer software]. https://CRAN.R-project.org/package=vegan

Özdemir, B. C., & Dotto, G.-P. (2017). Racial differences in cancer susceptibility and survival: More than the color of the skin? *Trends in Cancer*, *3*(3), 181–197. https://doi.org/10.1016/j.trecan.2017.02.002

Patel, T. A., Colon-Otero, G., Bueno Hume, C., Copland, J. A., & Perez, E. A. (2010). Breast Cancer in Latinas: Gene Expression, Differential Response to Treatments, and Differential Toxicities in Latinas Compared with Other Population Groups. *The Oncologist*, *15*(5), 466–475. https://doi.org/10.1634/theoncologist.2010-0004

Planchard, D., Loriot, Y., Goubar, A., Commo, F., & Soria, J.-C. (2009). Differential Expression of Biomarkers in Men and Women. *Seminars in Oncology*, *36*(6), 553–565. https://doi.org/10.1053/j.seminoncol.2009.09.004

Powell, C. A., Spira, A., Derti, A., DeLisi, C., Liu, G., Borczuk, A., Busch, S., Sahasrabudhe, S., Chen, Y., Sugarbaker, D., Bueno, R., Richards, W. G., & Brody, J. S. (2003). Gene Expression in Lung Adenocarcinomas of Smokers and Nonsmokers. *American Journal of Respiratory Cell and Molecular Biology*, *29*(2), 157–162. https://doi.org/10.1165/rcmb.2002-0183RC

Proctor, R. N. (2001). Tobacco and the global lung cancer epidemic. *Nature Reviews. Cancer*, *1*(1), 82–86. https://doi.org/10.1038/35094091

Riley, B. D., Culver, J. O., Skrzynia, C., Senter, L. A., Peters, J. A., Costalas, J. W., Callif-Daley, F., Grumet, S. C., Hunt, K. S., Nagy, R. S., McKinnon, W. C., Petrucelli, N. M., Bennett, R. L., & Trepanier, A. M. (2012). Essential Elements of Genetic Cancer Risk Assessment, Counseling, and Testing: Updated Recommendations of the National Society of Genetic Counselors. *Journal of Genetic Counseling*, *21*(2), 151–161. https://doi.org/10.1007/s10897-011-9462-x

Rubin, J. B., Lagas, J. S., Broestl, L., Sponagel, J., Rockwell, N., Rhee, G., Rosen, S. F., Chen, S., Klein, R. S., Imoukhuede, P., & Luo, J. (2020). Sex differences in cancer mechanisms. *Biology of Sex Differences*, *11*. https://doi.org/10.1186/s13293-020-00291-x

Schabath, M. B., Cress, W. D., & Muñoz-Antonia, T. (2016). Racial and Ethnic Differences in the Epidemiology of Lung Cancer and the Lung Cancer Genome. *Cancer Control : Journal of the Moffitt Cancer Center*, *23*(4), 338–346.

Schane, R. E., Ling, P. M., & Glantz, S. A. (2010). Health Effects of Light and Intermittent Smoking: A Review. *Circulation*, *121*(13), 1518–1522. https://doi.org/10.1161/CIRCULATIONAHA.109.904235

Short, R. V., & Balban, E. (1994). *The Differences Between the Sexes*. Cambridge University Press.

Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., DeMeo, D. L., Quackenbush, J., Glass, K., & Kuijjer, M. L. (2017). Understanding Tissue-Specific Gene Regulation. *Cell Reports*, *21*(4), 1077–1088. https://doi.org/10.1016/j.celrep.2017.10.001

Subramanian, J., & Govindan, R. (2008). Molecular genetics of lung cancer in people who have never smoked. *The Lancet Oncology*, *9*(7), 676–682. https://doi.org/10.1016/S1470-2045(08)70174-8

Suda, K., & Mitsudomi, T. (2014). Successes and limitations of targeted cancer therapy in lung cancer. *Progress in Tumor Research*, *41*, 62–77. https://doi.org/10.1159/000355902

Travis, W., Brambilla, E., & Riely, G. (2013). New Pathologic Classification of Lung Cancer: Relevance for Clinical Practice and Clinical Trials. *Journal of Clinical Oncology*. https://doi.org/10.1200/JCO.2012.46.9270

Troyanskaya, O. G., Garber, M. E., Brown, P. O., Botstein, D., & Altman, R. B. (2002). Nonparametric methods for identifying differentially expressed genes in microarray data. *Bioinformatics*, *18*(11), 1454–1461. https://doi.org/10.1093/bioinformatics/18.11.1454

Valle, I., Tramalloni, D., & Bragazzi, N. L. (2015). Cancer prevention: State of the art and future prospects. *Journal of Preventive Medicine and Hygiene*, *56*(1), E21–E27.

Vobr, R. (2013). Periods of Human Age. In *Anthropomotorics* (1st ed.). Masaryk University, Brno 2013.

Walser, T., Cui, X., Yanagawa, J., Lee, J. M., Heinrich, E., Lee, G., Sharma, S., & Dubinett, S. M. (2008). Smoking and Lung Cancer. *Proceedings of the American Thoracic Society*, *5*(8), 811–815. https://doi.org/10.1513/pats.200809-100TH

Weinberg, R. A. (1996). How Cancer Arises. *Scientific American*, *275*(3), 62–70.

Weinstein, J. N., Collisson, E. A., Mills, G. B., Shaw, K. R. M., Ozenberger, B. A., Ellrott, K., Shmulevich, I., Sander, C., & Stuart, J. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics*, *45*(10), 1113–1120. https://doi.org/10.1038/ng.2764

Wheeler, H. E., & Kim, S. K. (2011). Genetics and genomics of human ageing. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1561), 43–50. https://doi.org/10.1098/rstb.2010.0259

Woenckhaus, M., Klein-Hitpass, L., Grepmeier, U., Merk, J., Pfeifer, M., Wild, P. J., Bettstetter, M., Wuensch, P., Blaszyk, H., Hartmann, A., Hofstaedter, F., & Dietmaier, W. (2006). Smoking and cancer-related gene expression in bronchial epithelium and non-small-cell lung cancers. *The Journal of Pathology*, *210*(2), 192–204. https://doi.org/10.1002/path.2039

World Health Organization. (2021, March 1). *WHO Mortality Database—WHO*. https://www.who.int/data/data-collection-tools/who-mortality-database

Wu, M., & Ma, S. (2019). Robust genetic interaction analysis. *Briefings in Bioinformatics*, *20*(2), 624–637. https://doi.org/10.1093/bib/bby033

Yokota, J. (2000). Tumor progression and metastasis. *Carcinogenesis*, *21*(3), 497–503. https://doi.org/10.1093/carcin/21.3.497

Yoshino, I., & Maehara, Y. (2007). Impact of Smoking Status on the Biological Behavior of Lung Cancer. *Surgery Today*, *37*(9), 725–734. https://doi.org/10.1007/s00595-007-3516-6

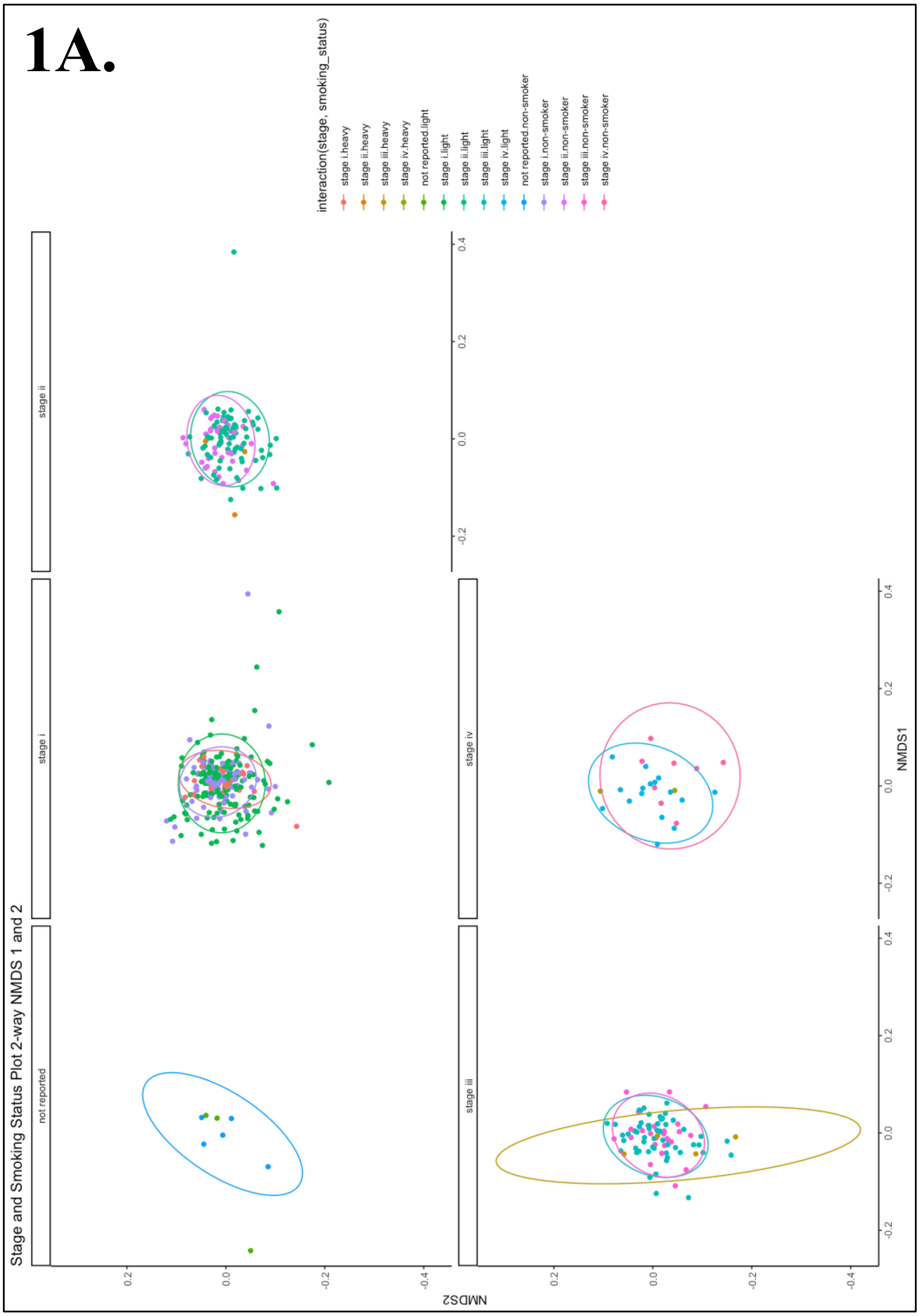
Zahm, S. H., & Fraumeni, J. F. (1995). Racial, ethnic, and gender variations in cancer risk: Considerations for future epidemiologic research. *Environmental Health Perspectives*, *103 Suppl 8*, 283–286. https://doi.org/10.1289/ehp.95103s8283

Zavala, V. A., Bracci, P. M., Carethers, J. M., Carvajal-Carmona, L., Coggins, N. B., Cruz-Correa, M. R., Davis, M., de Smith, A. J., Dutil, J., Figueiredo, J. C., Fox, R., Graves, K. D., Gomez, S. L., Llera, A., Neuhausen, S. L., Newman, L., Nguyen, T., Palmer, J. R., Palmer, N. R., … Fejerman, L. (2021). Cancer health disparities in racial/ethnic minorities in the United States. *British Journal of Cancer*, *124*(2), 315–332. https://doi.org/10.1038/s41416-020-01038-6

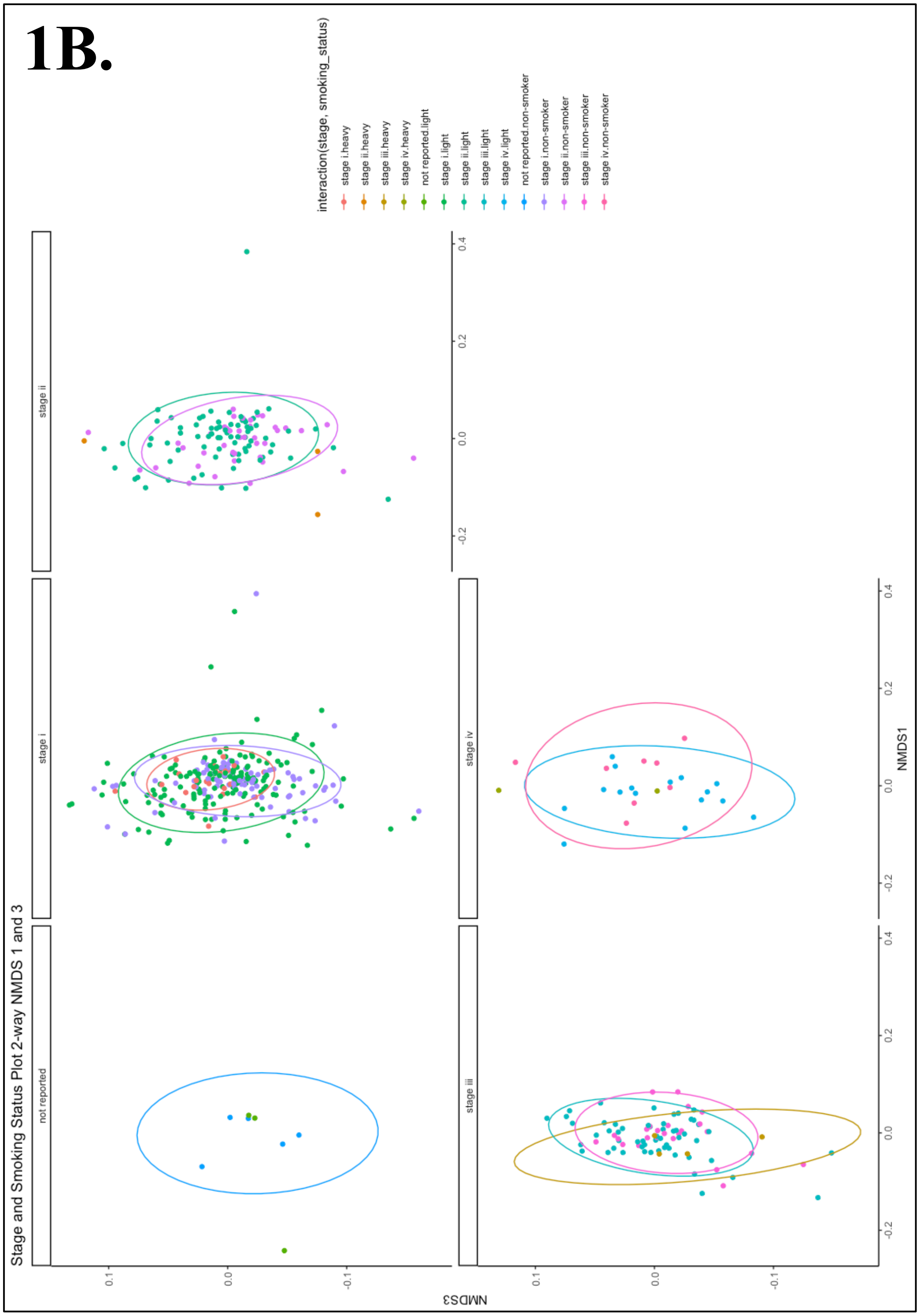
Zhou, J., Zhao, L.-Q., Xiong, M.-M., Wang, X.-Q., Yang, G.-R., Qiu, Z.-L., Wu, M., & Liu, Z.-H. (2003). Gene expression profiles at different stages of human esophageal squamous cell carcinoma. *World Journal of Gastroenterology*, *9*(1), 9–15. https://doi.org/10.3748/wjg.v9.i1.9

# Appendix

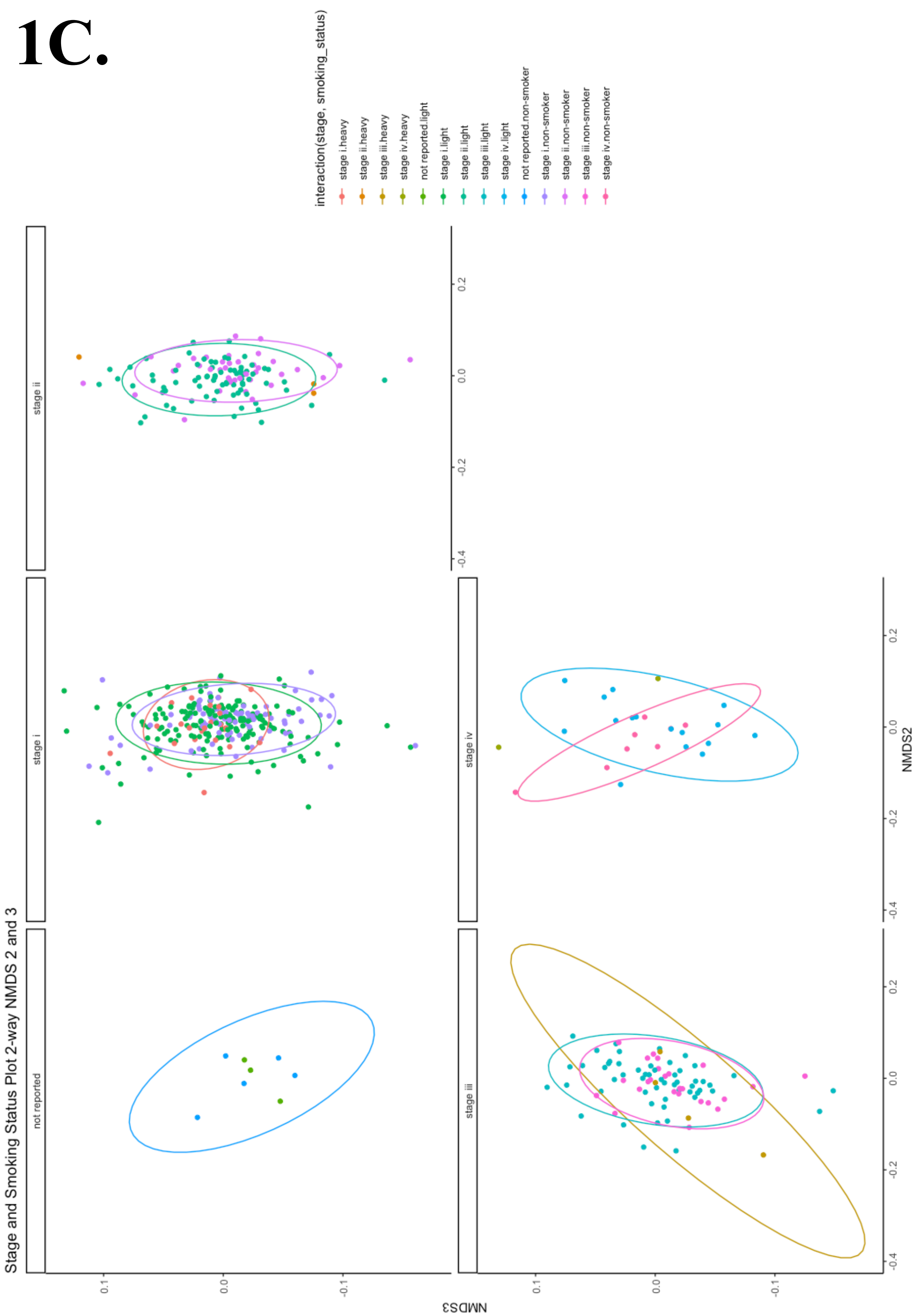
1A.


Stage and Smoking Status Plot 2-way NMDS 1 and 2
not reported
stage i
stage ii
stage iii
stage iv
NMDS1
NMDS2
-0.2
0.0
0.2
0.4
-0.4
interaction(stage, smoking_status)
stage i.heavy
stage ii.heavy
stage iii.heavy
stage iv.heavy
not reported.light
stage i.light
stage ii.light
stage iii.light
stage iv.light
not reported.non-smoker
stage i.non-smoker
stage ii.non-smoker
stage iii.non-smoker
stage iv.non-smoker

1B.

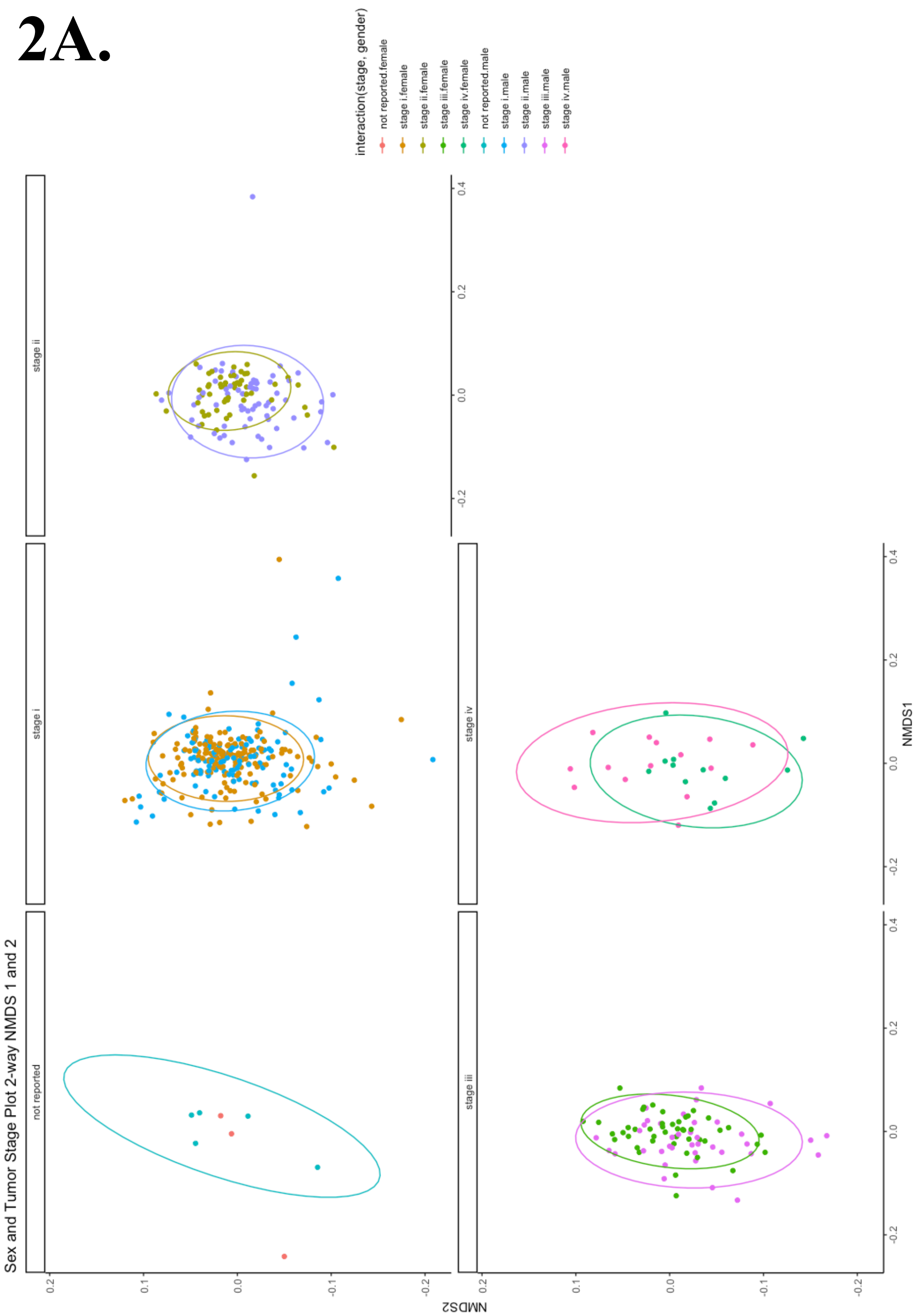

Stage and Smoking Status Plot 2-way NMDS 1 and 3
not reported
stage i
stage ii
stage iii
stage iv
NMDS1
NMDS3
-0.2
0.0
0.2
0.4
0.1
0.0
-0.1
interaction(stage, smoking_status)
stage i.heavy
stage ii.heavy
stage iii.heavy
stage iv.heavy
not reported.light
stage i.light
stage ii.light
stage iii.light
stage iv.light
not reported.non-smoker
stage i.non-smoker
stage ii.non-smoker
stage iii.non-smoker
stage iv.non-smoker

1C.

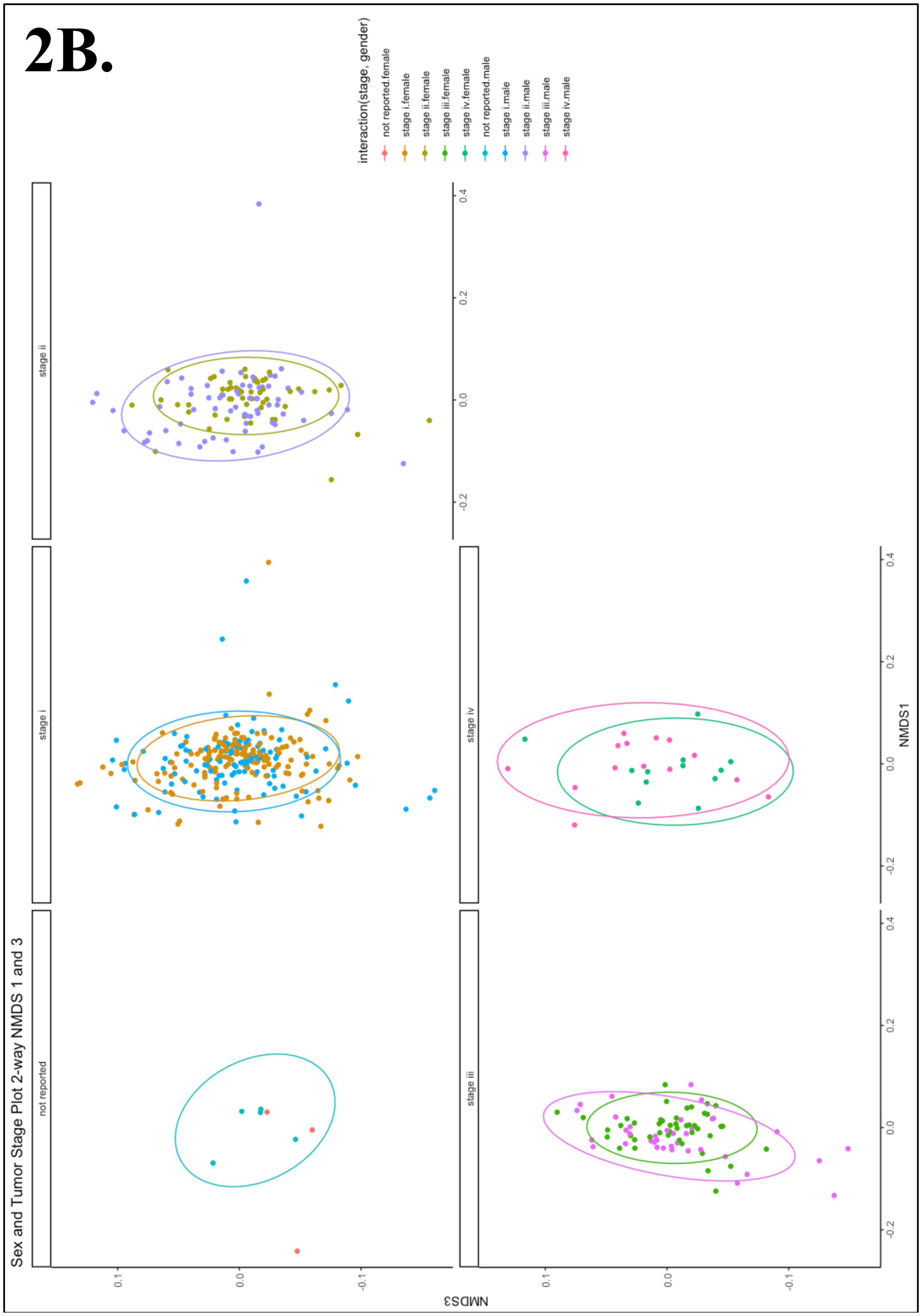


**Appendix 1 A-C.** Patient tumor stage and smoking status interaction gene expression ordination plots found in Figure 1 facet wrapped to show the differences within each individual stage studied.

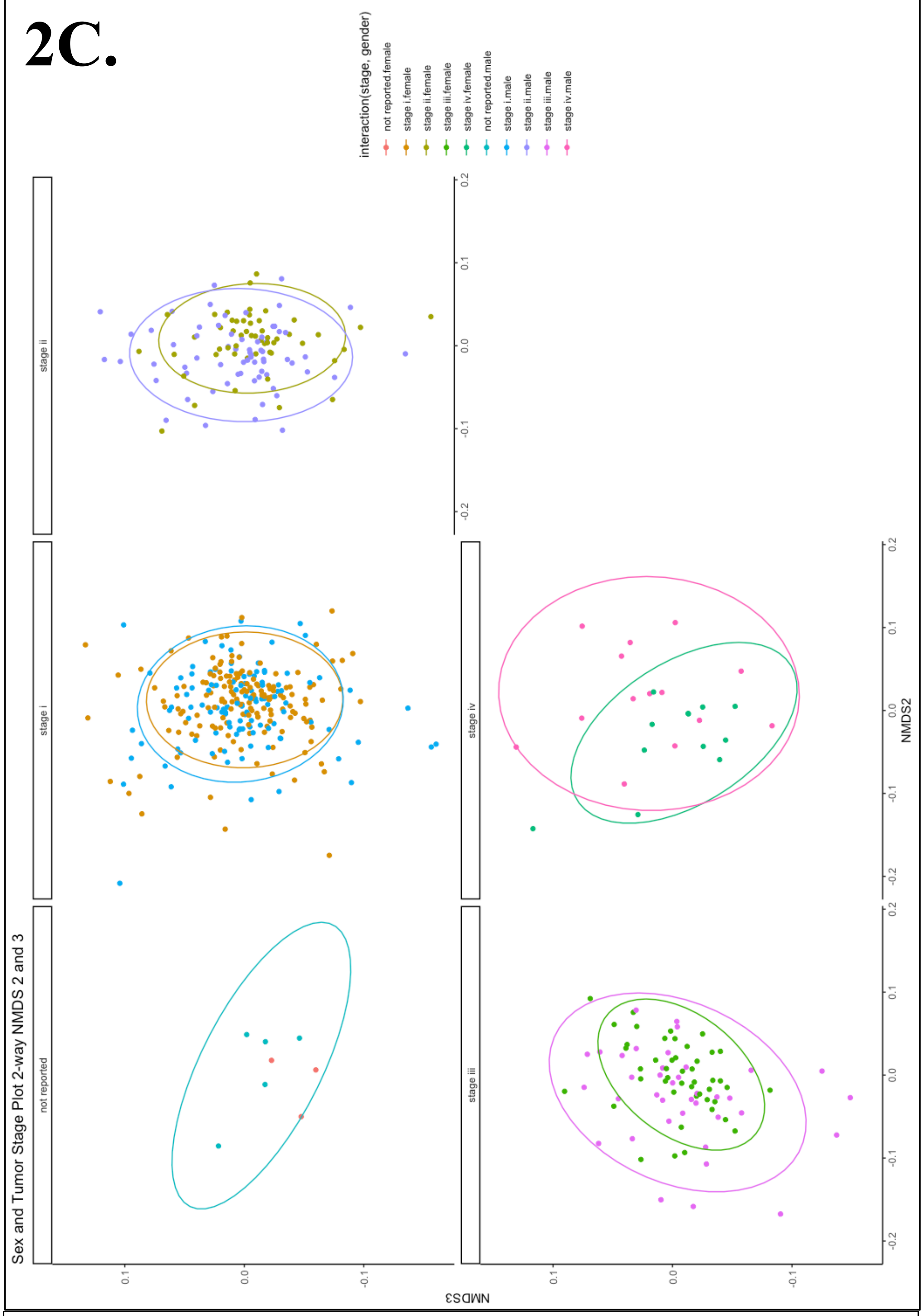
**2A.**


Sex and Tumor Stage Plot 2-way NMDS 1 and 2
not reported
stage i
stage ii
stage iii
stage iv
NMDS2
NMDS1
0.2
0.1
0.0
-0.1
-0.2
-0.2
0.0
0.2
0.4
interaction(stage, gender)
not reported.female
stage i.female
stage ii.female
stage iii.female
stage iv.female
not reported.male
stage i.male
stage ii.male
stage iii.male
stage iv.male

2B.

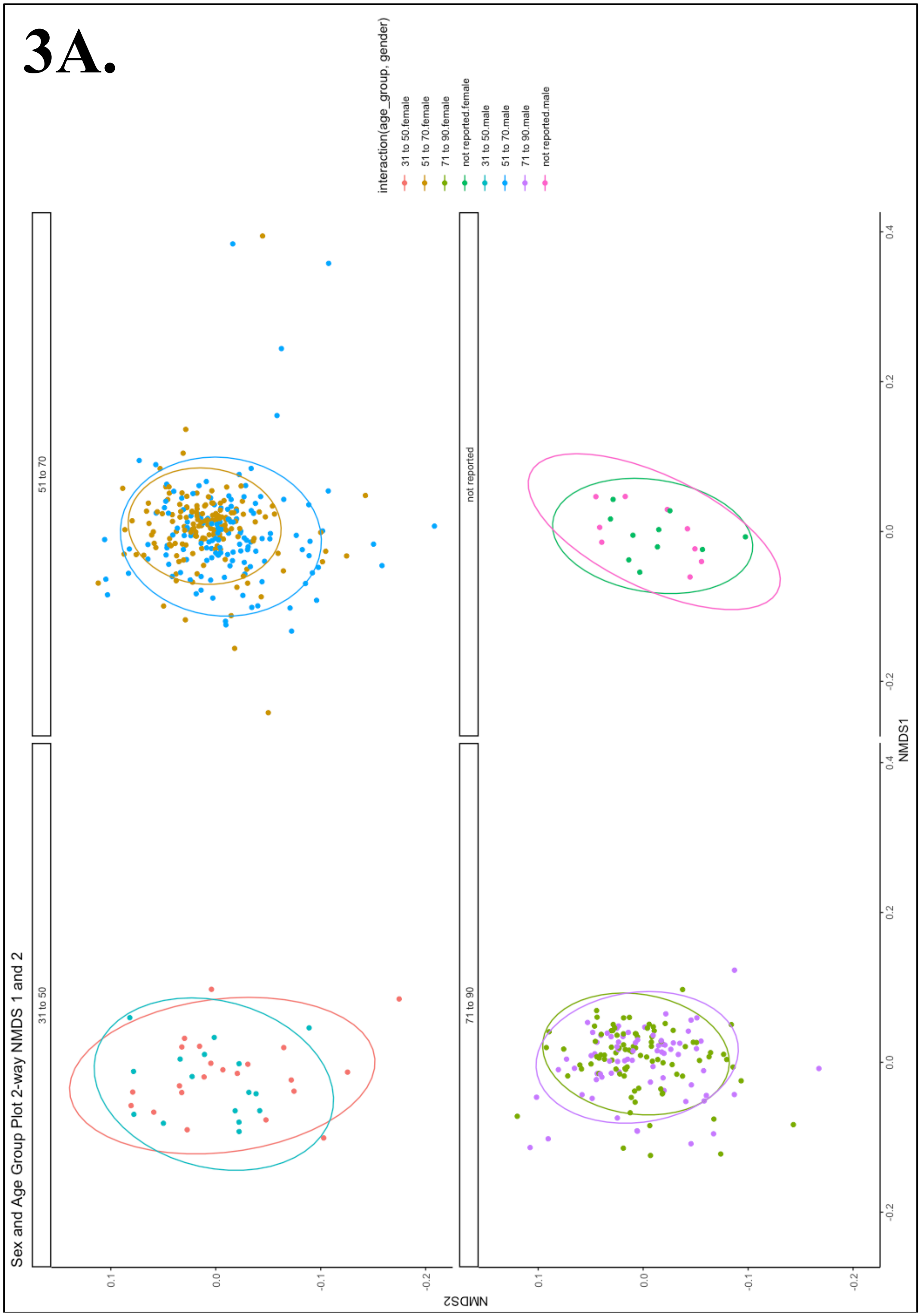

Sex and Tumor Stage Plot 2-way NMDS 1 and 3
not reported
stage i
stage ii
stage iii
stage iv
NMDS1
NMDS3
-0.2
0.0
0.2
0.4
0.1
0.0
-0.1
interaction(stage, gender)
not reported.female
stage i.female
stage ii.female
stage iii.female
stage iv.female
not reported.male
stage i.male
stage ii.male
stage iii.male
stage iv.male

2C.

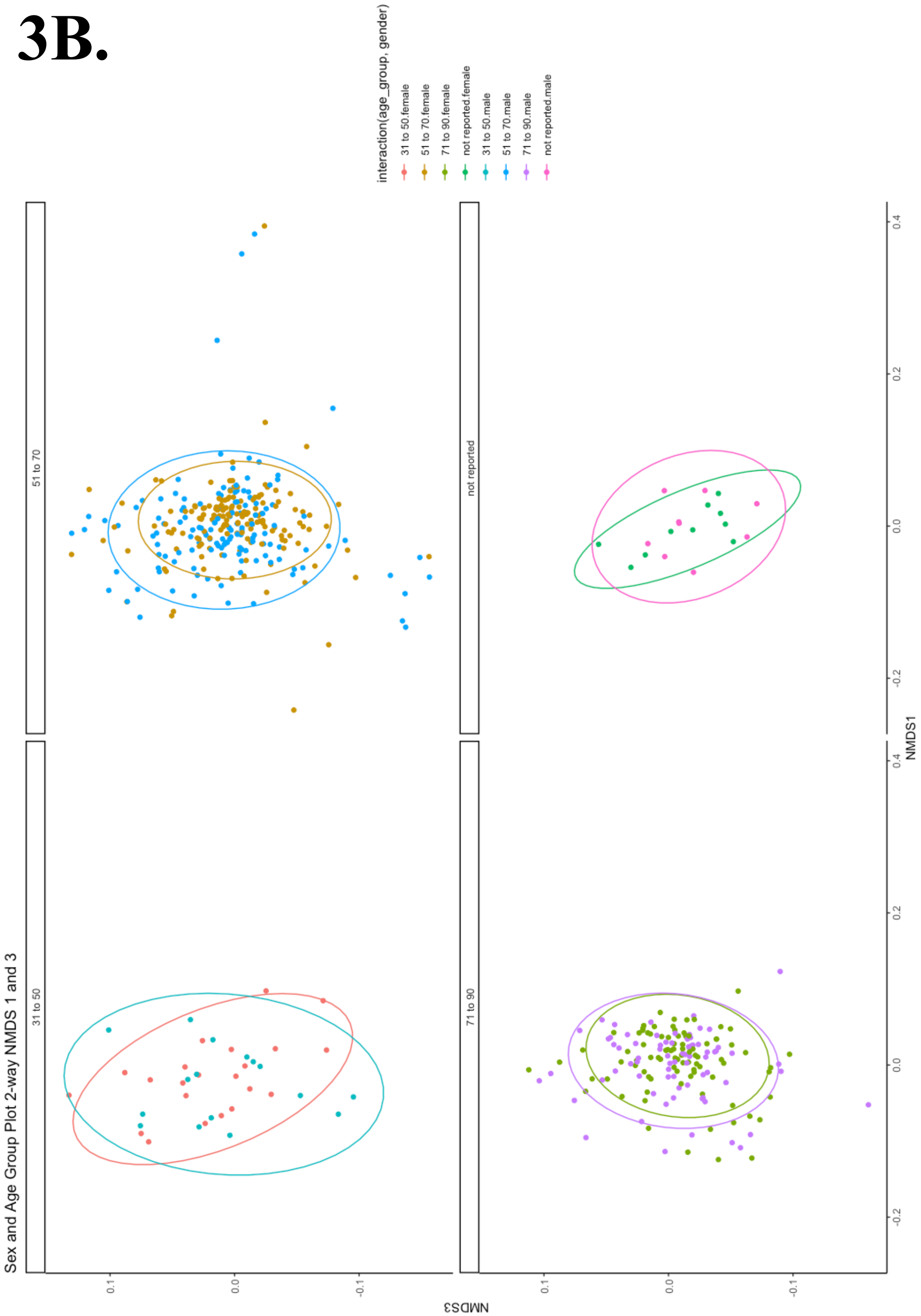


**Appendix 2 A-C.** Patient tumor stage and sex interaction gene expression ordination plots found in Figure 2 facet wrapped to show the differences within each individual stage studied.

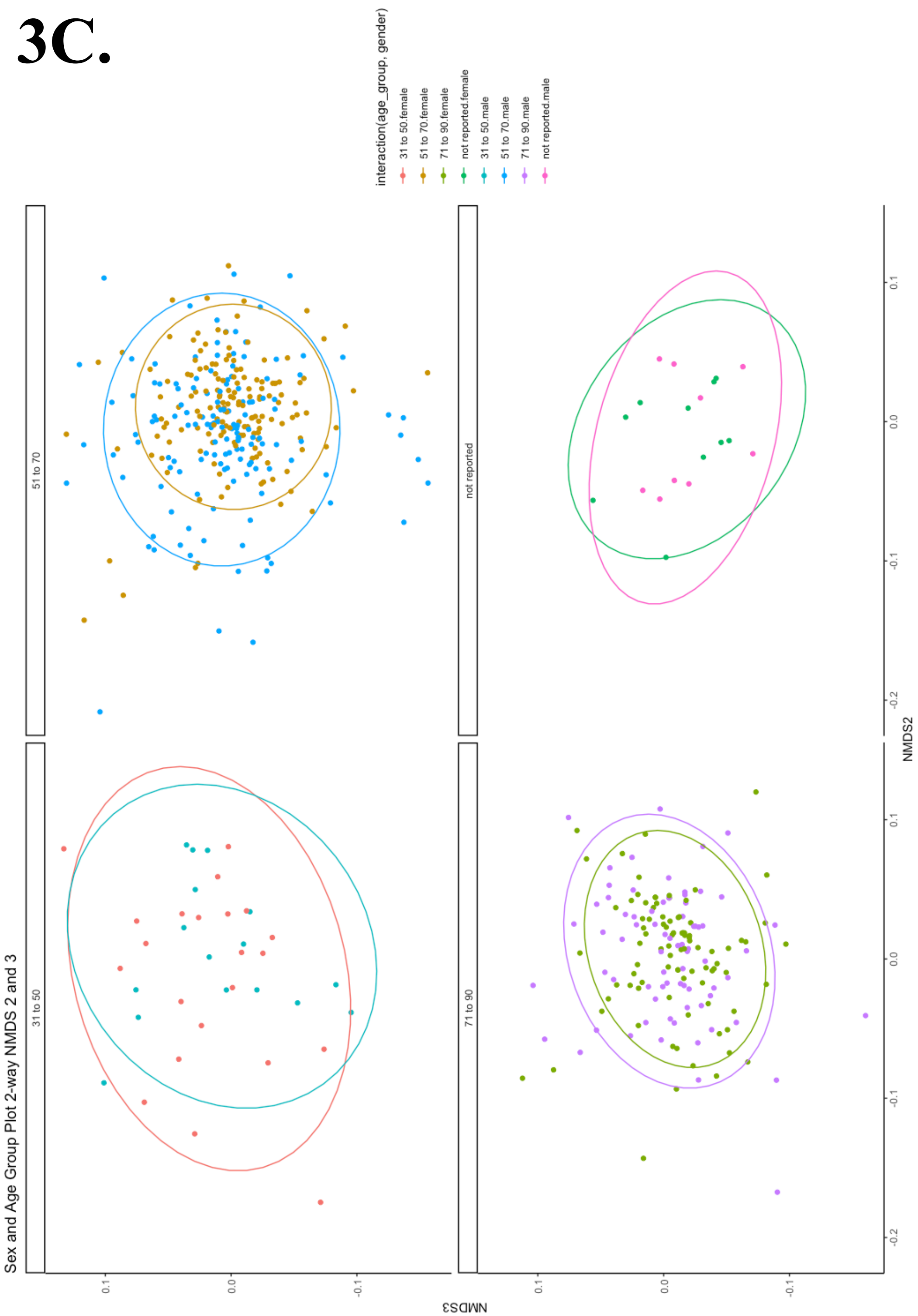
3A.


Sex and Age Group Plot 2-way NMDS 1 and 2
31 to 50
51 to 70
71 to 90
not reported
interaction(age_group, gender)
31 to 50.female
51 to 70.female
71 to 90.female
not reported.female
31 to 50.male
51 to 70.male
71 to 90.male
not reported.male
NMDS1
NMDS2
-0.2
0.0
0.2
0.4
0.1
0.0
-0.1
-0.2

3B.

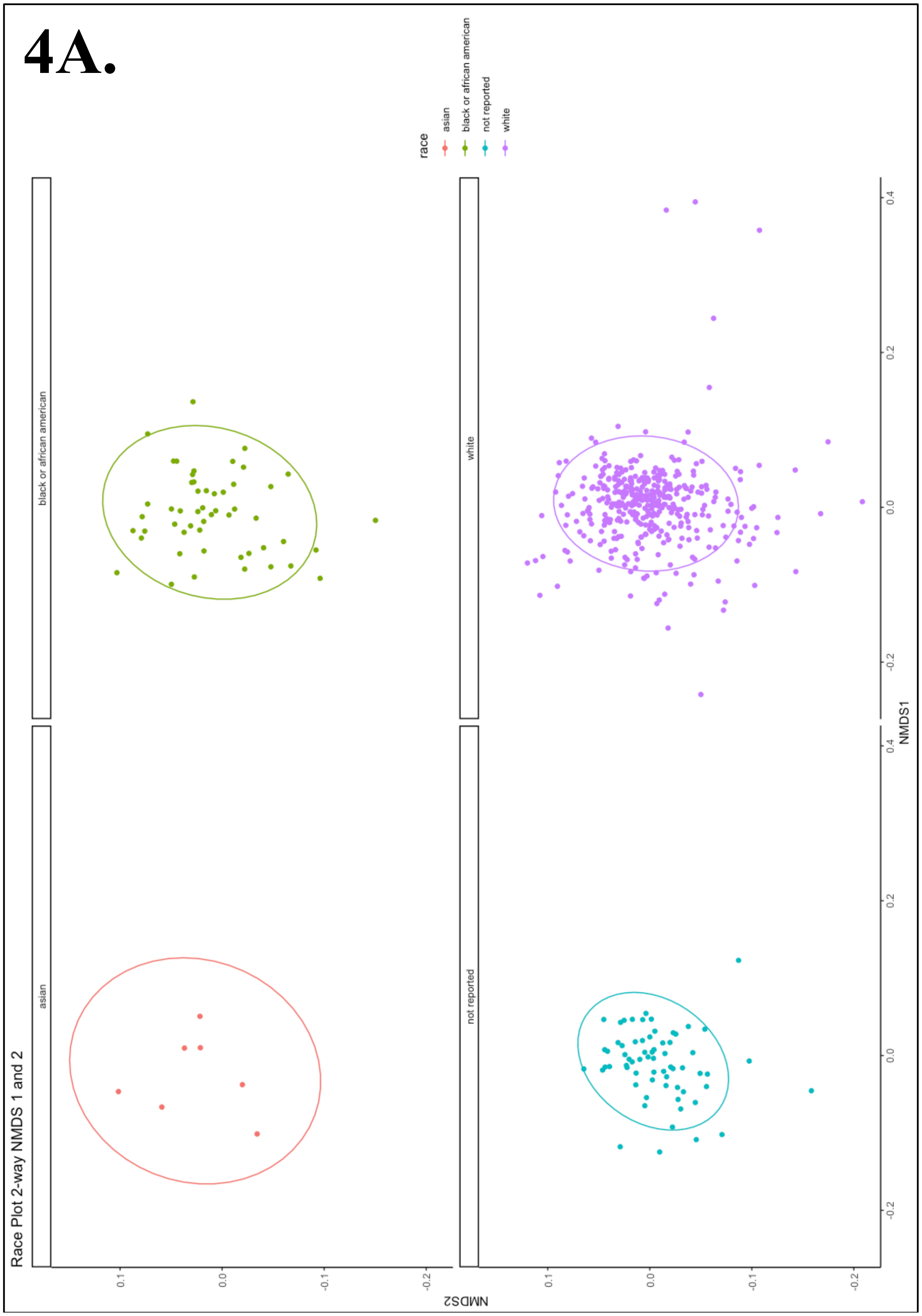

Sex and Age Group Plot 2-way NMDS 1 and 3
31 to 50
51 to 70
71 to 90
not reported
NMDS3
NMDS1
0.1
0.0
-0.1
-0.2
0.2
0.4
interaction(age_group, gender)
31 to 50.female
51 to 70.female
71 to 90.female
not reported.female
31 to 50.male
51 to 70.male
71 to 90.male
not reported.male

3C.

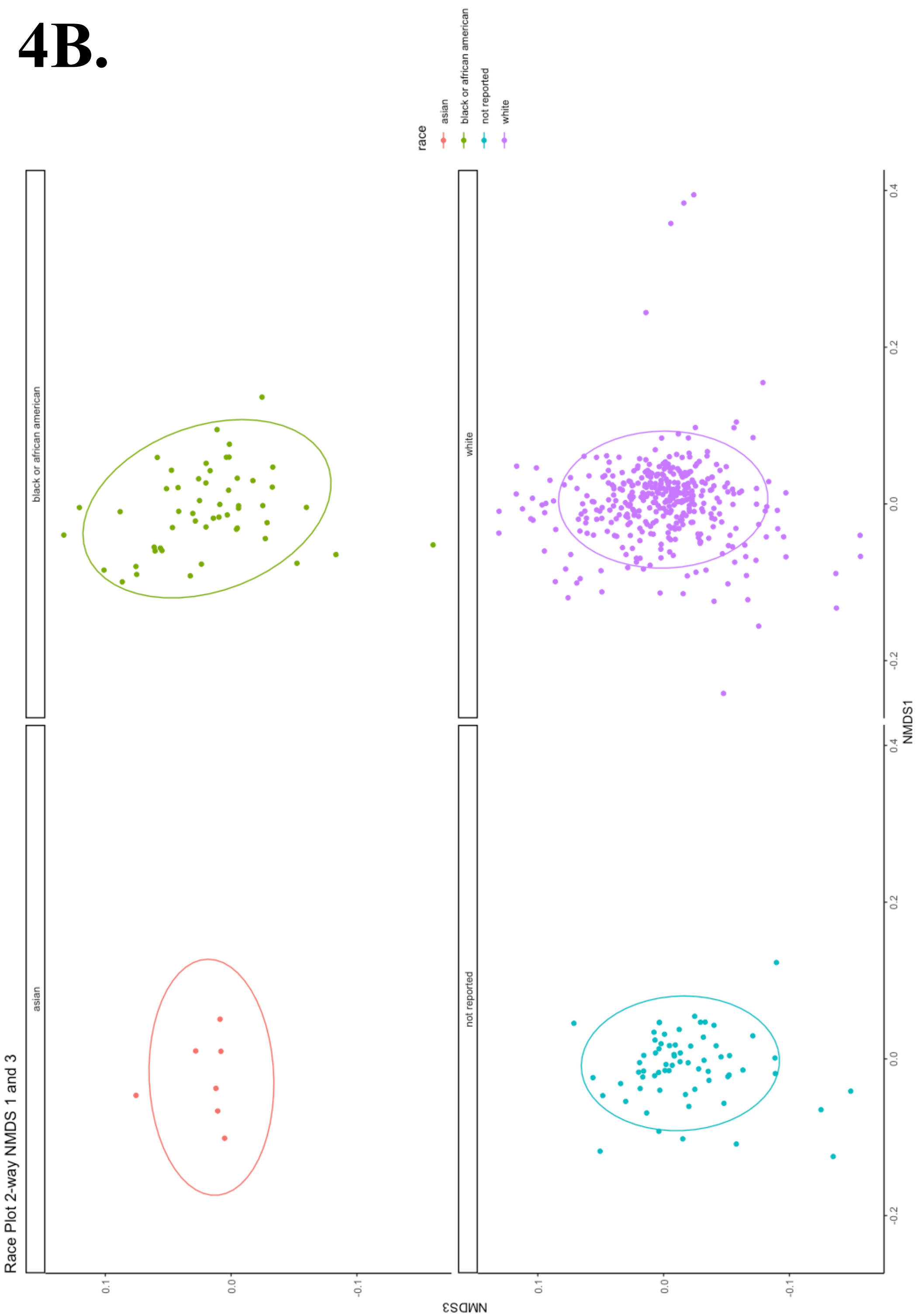


**Appendix 3 A-C.** Patient age group and sex interaction gene expression ordination plots found in Figure 3 facet wrapped to show the differences within each individual age group studied.

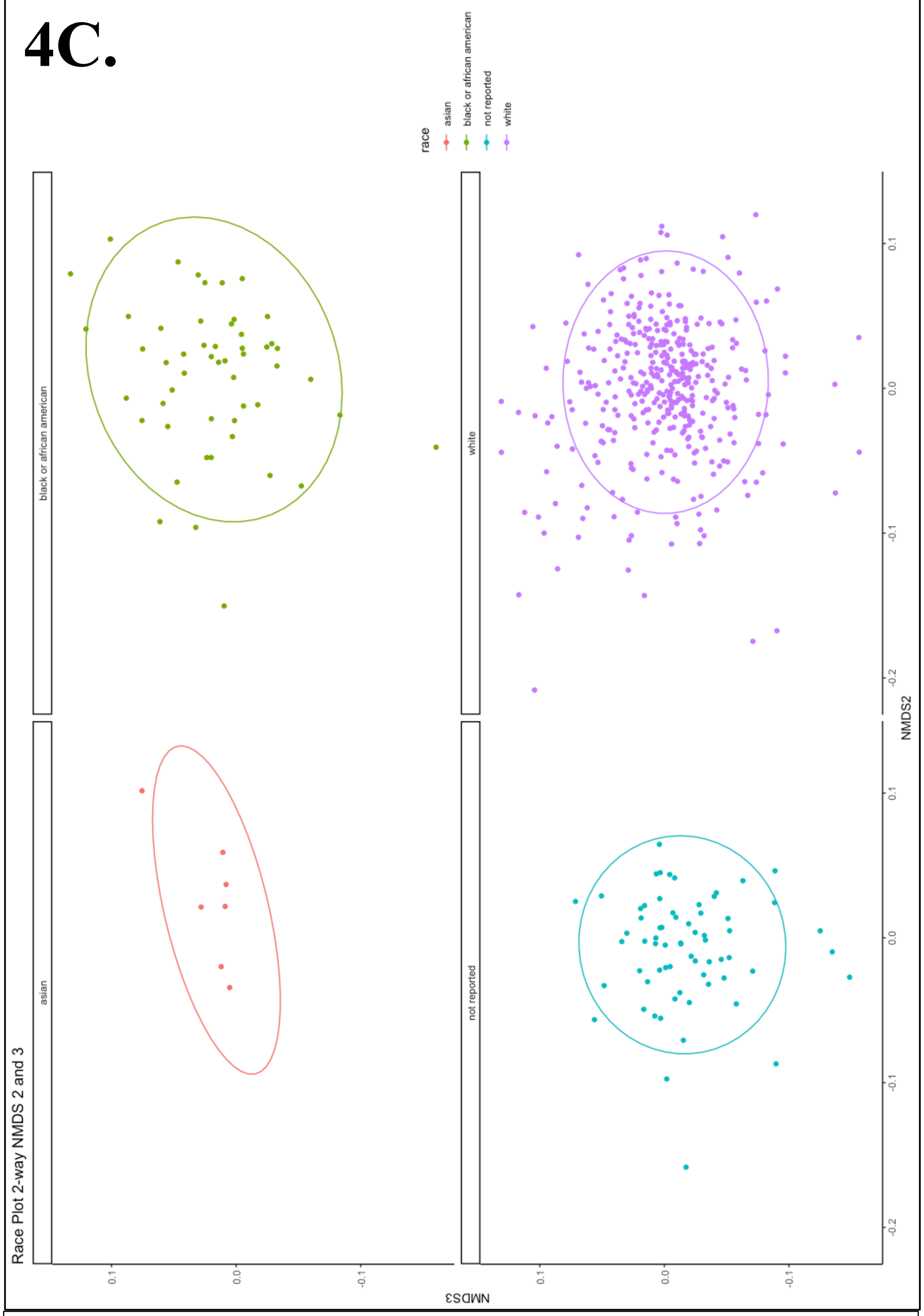
4A.


Race Plot 2-way NMDS 1 and 2
asian
black or african american
not reported
white
race
asian
black or african american
not reported
white
NMDS1
NMDS2
-0.2
0.0
0.2
0.4
0.1
0.0
-0.1
-0.2

4B.


Race Plot 2-way NMDS 1 and 3
asian
black or african american
not reported
white
race
● asian
● black or african american
● not reported
● white
NMDS3
0.1
0.0
-0.1
NMDS1
-0.2
0.0
0.2
0.4

4C.



**Appendix 4 A-C.** Patient race gene expression ordination plots found in Figure 4 facet wrapped to show the differences within each individual race studied.