The determination of biological trends in Bladder Urothelial Carcinoma

Anjali Menon, Hayden Scott, Rhea Kaul
Group 11

Abstract

Bladder Urothelial Carcinoma is one of the most common cancer types, representing the most common bladder cancer histotype [1], and further study of its trends would allow for better treatment. This study aims to isolate subgroups using genetic mutation data from the TCGA PanCancer Atlas and perform survival analysis on to determine if diagnosis age plays an important role. This is done using data pre-processing and clustering, as well as survival analysis. From this, clear clusters were obtained from RNAseq data and relatively unclear clusters were obtained from DNA mutation data, implying differences in transcription as diagnosis age increases, and 10 driver genes were obtained, with expression changing above and below the age of 65. Survival analysis resulted in inconclusive data regarding correlation between length of survival and diagnosis age.

Introduction

Carcinoma develops within an individual as a result of genomic faults, which cause rapid uncontrolled cell growth and affect other regulatory pathways within the affected cells. Broadly, this may occur due to exposure to carcinogenic substances or quite simply due to mutations acquired naturally during a lifetime. This combination of rapid growth and lack of regulation leads to the formation and potential spread of tumours within the individual.

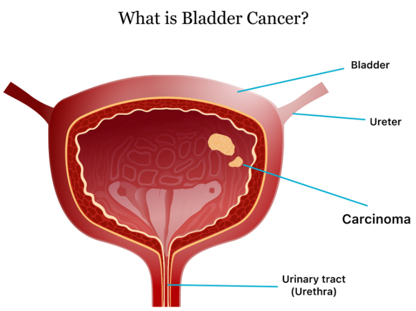


Figure 1: Diagram of a developing Carcinoma (1)

Bladder Urothelial Carcinoma (referred to henceforth as UCB) represents a significant amount of all bladder cancer cases and develops in the bladder's innermost cell layer known as the Urothelial cells [2] as seen in figure 1. Urothelial cells tend to expand when the bladder is full and contract when empty. These cells also line other parts of the body such as the ureters and urethra. For this reason, UCB is prone to metastasis around the bladder area.

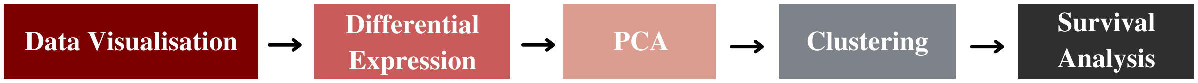
According to a recent study, age is the greatest single risk factor for developing UCB. [4] Furthermore, it was found that while men are 3 times more likely to develop UCB than women, women are predisposed with a more advanced case of the disease and have a lower survival rate [4].

The median diagnosis age of UCB is approximately 70 years and is considered a disease primarily affecting the senior population [5]. Due to the correlation between age and diagnosis of UCB, this poses a great challenge for the Canadian healthcare industry. Over the next 20 years, Canada's elderly population of age 65 and older is expected to grow by 68% [6]. For this reason, it is essential that such trends be analyzed so that the industry becomes equipped with technology to diagnose and cure such patients.

The goal of this research is to identify any correlation between gene mutation and the diagnosis age of patients. By using differential expression methods, we aim to determine when genes are more likely to mutate and develop cancer in a patient's lifetime. The results of this investigation are hoped to be used in the personalized diagnosis of patients and in preventive medicine.

Methods

Our analysis process was divided into the 5-stage process as seen below:



Data Visualization

In the preliminary data visualization stage, the data is plotted using pie graphs and histograms to identify any interesting trends. Next, the genes which were most frequently mutated were identified by determining the number of mutations across patients. These values were plotted and through visual examination of the graph, it was found that the first 300 most frequently mutated genes were selected for further analyzation.

Differential Expression

RNA sequence data was first accessed for the first 300 genes selected in the data visualization stage. A count matrix was formed consisting of these names and was compared with the clinical patient data to identify the age of each patient. The ‘age_at_index’ variable was used since this is the age at which the patient is first diagnosed with UCB. genes that were not selected as the first 300 were then discarded and the patient data was split into 2 categories: less than or equal to 65 years or over 65 years of age. The formation of these two categories is important to determine quantitative changes in expression levels between experimental groups. A volcano plot (Figure 4) was then made to determine any distinct correlation between the 2 groups. In the next stage, the number of results that correlate to a statistical p-value of 0.05 was found. A p-value of 0.05 indicates that these results were statistically significant. A new data frame was constructed to store these results including their gene names, symbols and Entrez Gene ID values. The log fold change was calculated to determine the degree of gene expression change and the data frame was sorted to reflect this. The five most up and down-regulated genes were found and used to generate a heatmap (Figure 3 [1]) after logarithmic normalization and hierarchical clustering. This was performed by using the DESeq2 package which conducts differential gene expression based on the negative binomial distribution. Differential expression analysis is pertinent to our research question as it helps us determine changes in expression levels between age groups diagnosed under 65 years and over 65 years. This can help us identify which genes are most commonly mutated and whether there are any trends in the data.

PCA | Clustering

PCA analysis was performed using the mutation patient data and RNA sequence data on patient age and lifespan after diagnosis. There were five major categories determined for patient age using the age_at_index variable (0-40, 40-60, 60-70, 70-80, 80+) after visual analysis of the data using the histogram (Figure 2 [2]). Each patient's vitality status was also retrieved to determine whether the patient is alive or deceased). In the mutation patient data, pre-processing of the data was performed first to ensure that the data matched that of the clinical TCGA dataset. To determine which genes were primarily being mutated and filter the corresponding samples, the melt and cast methods were used. In the RNA sequence data, the method was identical to that of the differential expression data however, the data were divided into five categories.

PCA was performed on the age_at_index variable and plotted using the ggplot() method. K-means clustering was then performed to determine any correlation between each of the 5 groups separately and the results were graphed using the fviz_cluster() method respectively as seen in figures 5 [1] and 6 [1].

Survival Analysis

Survival analysis was performed on the clinical patient data to analyse how survival is influenced by diagnosis age. 5 variables we used for the analysis: patient identifier, diagnosis age, disease specific survival status, months of disease specific survival, and last communication contact. The survival was measured in months and the last communication contact variable was converted from days to months to account for this. To ensure all patients were accounted for, months of disease specific survival was used only for patients who died with a tumor and last communication contact was used for all other patients. This ensured that survival analysis only accounted for deaths that were caused due to UCB complications and not other natural causes. Survival analysis was conducted for each group using survfit() and Surv(). The obtained data was plotted using ggsurvplot() to make comparison easier (Figure 7 [1]).To make visualization easier, the age groups were divided into sections of 10 years starting from 30 to 40 and ending at 80 to 90. As there was only one patient who was diagnosed at the age of 90, they were included in the 80 to 90 group.

Results

Table 1: Top Five Upregulated and Downregulated genes used for Differential Expression Analysis

Upregulated genes		Downregulated genes	
Symbol	Official Full Name	Symbol	Official Full Name
DNAH10	dynein axonemal heavy chain 10	SPHKAP	SPHK1 interactor, AKAP domain containing
CHD9	chromodomain helicase DNA binding protein 9	LRP2	LDL receptor related protein 2
FBXW7	F-box and WD repeat domain containing 7	FLG2	filaggrin 2
MUC17	mucin 17, cell surface associated	MUC4	mucin 4, cell surface associated
MALAT1	metastasis associated lung adenocarcinoma transcript 1	MUC16	mucin 16, cell surface associated

Through PCA analysis of gene mutations, as seen in figures 5,2 and 6,2,[*] one can see little statistical difference between the 2 sources of data being treated, DNA mutations and RNA sequencing data respectively. However, one observable trend is the difference in variation in the data. Examining the shape and quality of the clusters in figures 5,1 and 6,1 reveals that this between the 2 data types results in a large difference in how usable the data is, which in turn affects the quality of the clusters. The 2 main clusters using DNA mutation overlap significantly, and the other 3 clusters simply represent outliers

In contrast to the gene mutations data, the analysis done using RNAseq data generally shows a much larger variance, with the PCA plot having a much higher range and the clusters being far more disparate. This separation between the 5 clusters allows for a clear visual distinction, despite some overlap still being present. This may indicate that the variation between patients that results in subtypes within the UCB histotype takes place in the RNA transcriptomes, not the DNA itself.

[*] Figures are attached in Appendix A

Additionally, as seen in the heatmap in figure 3[2], it can be observed that upregulated genes are mostly less expressed than downregulated genes. This fits with expected genetic trends, as downregulated genes are affected early on in the translation process. Furthermore, as seen in table 1, it was noted that one of the primary upregulated genes is MALAT1, which is a transcriptional regulator involved in cancer metastasis and cell migration [8]. Of these upregulated genes, MALAT1 is the most commonly expressed, with CHD9 and FBXW7 following after. An additional observation made with this heatmap is that SPHKAP is expressed in very few samples.

Lastly, by use of survival analysis, it was found that there is no direct correlation between survival and diagnosis age. However, the general trend observed was that people who were diagnosed at an older age had a lower survival rate. It is unclear whether it is due to natural causes or due to UCB progression and complications. The group with the highest probability rate was group 30-40 (figure 7[2]), which may be due to the small sample size of that group. The two groups with the lowest survival probability were 50-60 (figure 8[1]) and 80-90 (figure 8[2]). This is expected for the latter group but similar results were seen in the 50-60 age group, which cannot be easily explained by natural and biological factors.

Discussion

After undertaking a comprehensive analysis of the TGCA Bladder Urothelial Carcinoma database, trends related to activity/mutation of certain genes were isolated and any correlation between gene mutation and the diagnosis age of patients were identified

Data Trends:

Most interestingly, this study found no obvious and direct correlation between diagnosis age and length of survival, besides slight differences in the range of survival length between age groups. Additionally, this study's use of Principal Component Analysis and K-Means Classification found that despite a low variance in the tumors' DNA between patients, there was some noticeable distinction between patients' tumorous RNA transcriptomes, possibly leaving the possibility that differences causing subtypes of UCB lie within the transcription process. This, as well as the prevalence of mutations affecting MALAT1, a gene responsible for regulating transcription relating to cell migration, seems to imply that the genomic faults that cause UCB affect mRNA more than they affect DNA. On the topic of affected genes, a noticeable trend is the fact that, despite 300 genes total being affected by some mutation across the entire patient population, there are only roughly 10 that are broadly expressed. This may be because most of the 300 genes affected are only passenger genes, and do not worsen the carcinoma themselves. However, among common cancer types, 10 driver mutations is relatively high. This follows with noticed trends in another study from the university of Illinois, which found that bladder cancer has a high number of driving mutations when compared to other types of carcinoma [10].

Biological/Environmental factors:

Bladder cancer has a complex chromosome number and structural variation. Partial or complete loss of genetic material at chromosome 9 can lead to the loss of genes such as p16 that act as tumour suppressors and indicators of recurrences of low-grade bladder cancer [11]. Chromosome 3,7,13, and 17 abnormalities are also common in bladder cancer[11]. This can help explain the low variance between the tumour's DNA due to the fact that the complex structural variation is very common.

One notable gene was MALAT1, as shown by the prevalence of mutations affecting it in the patient data. A study shows that MALAT1 plays a role in the mechanism by which TGF-β promotes tumour invasion and metastasis by inducing an epithelial-mesenchymal transition [12]. This explains why MALAT1 is an upregulated gene in bladder cancer that was used for differential expression analysis. Gene expression of CHD9 and FBXW7 is similar for both patients older than 65 and younger than 65, but DNAH10, FLG2, MUC16, MUC17, LRP2, and SPHKAP are expressed to a lesser extent in the older group. MALAT1 and MUC4 are both expressed at a higher amount in patients older than 65. This difference in genetic expression means the viability of drugs targeting specific genes will vary with age. The fact that 10 driver genes were found aligns with accepted literature despite being much larger than the usual amount of driver genes caracteristic of other cancers.

Furthermore, there are also other biological factors that may have caused the variation in the data set, including bias in the race proportion due to demographics of a predominantly Caucasian greater population as seen in figure 9, which may not be representative of the general population as a whole. This causes the data to appear skewed and may lead to the false interpretation that Caucasian people are predisposed to UCB.

Challenges

Throughout the duration of the study there were several challenges posed due to limitation in datasets and the size of the data. When conducting PCA analysis for the clinical dataset, age related information was present for almost all patients, however there were three that did not have age as a variable. For this reason, those variables were set to 0 automatically to avoid null point errors during the analysis. Furthermore, when attempting to conduct linear regression, the large dataset of diagnosis age made it harder to find a direct correlation as taking the average of ages in each 10 year group would not be representative of the population as a whole. For this reason, linear regression was not taken into consideration due to non representative results.

Future Work

The results of survival analysis in this study were relatively inconclusive. Similar results of lowest survival probability were determined for the 50-60 and 80-90 age groups. In order to analyse the biological reasoning behind this result, it would be necessary to determine which tumour stage is the most prominent amongst these groups. From this, a sister study may be produced which can help to determine whether survival probabilities change based on tumor stage and age.

Conclusion

According to this study's data, it is unreasonable to claim that there is any clear correlation between diagnosis age and length of survival after diagnosis. A further study that considers tumor stage as an additional feature may produce results that better ascertain the relationship between diagnosis age and survival length.

The aim of this study was to identify any correlation between gene mutation and the diagnosis age of patients and through the use of differential expression analysis, determine which genes are most likely to mutate and develop cancer. The results of this investigation are hoped to be used in the personalized diagnosis of patients and in preventive medicine.

Contributions

Data Visualization: Rhea Kaul
PCA: Anjali Menon
Clustering: Anjali Menon
Survival Analysis: Rhea Kaul
Differential Expression: Anjali Menon
Report:

- Abstract: Hayden Scott
- Introduction: Anjali Menon
- Methods: Anjali Menon, Rhea Kaul
- Results: Hayden Scott
- Discussion: Rhea Kaul, Hayden Scott, Anjali Menon
- Conclusion: Anjali Menon, Rhea Kaul, Hayden Scott

Presentation:

- Script: Rhea Kaul, Hayden Scott
- Slides: Anjali Menon

References

[1] “Bladder Cancer: Introduction”, cancer.net. https://www.cancer.net/cancer-types/bladder-cancer/introduction [Accessed: November 8th]

[2] “What is Bladder Cancer?”, Cancer.org. https://www.cancer.org/cancer/bladder-cancer/about/what-is-bladder-cancer.html [Accessed: November 8th]

[3] “Bladder Cancer Treatment”, National Cancer Institute. https://www.cancer.gov/types/bladder/patient/bladder-treatment-pdq [Accessed: November 21st]

[4] Shariat, S. F., Sfakianos, J. P., Droller, M. J., Karakiewicz, P. I., Meryn, S., & Bochner, B. H., “The effect of age and gender on bladder cancer: a critical review of the literature”, in BJU international, 105(3), 300–308, 2010. https://doi.org/10.1111/j.1464-410X.2009.09076.x [Accessed:]

[5] Messing E.M., “Urothelial tumors of the bladder”, in Campbell-Walsh Urology. Wein AJ, Kavoussi LR, Novick AC, Partin AW, Peters CA, editors. Ninth Edition. Saunders-Elsevier; Philadelphia: 2008. pp. 2407–46, Chapter 75. [Accessed:]

[6] “Infographic: Canada’s seniors population outlook: uncharted territory”, Canadian Institute for Health Information. https://www.cihi.ca/en/infographic-canadas-seniors-population-outlook-uncharted-territory

[7] Cerami et al, “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data”, Cancer Discovery, 401. PubMed. May 2012 2 [Accessed: November 3rd]

[8] Gao et al, “Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal”, in Sci. Signal, 6, pl1 (2013). PubMed. [Accessed: November 3rd]

[9] “Metastasis Associated Lung Adenocarcinoma Transcript 1 database, MALAT1”, in National Center of Biotechnology Information. https://www.ncbi.nlm.nih.gov/gene/378938 [Accessed: December 5th]

[10] J. Iranzo, “Cancer-mutation network and the number and specificity of driver mutations”, in PNAS, https://doi.org/10.1073/pnas.1803155115 [Accessed: December 7th]

Figure 1: Graphic of a carcinoma developing within a bladder, Drugwatch https://www.drugwatch.com/health/cancer/bladder-cancer/

[11]Y. Li, L. Sun et al, “Frontiers in Bladder Cancer Genomic Research”, in Frontiers in Oncology. https://doi.org/10.3389/fonc.2021.670729 [Accessed: December 7th]

[12]Y. Fan, B. Shen, et al, “TGF-β Induced Upregulation of MALAT1 Promotes Bladder Cancer Metastasis by Associating with SUZ12”, in Clinical Cancer Research. https://doi.org/10.1158/1078-

Appendix A: Graphs

Data Visualization

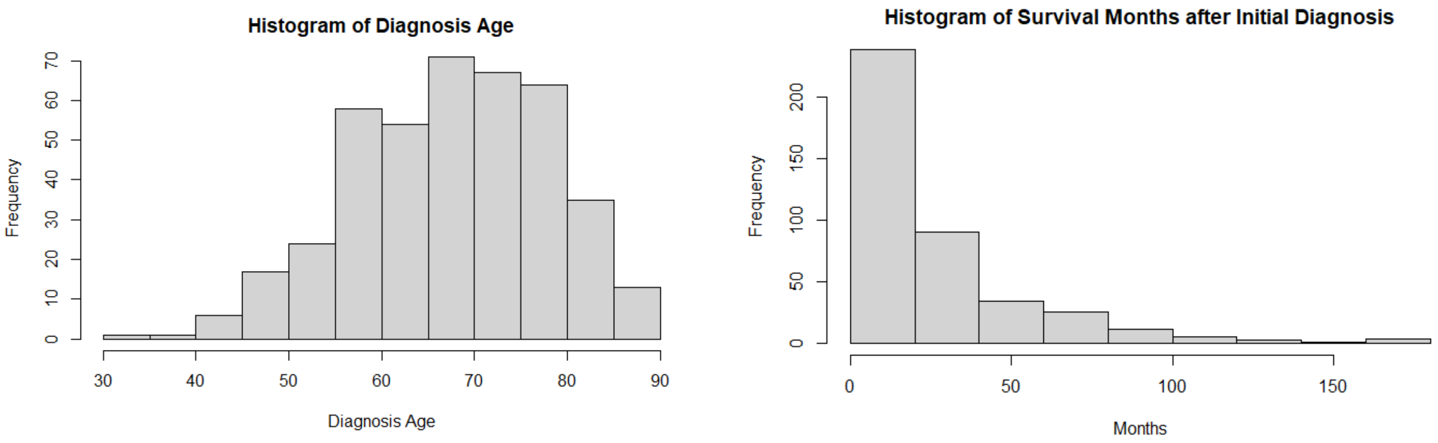


Figure 2: [1] Survival months after initial diagnosis of UCC patients (Right) and [2] Frequency of patients diagnosed with UCC at a specific age (Left)

Differential Expression

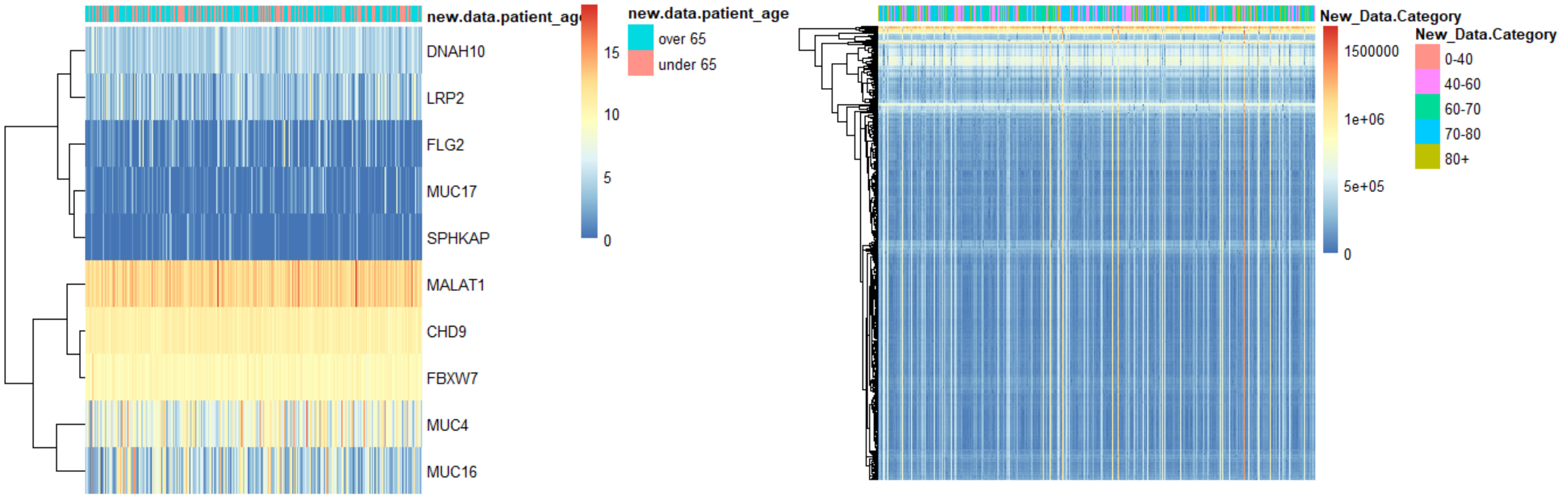


Figure 3: [1] Heatmap of top 300 expressed genes (Right) and [2] Heat map of top 5 upregulated and downregulated genes (Left)

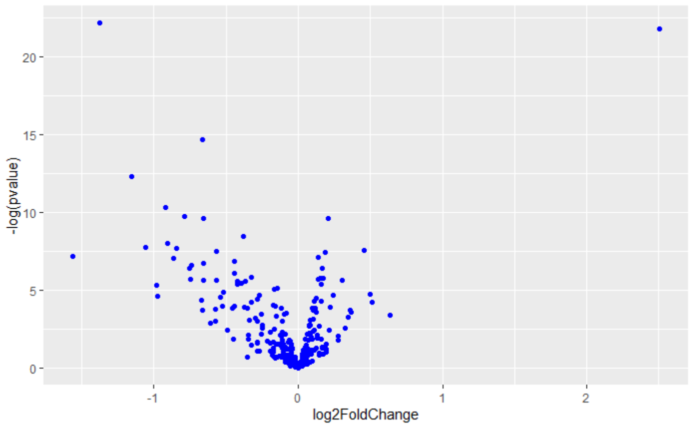


Figure 4: Volcano plot depicting the statistical significance versus the magnitude of change

PCA | Clustering

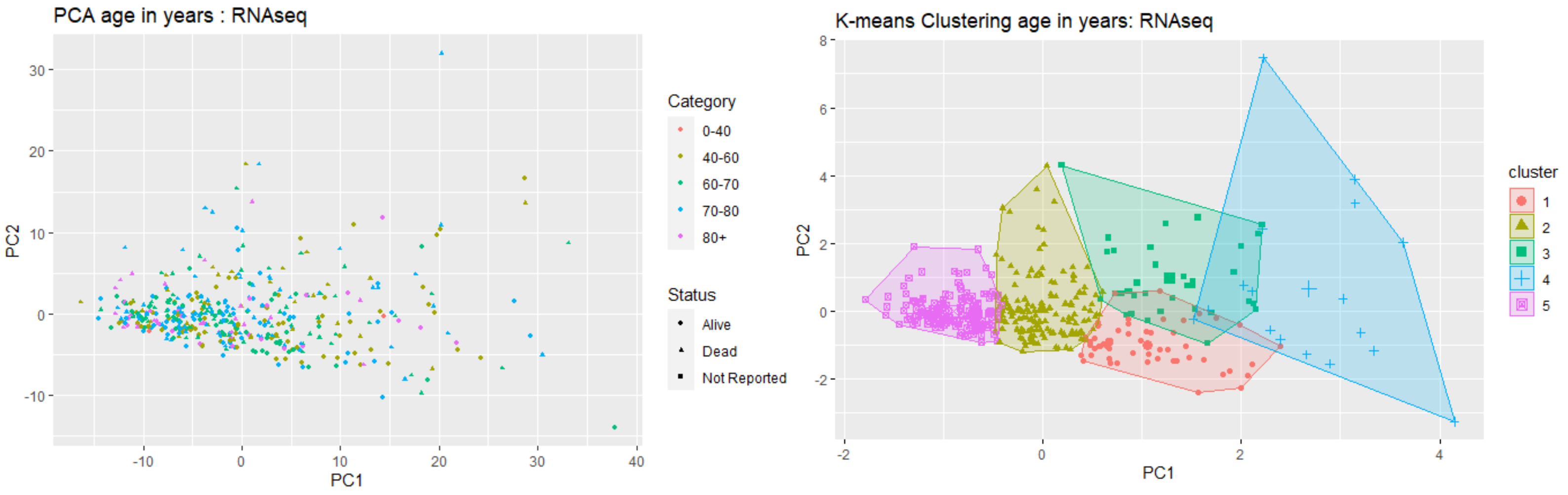


Figure 5: RNA sequenced data: [1] K-means clustering plot of age in years (right) and [2] PCA plot of age in years along with vital status

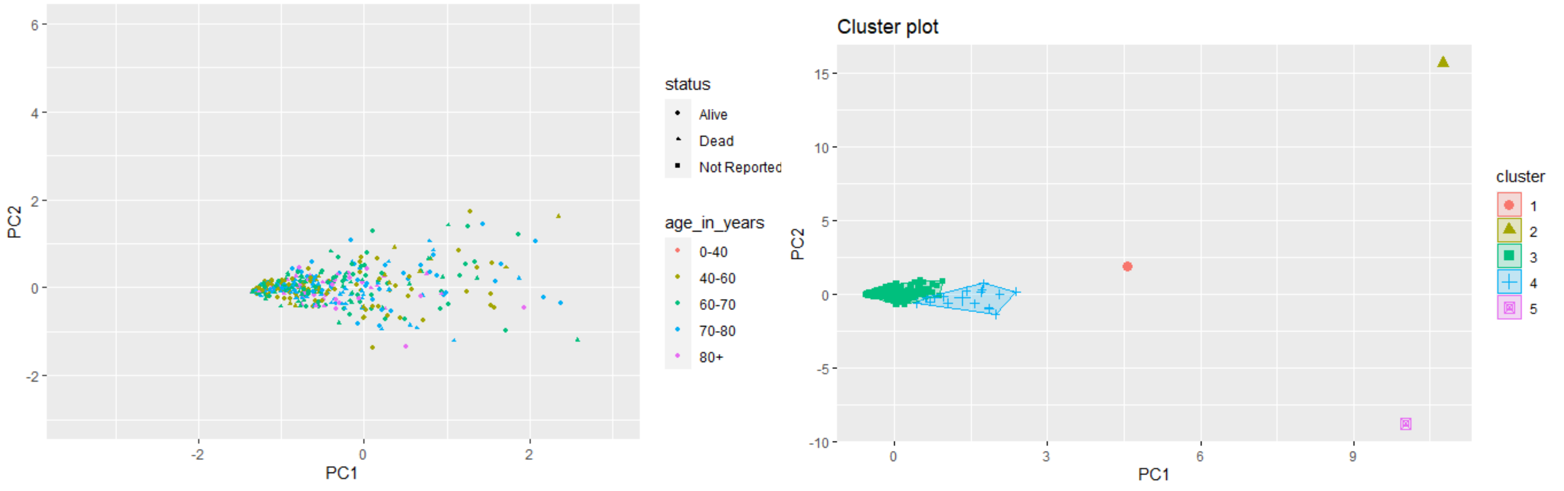


Figure 6: Clinical Mutation data: [1] K-means clustering plot of age in years (right) and [2] PCA plot of age in years along with vital status

Survival Analysis

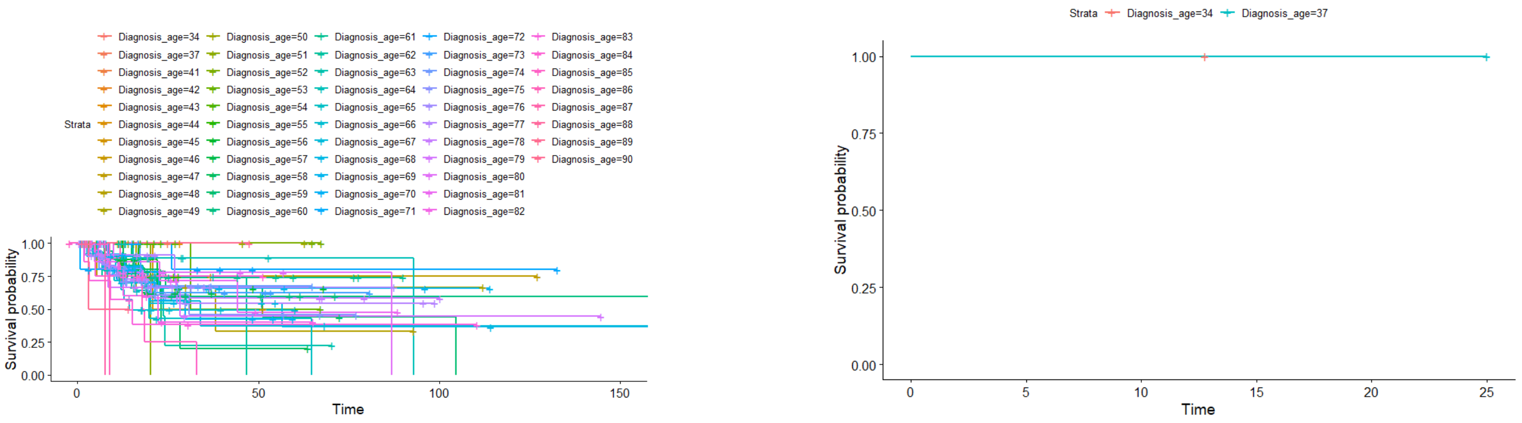


Figure 7: Survival Analysis data: [1] Survival analysis graph for all diagnosis ages (Left) and [2] Survival analysis graph for diagnosis ages 30-40 (Right)

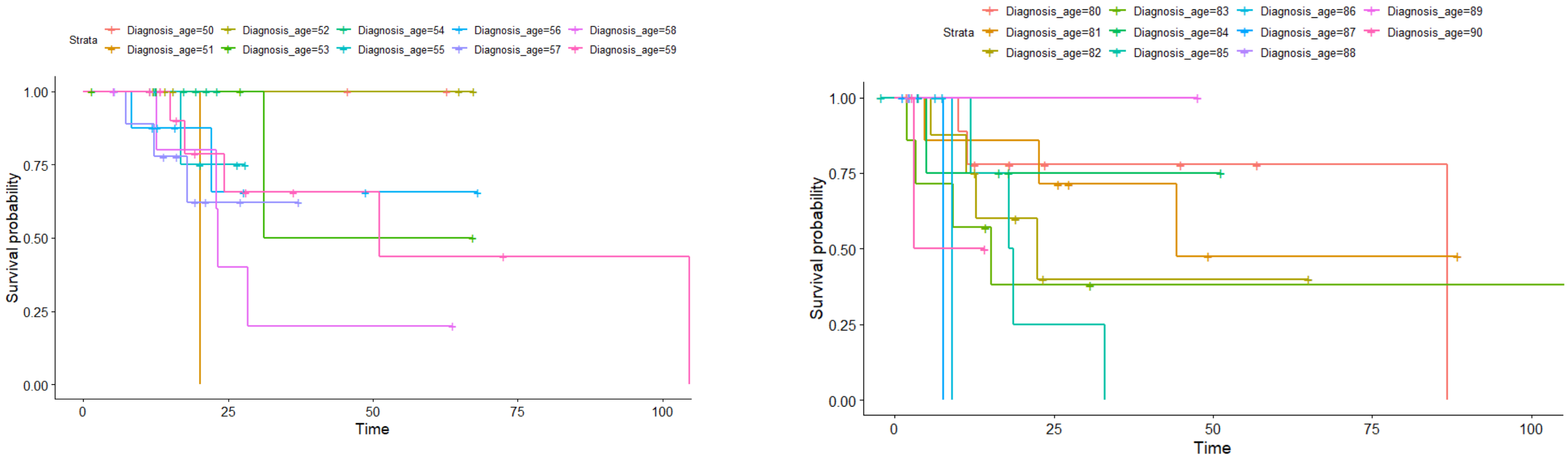


Figure 8: Survival Analysis data: [1] Survival analysis graph for diagnosis ages 50-60 (Left) and [2] Survival analysis graph for diagnosis ages 80-91 (Right)

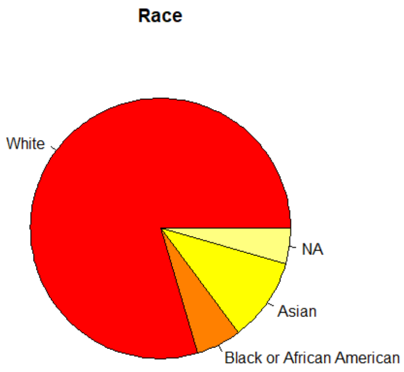


Figure 9: Pie chart showing racial distribution of data