BMI 701 Final Project: Analysis of TCGA-MESO Samples Based on Mutation Variants and Gene Expression

Xiao (Katrina) Liu
Harvard Medical School
xliu@hms.harvard.edu

Abstract

The TCGA dataset is comprised of a variety of patient sample information collected by many project. The project TCGA-MSEO in particular focused on mesothelioma, a type of cancer which mostly arises in the mesothelial surfaces of tissues in the pleura. In this project, sample information, gene expression level, and mutation variants information associated with the disease are collected by the TCGA-MESO project, providing the foundation of this study. We conducted principle component analysis and clustering to see if there are correlation between the gene expression data and mutational variants on the general level.

1. Introduction

Mesothelioma is a relatively uncommon cancer disease that arises in the mesothelial surfaces of tissues in the pleura and possibly in the peritoneum and the tunica vaginalis [2]. It has a long latency period, over 40 years from exposure to asbestos, and is difficult to diagnosis [1]. In this study, we hope to analyze the mutation variants information associated with mesothelioma and gene expression data of the collected samples of TCGA-MESO project to hopely provide insights of how they are associated with characteristics of samples through high throughput gene expression analysis.

2. Methodology

2.1. Data Source

The results here are in whole or part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga. TCGA-MESO project contains 74 samples characterized of 87 total samples in the marker paper, which is an ideal size for the scale of this study.

Figure 1 is a overview of how samples age group are related to vital status. It appears the majority of the cases in this project are dead, indicating the high fatality of the disease.

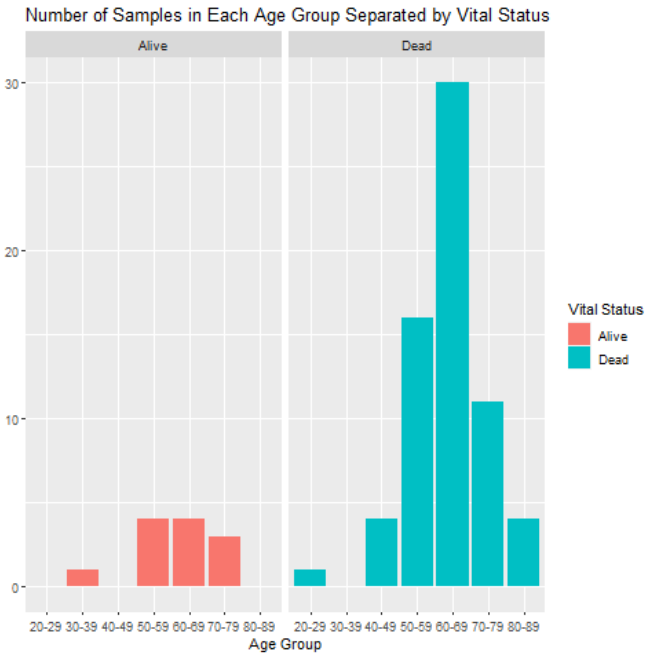

Figure 1

Figure 2 is a brief overview of the numbers of gene of each gene type included in this study. Three major gene types are lncRNA genes, processed pseudogenes, protein encoding genes. Therefore, a reasonable assumption for number of the principle components used in gene expression study would be around 4-6.

We want to focus on the impact level and consequences of each mutation variants, so we plot out the number of variants associated with each consequences separated by impact level in Figure 3. We noticed that, for the majority of the consequences, the variant associated with them falls into one impact level. The synonymous variant dominates the low impact level variants; the missense variant dominates

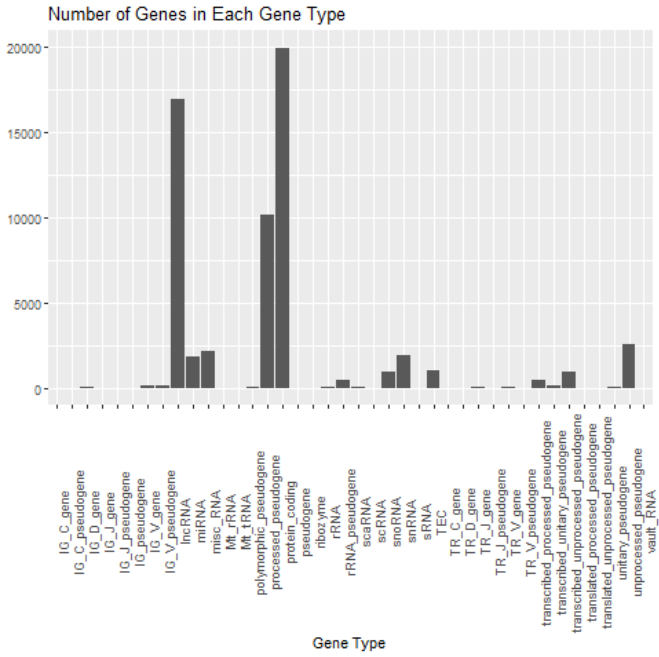

Figure 2

the moderate impact level; and the modifier impact level consists of only the intron variant. Therefore, it is reasonable to assume the consequences might be fitted into four clusters.

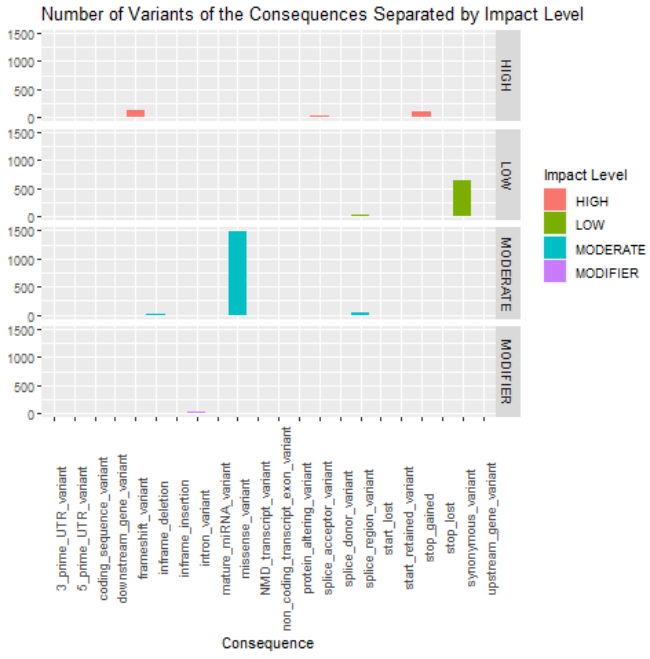

Figure 3

2.2. Data Preparation

The original consequences of each mutation variant type are stored in one column of the dataset We filtered the samples and genes based their appears in mutation variant table. Then, we used them to subset the original scaled gene expression table to obtain a new gene expression table with the condition of original gene express level ¿ 10 for more than 80 out of 87 original samples. We performed scaling based on each consequence and each gene to mean 0 and standard variation 1.

2.3. Clustering

We conducted hierarchical clustering on the consequences of each mutation variant based on the number of samples having the variant and the number of genes involved in the variant. The distance metric is Euclidean distance and the linkage method used in hierarchical clustering is average linkage. We also applied principle component analysis to reduce the dimensions of consequence and apply k means clustering algorithm on the matrix obtained by PCA with number of components that explain at least 90 percent of the variance. Following a similar procedure, we also analyzed the filter gene expression data in the same way.

3. Result

3.1. Hierarchical Clustering on Consequences of Mutation Variants

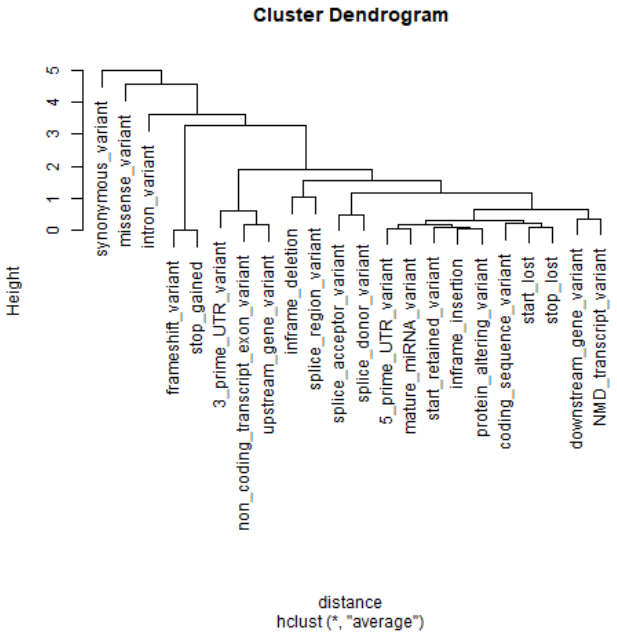

Figure 4

In Figure 4, we draw the dendogram based on the result of the hierarchical clustering. The figure validates our assumption in Section 2.1, where if we cut the dendogram with height 2.5, three dominant consequences were identified and formed three singleton cluster and the rest of consequences clustered in the remaining group. The three singleton groups consist of synonymous variant group, missense variant group, and the intron variant group, which corresponds to the dominant variants in each of the low, moderate, modifier impact level.

3.2. PCA and KMeans Clustering on Gene Expression Levels

The second major analysis we performed is to process the filtered gene expression level data table. We performed PCA to reduce the dimension across genes and find that 6 principle component would be able to explain 90% of the variation across gene expression data. The Kmeans clustering algorithm were then performed on the first 6 columns of PCA processed data. From Figure 5, we can see that the KMeans clusters mostly separate the samples along axis of the second principle component. We then hypothesized if values of PC2 indicate certain features of samples.

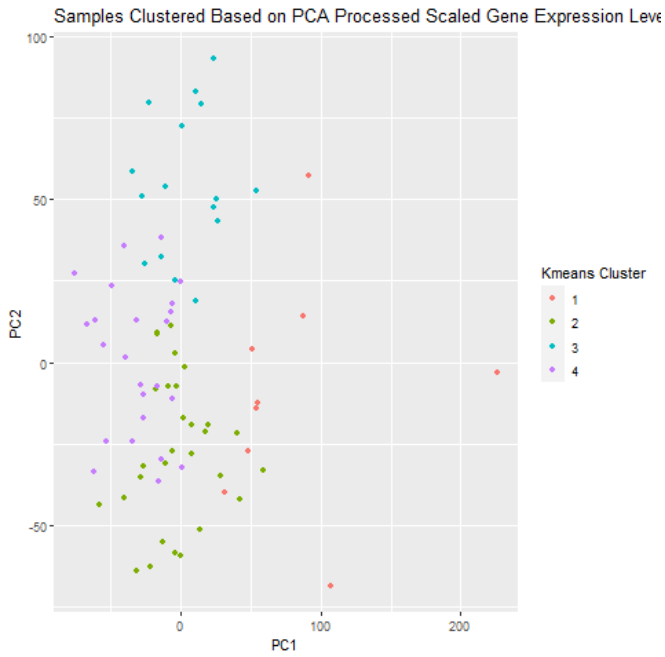

Figure 5

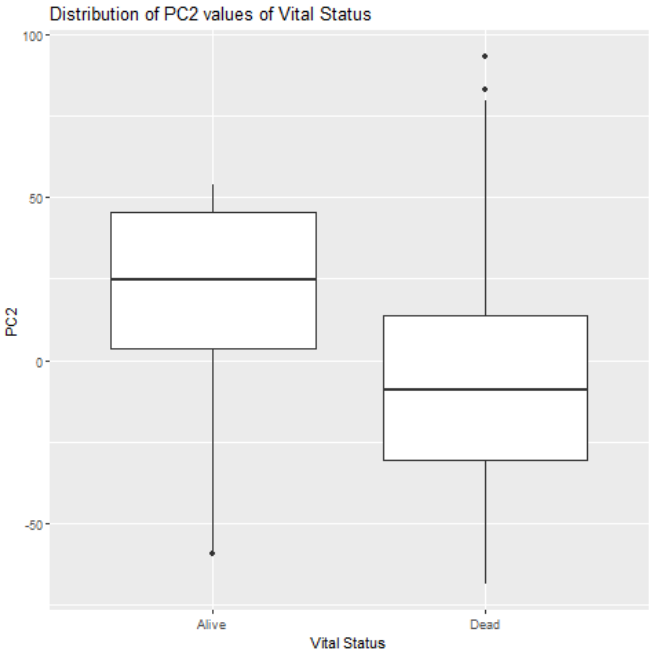

Figure 6

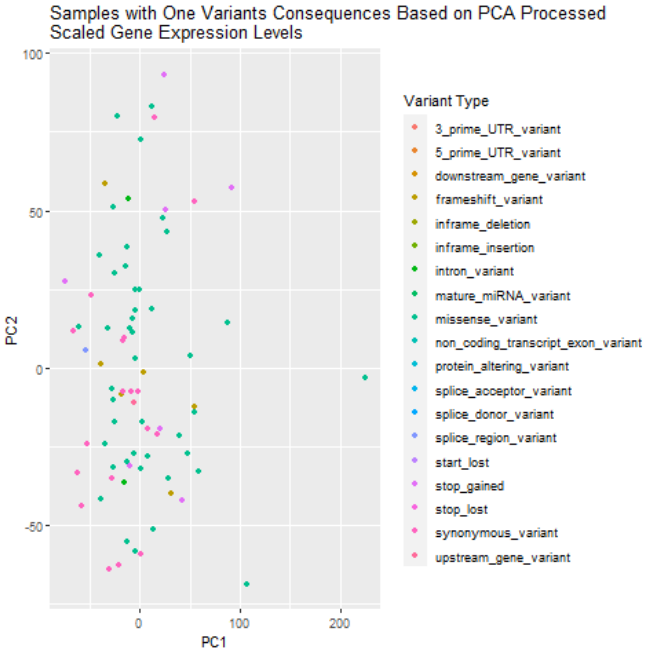

Figure 7

One significant results is how the distribution along the second principle component varied between different vital status. The distributions are shown in Figure 6. Analysis of the PCA processed gene expression data were performed with different attributed of samples and mutations were performed. However, there were no noticeable results produced and therefore results in a lack of associations between the mutation variants and gene expressions from a high throughput analytical perspective. From Figure 7, implicit groups of consequences are present along the second dimension of the sample, but the distinguish-ability is low.

4. Discussion

Additional analysis were performed but only a part of the results were discussed in detail within this writeup. Please reference the RMarkdown file for the complete methodology and analysis. All figures produced are either used in the writeup or included in the Supplementary Figures section 5. All figures are also included in the GitHub repository images folder.

In summary, through this study, we managed to find the clustering of the consequences are highly correlated to impact levels of the variants.But overall, high throughput gene expression analysis has not fully proven its robustness in predicting pathological features of mesothelioma and we have shown that no particular association between mutation variants and high throughput gene expression data were discovered.

Certain bias might exist in the data set as the sample size of the project is limited and might not be able to reflect the real distribution of certain sample features such as the vital status of the sample. Therefore, future work on other dataset is required to verify the associations noticed in this study. Also, the relationship between samples and variants, variants and consequences can better be fitted with mixed-membership models, which requires more complicated data processing, statistical model and analytical procedures to fully understand their relationships.

References

[1] F. E. Mott. Mesothelioma: a review. *Ochsner J*, 12(1):70–79, 2012. 1

[2] M. J. Teta, P. J. Mink, E. Lau, B. K. Sceurman, and E. D. Foster. US mesothelioma patterns 1973-2002: indicators of change and insights into background rates. *Eur J Cancer Prev*, 17(6):525–534, Nov 2008. 1

5. Supplementary Figures

5.1. Sample Features

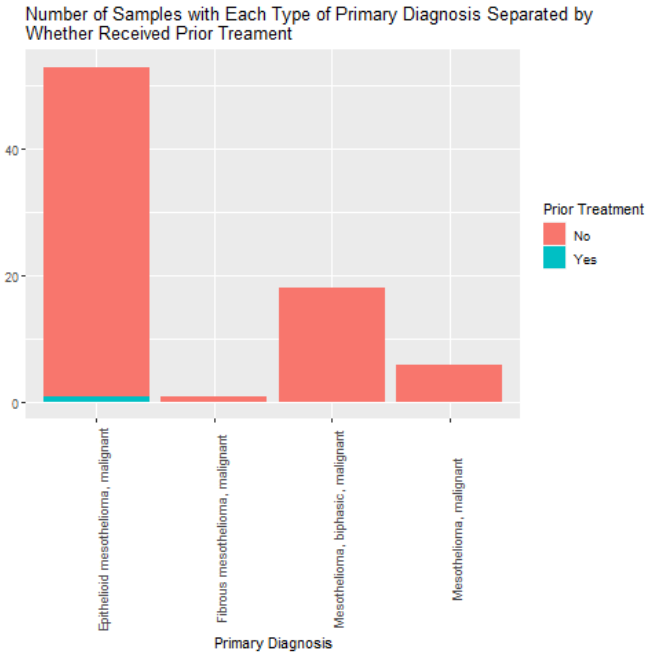

Figure 8

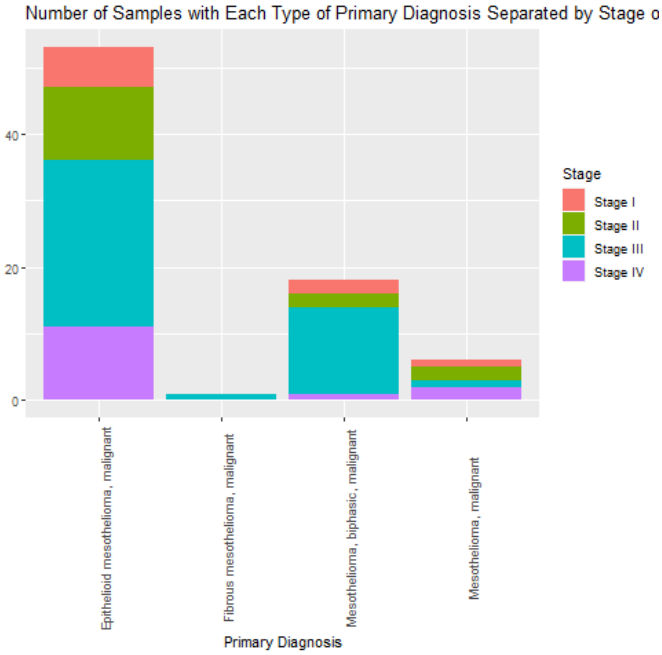

Figure 9

5.2. Variant Features

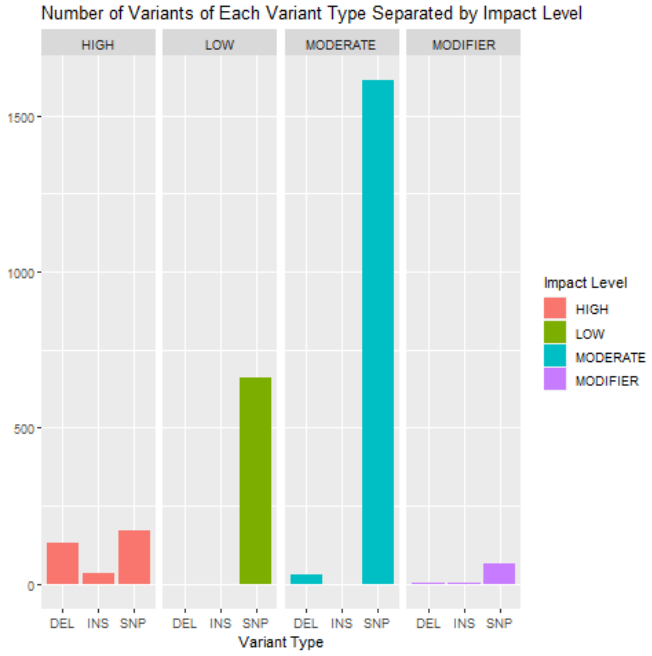

Figure 10

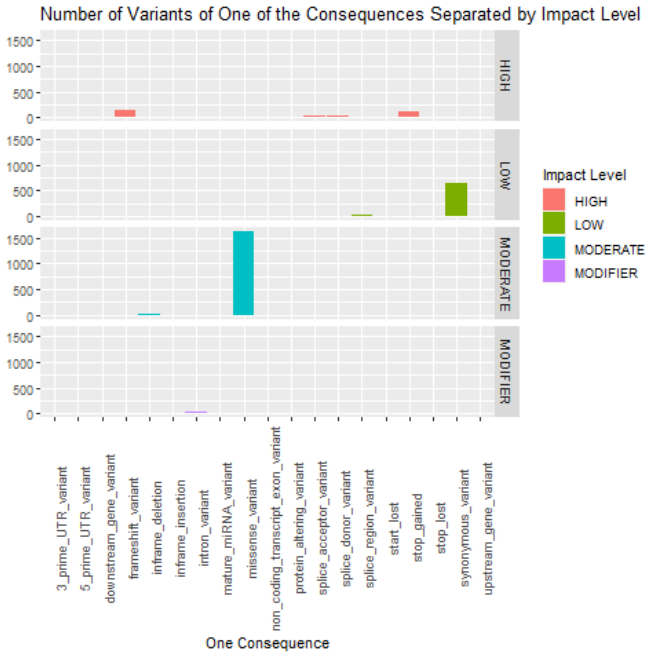

Figure 12

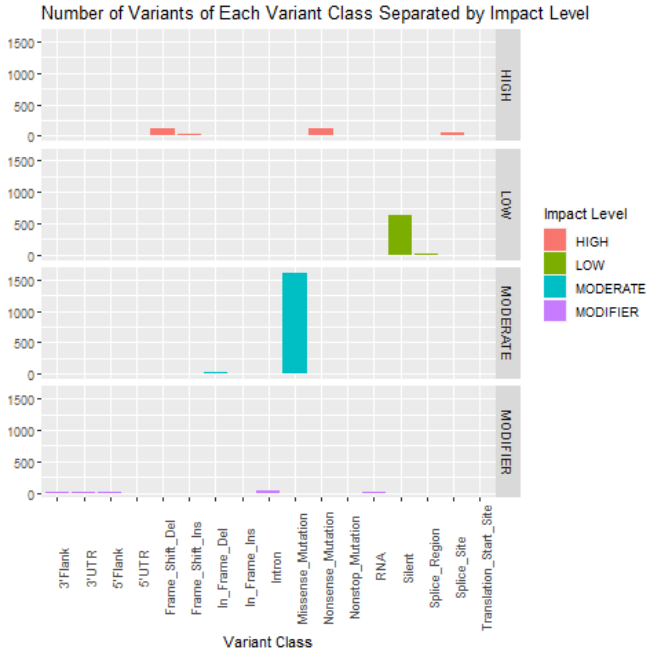

Figure 11

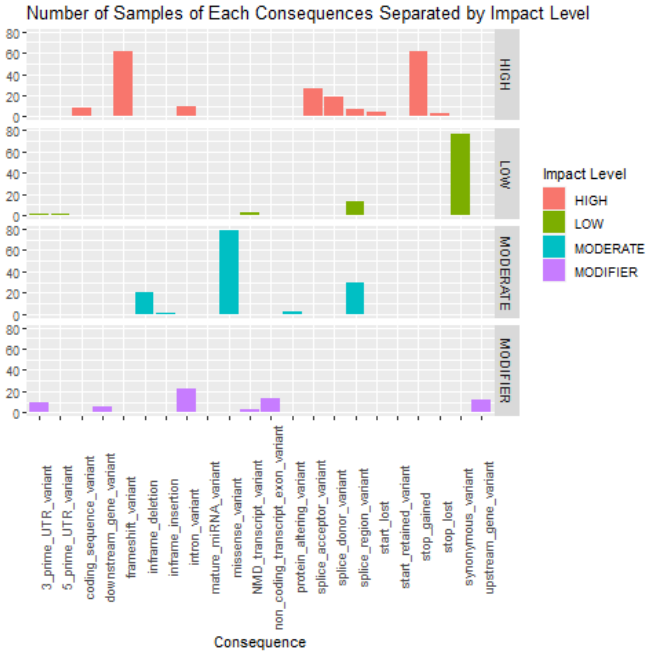

Figure 13

5.3. Consequence PCA and KMeans

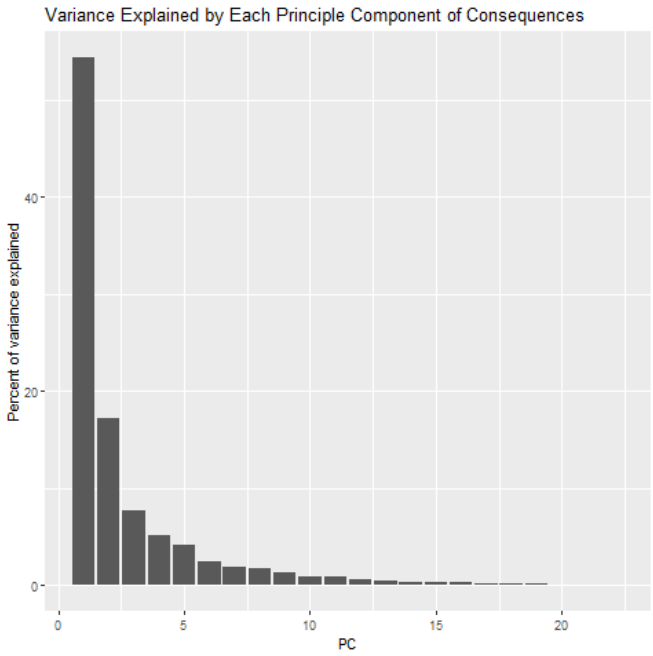

Figure 14

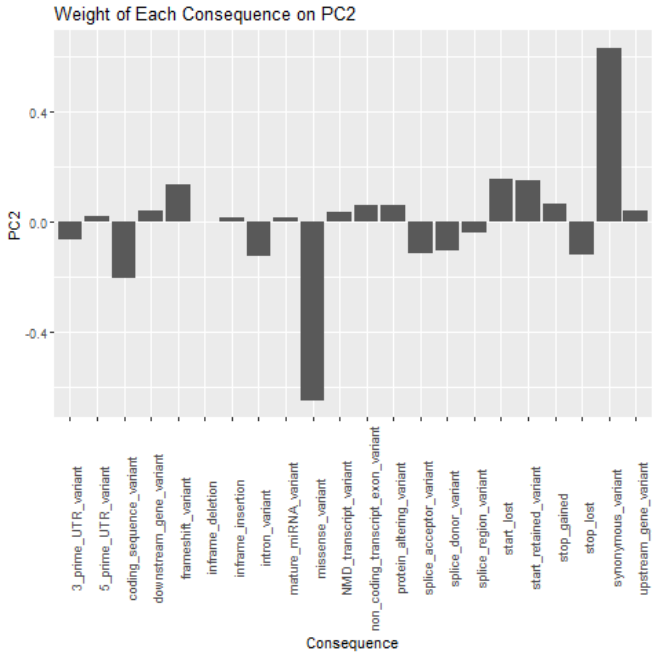

Figure 16

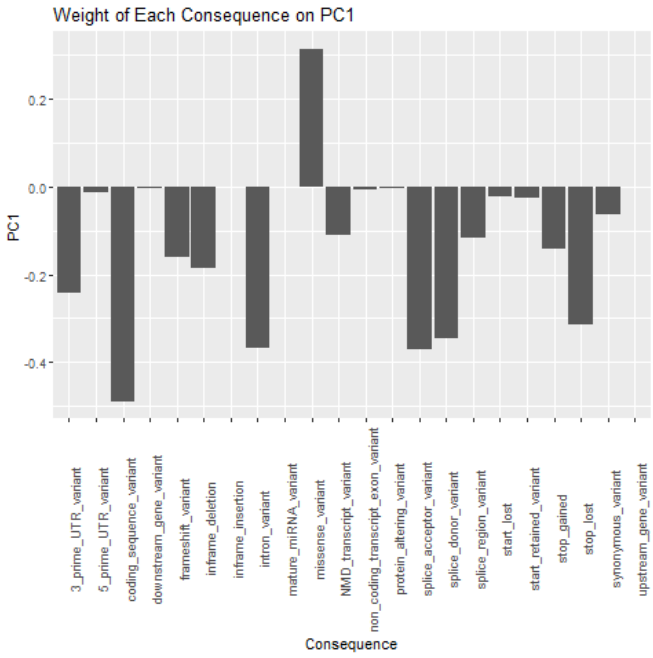

Figure 15

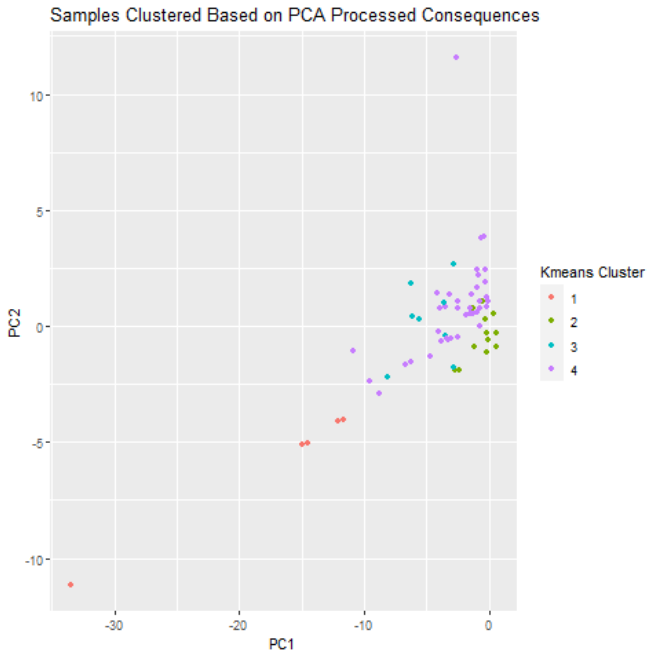

Figure 17

5.4. Gene Expression PCA and KMeans

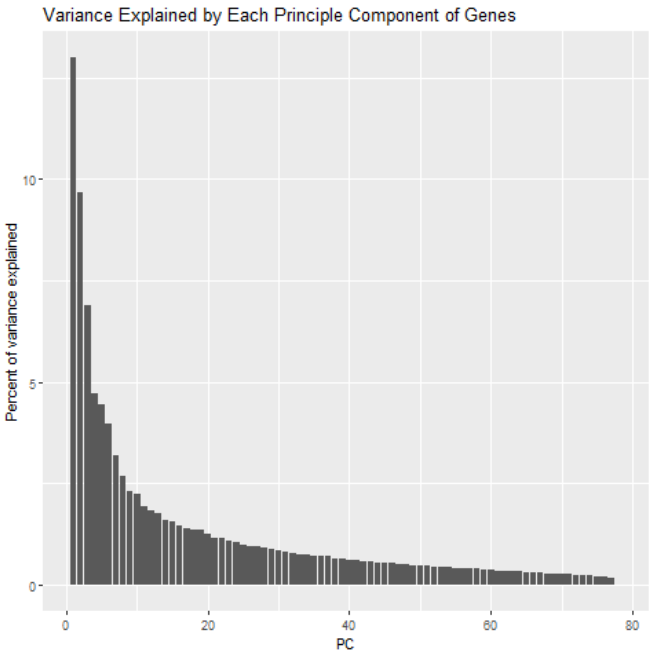

Figure 18

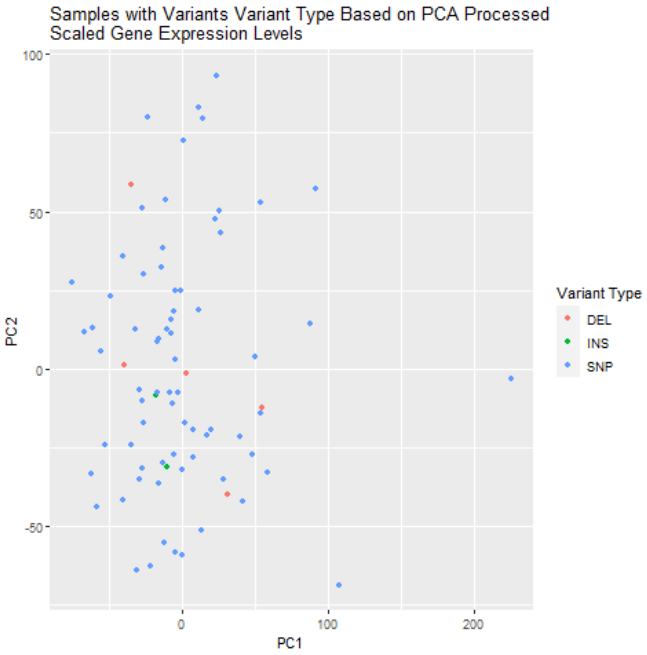

Figure 20

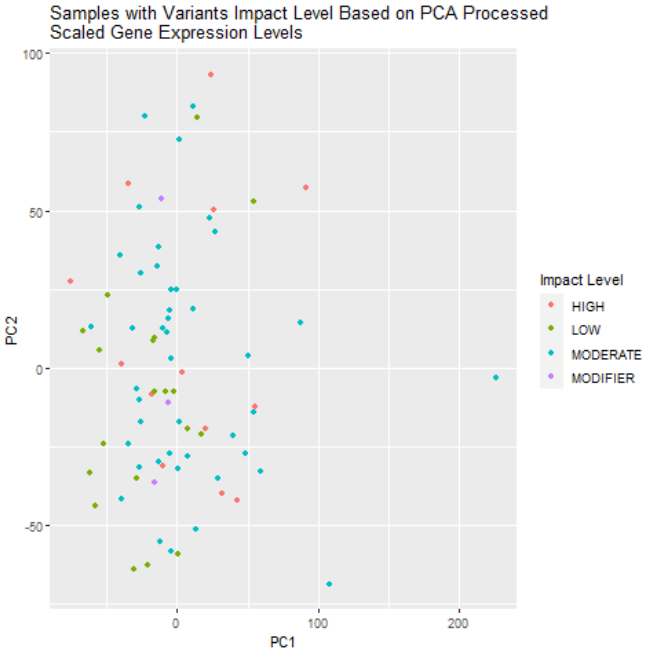

Figure 19