BMEG 310 Final Project

An Analysis on Lung Squamous Cell Carcinoma Patient Samples

Group 19: Arnav Savla, Siqi Da, Ruini Xiong

Abstract

This research project investigates how clinical factors and genes influence the patient survival of lung squamous cell carcinoma patients.

Introduction

Lung Squamous Cell Carcinoma (LUSC) can occur in many parts of the lungs including the bones, and the lymph nodes in, around and between the lungs. However, carcinoma can mostly be found in the trachea and the bronchi. This type of lung cancer is genetically heterogeneous. In most cases, LUSC firstly occurs in the epidermal layers, where any response to specific types of epithelial injury could cause mutations [1].

In this research report, we will discuss the analysis of 487 lung squamous cell carcinoma patient samples on the TCGA Cancer dataset. We try to analyse different RNA expressions through the analysation of clinical data, mutation data and RNA sequence data. The key analysis focuses on the relevance in mutation, expression, and clinical data. The goal of the study is to show how particular trends may relate to characteristics of certain genetic changes, how clinical factors affect patient survival, and the potential mechanisms implicated.

Methods of Analysis

I. Principal Component Analysis (PCA) on Mutation Data

Performed PCA on the mutation data, specifically on the isolated tumor reference and alternate allele counts (t_ref_counts and t_alt_counts respectively) to find the most dominant tumour causing gene mutations, namely through Hugo Symbol ID. In this PCA, we do not need to scale as we have already isolated two factors.

II. Survival Analysis on Clinical Patient Data

Performed survival analysis on the clinical patient data to find the factors that may affect patient survival time. The analysis was focused on correlations between survival time and gender, age, race and tumor stage respectively. We chose overall survival (OS) instead of other survival times, as OS is internationally accepted as being unambiguous, unbiased, and with a defined end point of paramount clinical relevance[2]. Kaplan-Meier plots and p-values of logrank test were produced to visualize the survival states.

III. k-Means Clustering on Patient Data

We first performed k-Means clustering on the Clinical patient data to find the highlighted relationship between OS and DFS months, Age and Day to the last follow-up. We analyse this data by clustering it, finding correlations in the data and observing how they influence each other. The analysis was focused on numerical data already provided in the data to obtain more than just binary values that can be obtained from text. We can then use the clusplot function in R to find the variation, which helps find the correlation in the data.

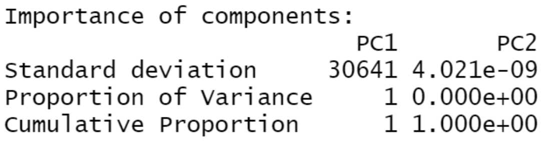
IV. k-Means Clustering on Mutations Data

We also conducted the same, k-means clustering, on the mutations data for Lung Squamous Cell Carcinoma. Again, we use numerical data to avoid simple binary values. Mutation data provides the mutated genes and their data to find the presence in tumor cells. We use clustering to, again, find the correlation between the data value provided so we can further explore this data.

Results

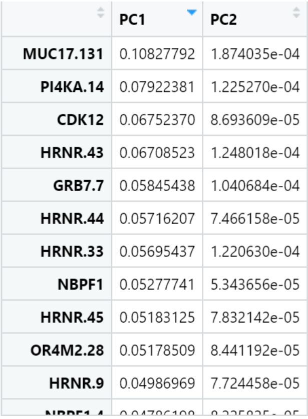
I. Principal Component Analysis (PCA)

Upon performing PCA on the subset containing these two factors, we find that PC1 has far greater dominance on the dataset with absolute value 21666.47 and 1.606237e-09, therefore we can neglect PC2 altogether.



```
Importance of components:
                         PC1       PC2
Standard deviation     30641 4.021e-09
Proportion of Variance     1 0.000e+00
Cumulative Proportion      1 1.000e+00
```

Taking the first column of the rotation from the PCA results, we can find the principal component scores for each mutation. These results show us our most dominant mutations in lung squamous cell carcinoma patients are MUC17.131, PI4KA.14, CDK12, HRNR.43, and GRB7.7.



	PC1	PC2
MUC17.131	0.10827792	1.874035e-04
PI4KA.14	0.07922381	1.225270e-04
CDK12	0.06752370	8.693609e-05
HRNR.43	0.06708523	1.248018e-04
GRB7.7	0.05845438	1.040684e-04
HRNR.44	0.05716207	7.466158e-05
HRNR.33	0.05695437	1.220630e-04
NBPF1	0.05277741	5.343656e-05
HRNR.45	0.05183125	7.832142e-05
OR4M2.28	0.05178509	8.441192e-05
HRNR.9	0.04986969	7.724458e-05
NBPF1.4	0.04786108	8.325835e-05

Figure 1: PC1 values for gene mutations in descending order of relevance.

II. Survival analysis on Clinical Patient Data

Sex and survival time

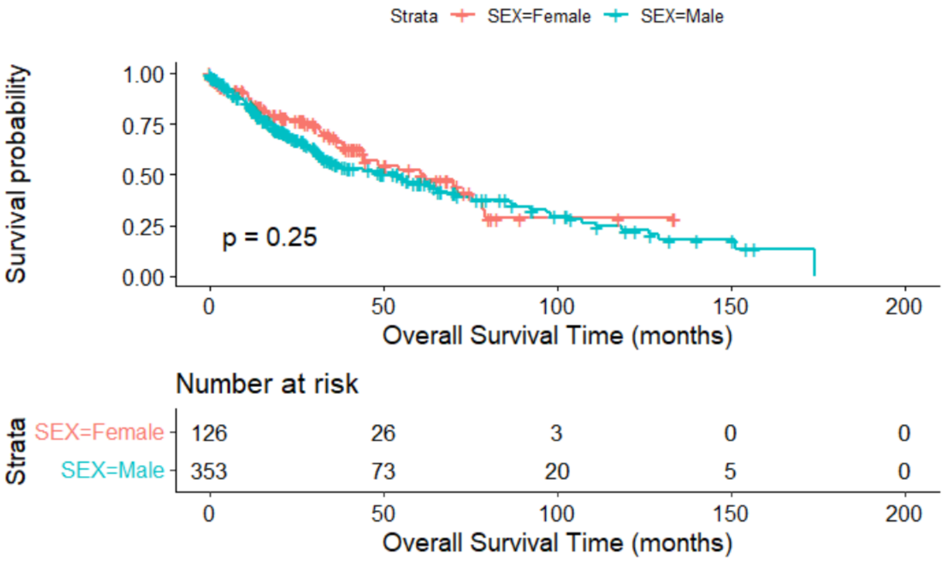


Figure 2: Survival analysis on patient sex and overall survival time

Survival analysis on patient sex and survival was performed. The p-value of logrank test is non-significant (>0.05), so the gender of the patient does not significantly affect survival time. Though the curve looks different after the 100-month mark, the at risk table below showed that most of the patients die or are censored before 100 months. Therefore, the overall survival states of patients won't be really affected by patient gender.

Age and survival time

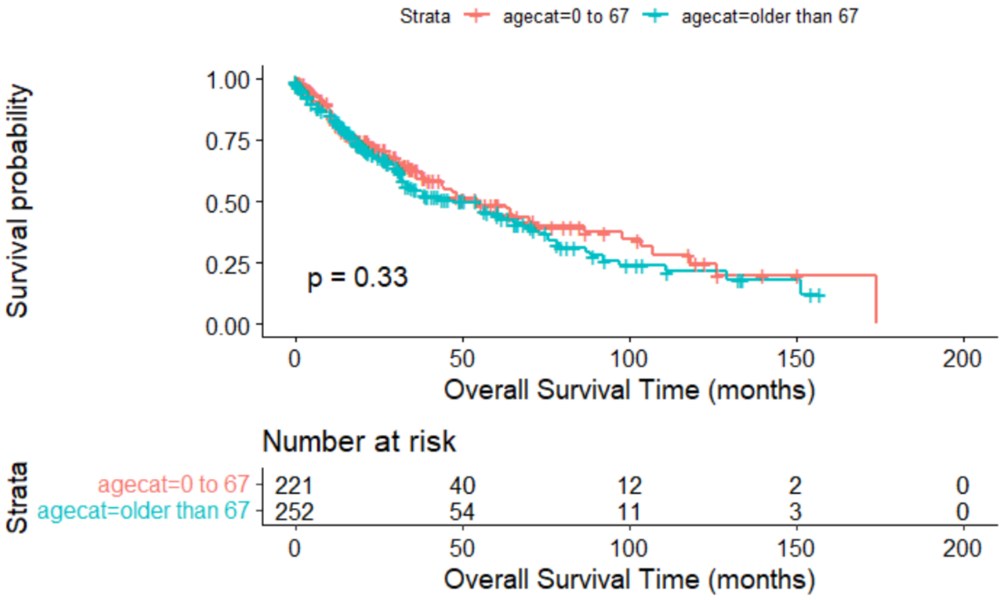


Figure 3: Survival analysis on patient age and overall survival time

Survival analysis on patient age and survival was performed. Patients were separated by the mean age (67) into two groups and compared survival probability. The p-value of logrank test is 0.33, which is non-significant (>0.05), so the age of the patient does not significantly affect survival time.

Race and survival time

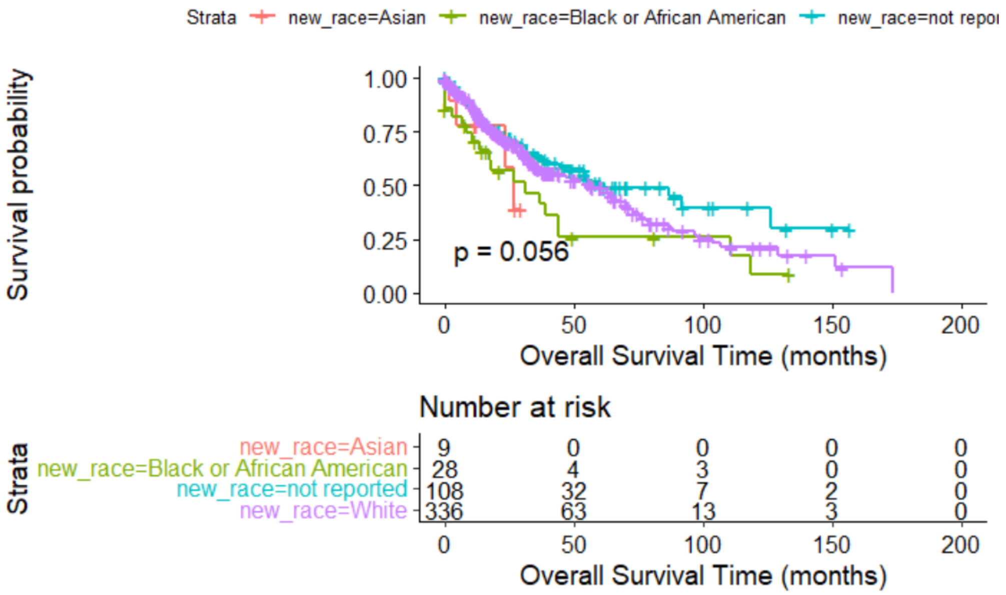


Figure 4: Survival analysis on race and overall survival time

Survival analysis on patient race and survival was performed. The data was separated into four groups: Asian, Black or African American, White and not reported. According to the Kaplan-Meier plot, Black or African American patients have a worse survival probability. The p-value of logrank test is small and close to 0.05. We can suppose that race is related to survival time. However, the sample numbers of races differ a lot and most samples are white. So p-value may not be reliable. The effects of race to survival probability still cannot be determined.

Tumor stage and survival time

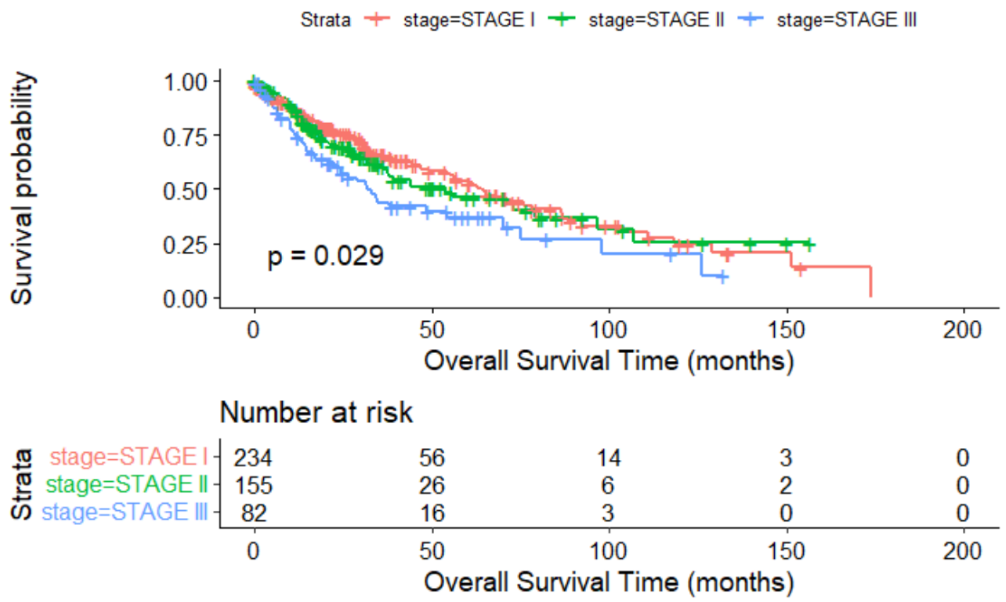


Figure 5: Survival analysis on tumor stage and overall survival time

Survival analysis on the tumor stage and survival time was performed. The tumor sub-stages data were joined together to increase the group size and increase the reliability of logrank test. In this case, the p-value is small (<0.05), so tumor stage will affect survival and patients in later tumor stage have a worse survival probability.

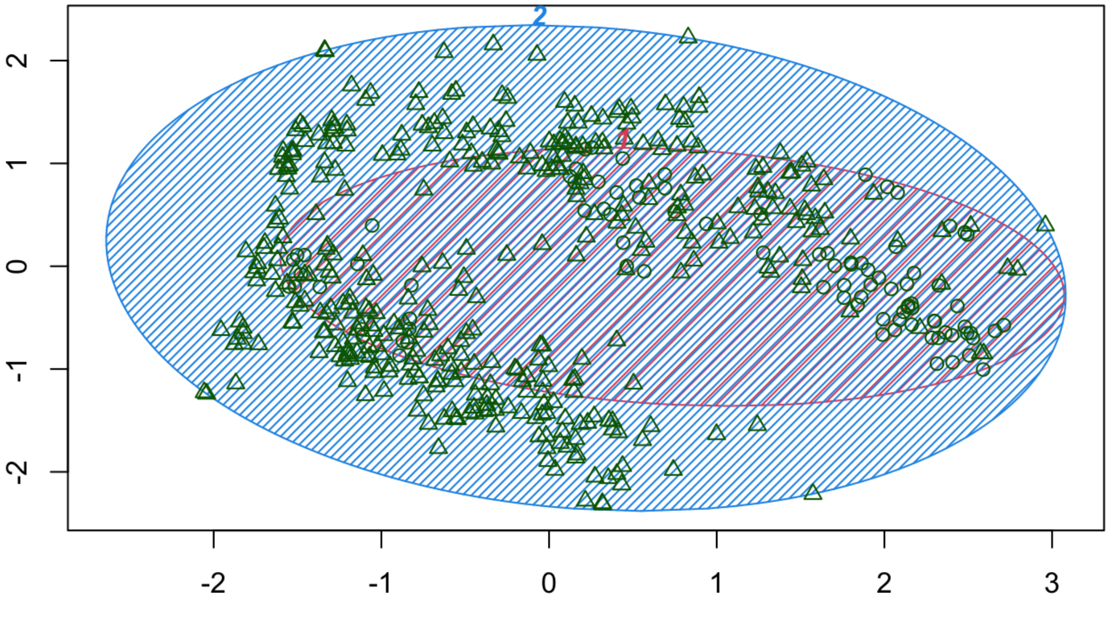
Clustering the Data

III. k-Means Clustering on Patient Data

This data was primarily clustered for the presence of numbers in the data. Initially, we wanted to find the correlation between the disease free survival (DFS_MONTHS) months of the 487 patients and the number of months of treatment (OS_MONTHS)[3]. This gave us a 100% variation in the data with two smooth line clusters. We then decided to cluster the 2 with days since the patients' last follow up, giving us an 82.38% variability.

Our team then thought we could do better, clustering 4 things, adding age into the mix to get 63.24%.

We observed 2 strong clusters forming, proving a strong correlation between age, treatment time, days since the last followup and disease free survival months.



These two components explain 63.24 % of the point variability.

Figure 6: 2D k-means clustering for OS Months, DFS Months, Age and Days since last follow-up

IV. k-Means Clustering on Mutations Data

We start with clustering t_ref_count, which is the read depth supporting the reference allele in tumor BAM and t_alt-count, read depth supporting the variant allele in tumor BAM. Again, this gave a variation of 100%. To then reduce the variation, we start by clustering the n_ref_count, read depth supporting the reference allele in normal BAM (cleared in somatic MAF) which gives us a variation of 93.07%. This variation of three factors gives us one cluster. We can then cluster with a 4th variable, n_alt_count, the read depth supporting the variant allele in normal BAM (cleared in somatic MAF). This still gives us a variation of 80.03% so we can add another data column. We then add NCALLERS to the cluster, which is the number of calls.

This gives us a cluster of 5 factors, giving a variation of 68.21% and shown in figure 7 below.

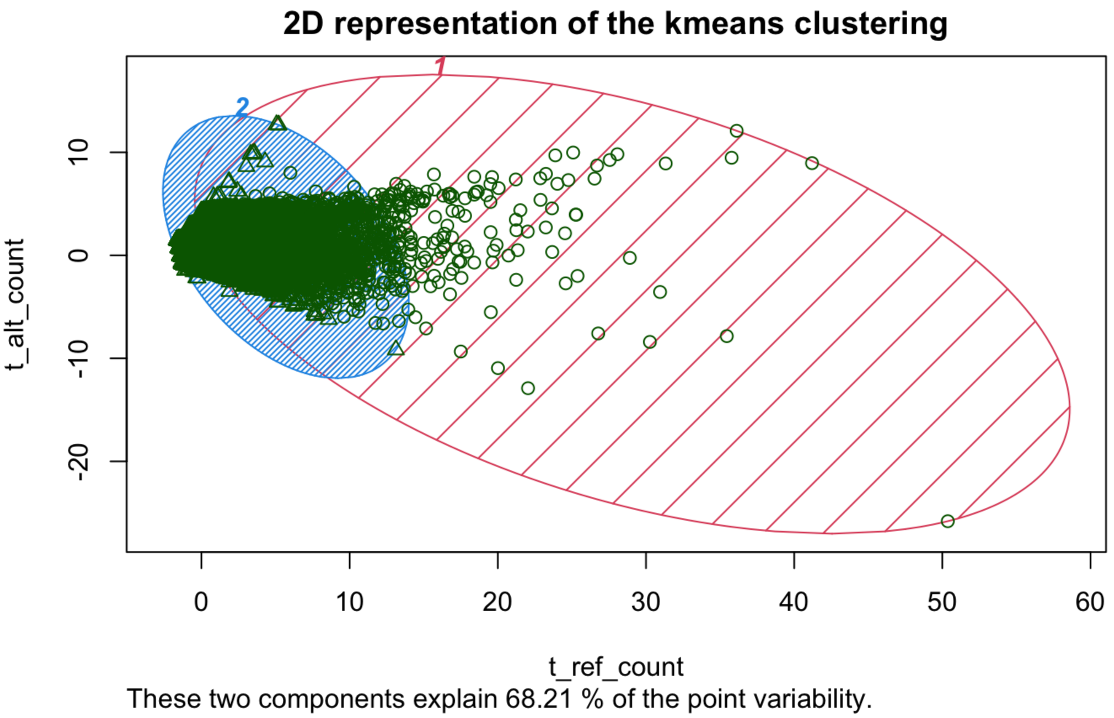


Figure 7: k-Means cluster of 5 data columns, t_ref_count, t_alt_count, n_ref_count, n_alt_count, NCALLERS.

Discussion

Passing the mutation data through a principal component analysis (PCA), we find the mutations MUC17.131, PI4KA.14, CDK12, HRNR.43, and GRB7.7 most dominant. According to literature, the above named mutations are relevant to LUSC as follows:

1. MUC17.131: Mucin 17 is a cell surface associated gene, very common in cancers starting from the epithelial layers.
2. PI4KA.14: P14KA is a protein-encoding gene that focuses on membrane proteins. Mutations causing overexpression of this gene identified as cellular factors associated with cancer progression. [4]
3. CDK12: CDK12 is altered in 1.72% of squamous cell lung carcinoma patients, making the mutated gene an extensive factor in small cell lung cancers [5]. It is generally a gene that encodes a protein that controls genomic stability and process of gene repair [5].
4. HRNR: promotes tumor progression in squamous cell lung carcinoma. [6]
5. GRB7: GRB7 was found to promote lung cancer progression [7]. In another study, high Grb7 expression was strongly associated with decreased survival [8].

Because these results match that of literature, we can conclude that the PCA had successfully found the most dominant mutations within the specific data set. However, there were a few differences from our results and the most common top 5 genes. It is possibly due to the specificity of the mutation IDs. For example, HRNR itself is a huge tumor profession promoter in squamous cell lung carcinoma. However, because it is separated into several different IDs (HRNR.44, HRNR.33, HRNR.45, HRNR.9…). Therefore, it may be difficult to find the actual top 5without further filtering or manipulation on the data.

Our k-Means analysis tries to partition the observed data into a certain amount of clusters. We can use k-Means analysis to characterize the data into relation clusters that can define the correlation that they present between different frames.

In the patient data, we can start observing the correlation shown by "OS_MONTHS", "AGE", "DAYS_LAST_FOLLOWUP", and "DFS_MONTHS". This is done to measure the correlation between age and the other data frames. Since most of the data is in the units of time, it is then easy to observe the relation between them. k-Means clustering reduces the inter-cluster distances to reduce the overall distances between points. Analysing the 4 variables, we see that there is an extensive correlation between age and the other 3 factors. To be specific, a 63.24% correlation can be seen which shows a strong correlation between the 4 variables. 2 clusters clearly form, showing a linear decrease in the values. As age increases, we see an increase in months since treatment began but we see a decrease in the days since the last follow up and in the disease-free survival months too.

In mutation data, we see a correlation appear between, "t_ref_count","t_alt_count", "n_ref_count", "n_alt_count", and "NCALLERS". This data primarily tries to find the relation between t_ref and n_ref along with a relation between t_alt and n_alt. The data then shows this cluster to be very strongly related, showing a variation of only 68.21% (the greater the variation, the denser the cluster), which shows strong linearity between n and t values.

Some of the limitations in clustering occur when clustering more than 3 datasets together, it becomes harder to read the clusters and analyse them successfully. We can then cluster 2 factors together to understand them better. For instance, the appendix shows the relation between t_ref and t_alt clearer than in figure 2. This is harder to do with patient data due to the incomplete sets of data provided.

k-Means clustering provides an estimation of the relationship between data, allowing for rough analysation that can lead to more advanced research. It essentially allows for the visualization of data position before further reading.

By survival analysis on clinical data, we can get the influencing factors affecting the survival probability of patients. The tumor stage figured out to have a correlation with patients' survival probability. As the Kaplan-Meier plot showed, patients in tumor stage III have an obviously worse survival probability. We got an overall p-value testing the null hypothesis that all the curves are similar at every time point. P-value of the logrank test is 0.029 (<0.05), so that we can reject the null hypothesis. Therefore, there is indeed a correlation between tumor stage and survival time, and the survival time of patients with more advanced tumors is shorter. For many types of cancer, the prognosis is often expressed as a survival rate. Survival rates are nearly always based on the stage of cancer at the time of diagnosis[9].

The survival analysis on race also gave out a small p-value, but there are some limitations of this analysis. The sample size difference is too large. The numbers of Asian patients and black or African American patients are far less than that of white patients, and nearly 30% of the race of patients was not reported. Therefore, the power of the logrank statistics is not strong enough and p-value is not very reliable. Although we can see from the plot, the survival probability of black or African American is lower than that of other groups, we still cannot conclude that there is a strong correlation between race and patient survival.

Contribution

Arnav Savla performed k_means clustering on both patient, and mutation data, Siqi Da performed survival analysis on the patient data, Ruini Xiong performed PCA on mutation data. Everyone worked on the report and slides together, finalizing ideas through call.

References

[1] C. F. Brainson and C. F. Kim, “Lung Cancer Stem Ce
lls: Drivers of a Genetically and Histologically Complex Disease,” 2016, pp. 149–175,.

[2] J. J. Driscoll and O. Rixe, “Overall survival: Still the gold standard: Why overall survival remains the definitive end point in cancer clinical trials,” *Cancer J.*, vol. 15, no. 5, pp. 401–405, 2009.

[3] “File Format: MAF - GDC Docs,” *Cancer.gov*. [Online]. Available: https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/. [Accessed: 10-Dec-2021].

[4] A. Ilboudo *et al.*, “Overexpression of phosphatidylinositol 4-kinase type IIIα is associated with undifferentiated status and poor prognosis of human hepatocellular carcinoma,” *BMC Cancer*, vol. 14, no. 1, p. 7, 2014.

[5] “CDK12 - My Cancer Genome,” *Mycancergenome.org*. [Online]. Available: https://www.mycancergenome.org/content/gene/cdk12/ . [Accessed: 10-Dec-2021].

[6] S.-J. Fu *et al.*, “Hornerin promotes tumor progression and is associated with poor prognosis in hepatocellular carcinoma,” *BMC Cancer*, vol. 18, no. 1, p. 815, 2018.

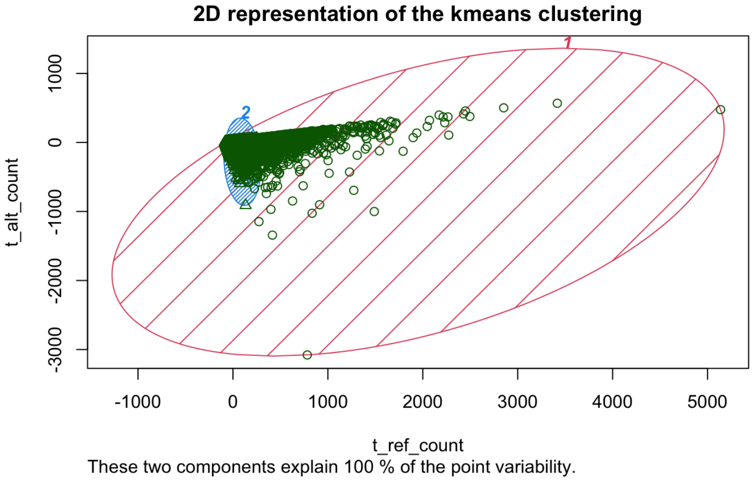
[7] Y. Nadler, A. M. González, R. L. Camp, D. L. Rimm, H. M. Kluger, and Y. Kluger, “Growth factor receptor-bound protein-7 (Grb7) as a prognostic marker and therapeutic target in breast cancer,” *Ann Oncol*, vol. 2010;21(3):466–473, 1093.

[8] P.-Y. Chu, Y.-L. Tai, M.-Y. Wang, H. Lee, W. H. Kuo, and T.-L. Shen, “EGF-activated Grb7 confers to STAT3-mediated EPHA4 gene expression in regulating lung cancer progression,” *bioRxiv*, 2020.

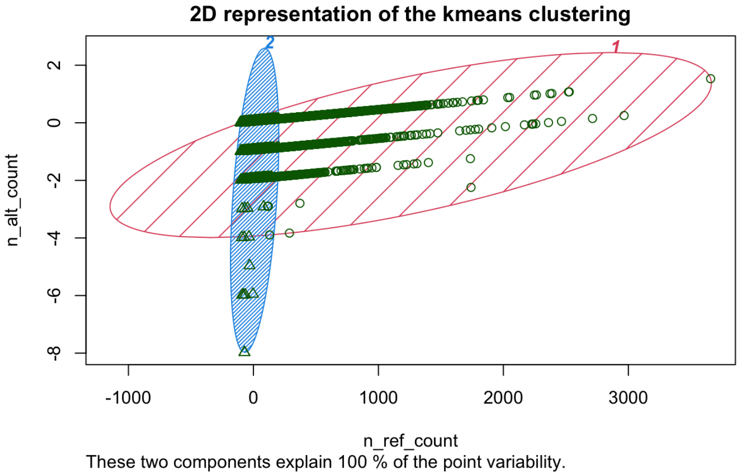
[9] “Cancer Staging,” *Cancer.org*. [Online]. Available: https://www.cancer.org/treatment/understanding-your-diagnosis/staging.html. [Accessed: 10-Dec-2021].

Appendix

[1] t_ref vs t_alt



[2] n_ref vs n_alt


2D representation of the kmeans clustering
n_alt_count
n_ref_count
These two components explain 100 % of the point variability.