Analyzing Genetic Data to Determine Factors for Gene Expression and Liver Hepatocellular Carcinoma Patient Survival

BMEG 310 - Group 16 Final Report

Ryan Choi, Tiffany Huang, Adi Schlager
9 December, 2022

Abstract

Liver Hepatocellular Carcinoma (LIHC) is one of the most common types of liver cancer and is currently the third leading cause of cancer death worldwide, with over 700,000 deaths every year [1]. Many studies have been conducted on the pathogenesis of LIHC [2], but distinct factors and genes contributing to the promotion of LIHC remain unclear. Thus, this report aims to explore whether there are genetic mutations specific to each stage of cancer. Our research involved the bioinformatic analysis on the RNA sequence, mutations and clinical patient datasets of LIHC obtained from The Cancer Genome Atlas (TCGA) in R. The study followed an analysis workflow that included initial data exploration, DE analysis on expression and mutation data using the "DESeq2" pipeline, survival analysis using the "CRAN "package and pathway analysis using the "GAGE" package. All three datasets were used by conducting DE analysis based on specific gene mutations and clinical patient factors. This study found no conclusive correlation between any investigated factors and gene expression rates. It did not identify any significant gene mutations either.

Introduction

Liver Hepatocellular Carcinoma (LIHC) is the most common liver cancer and causes over 700,000 deaths annually [1]. While liver cancer may be caused by the metastasis of tumors in other areas of the body, HCC originates in the liver. It severely affects the liver's functionality. As the number and size of tumors grow, the liver gets less effective at removing toxins from the blood. It also cannot provide the necessary hemoglobin, amino acids, or platelets to the bloodstream. Additionally, the metabolic functions of storage and bile production get reduced as the amount of healthy tissue decreases [3].

LIHC occurs most in people with damaged livers or with liver diseases. Specifically cirrhosis, hepatitis B and C, Type-2 diabetes, and extreme accumulation of fat in the liver have been proven to severely increase risk [4]. Smoking and consuming alcohol introduce toxins to the liver and thus can increase risk. LIHC tends to progress slowly with few symptoms in early stages but speeds up as it grows [5].

Detecting LIHC as early as possible is very important to ensure the most appropriate treatment is provided. The current issue with detecting LIHC is that symptoms do not appear immediately and it is very easy for medical imaging to miss early tumors. The best method, MRIs, still missed up to 25% of early stage tumors. CT scans and ultrasound were found to be worse at definitively identifying early stage tumors [6].

LIHC treatments include surgery or liver transplants to remove or replace the diseased tissue. Alternatives include various targeted plans such as ablation, immunotherapy, and radiation therapy [5]. Each case requires its own treatment plan based on several factors such as eligibility for transplant, tumor size and number. Treatment type also affects future allocation of resources to ensure a favorable result [6].

The goal of this project is to determine whether there are genetic mutations that are specific to certain stages of cancer. This would theoretically allow for preemptive targeting of late stage mutations during early stage liver cancer. The report discussed how the RNA sequence, clinical patient, and mutation data was explored. It will also present the findings and how they may relate to biological or clinical applications.

Methods

First, the “RNA sequence,” “Clinical Patient,” and “Mutation” datasets for LIHC were downloaded from the Bioconductor package TCGAbiolinks.

To clean the dataset, all “RNAseq” samples that did not have an associated patient in “clinical_patient” were removed. All genes that had a row sum less than one were removed as well because they were insignificant compared to other genes and could weaken our analysis.

The primary approach involved clustering the samples based on their respective cancer stage, and attempting to find the main differentiating genes using DDSeq. This method yielded insignificant results. The secondary approach involved generating clusters based on gene expression, and determining whether the clusters were more specific to certain cancer stages.

To minimize noise from less significant genes, the top 500 genes with the most variance were sorted based on their standard deviation. To speed up clustering, Principal Component Analysis (PCA) was applied to the new “RNAseq” dataset. The first 104 Principal Components, which accounts for 90% of the variability, were selected.

Hierarchical clustering was used for the versatility in number of clusters. Initially, the standard hierarchical method was used, however, using “ward.D” produced much more even clusters. Eventually, four clusters were selected as the difference between including five and six clusters was insignificant.

To evaluate the significance of our clusters, Differential Analysis was implemented using the “DDSeq” package to determine if any main genes had extremely variable expression levels between clusters. To do this, genes with an adjusted p-value of 0.05 were considered, as a 5% false positive rate was deemed acceptable. To choose the most variable genes, only significantly up- or down-regulated genes were considered ($log2FoldChange > |1|$). This meant that the gene expression in the comparison groups was either double or half of the control groups.

A heatmap and pca plot was generated to visualize the significance of the clusters (Figure 3a and 3b). From these, cluster 3 was determined to be the only significant cluster of the four. It was also noted that the large majority of genes showed minimal difference between the four clusters. As the other clusters were not well defined, clusters 1, 2, and 4 were combined into one cluster which could be compared to cluster 3.

The “GAGE” package was used to look at affected pathways for the differential analysis.The “GAGE” package failed when limiting the genes used to those with less than 5% false positive rate so it was increased to 6%.

Performing survival analysis on all four initial clusters showed that patients in cluster 3 had a slightly lower survival rate than patients in all other clusters (Figure 4). Given that the false positive rate was quite small ($p = 0.00016$), these results indicated that the difference in gene expression has significant consequences.

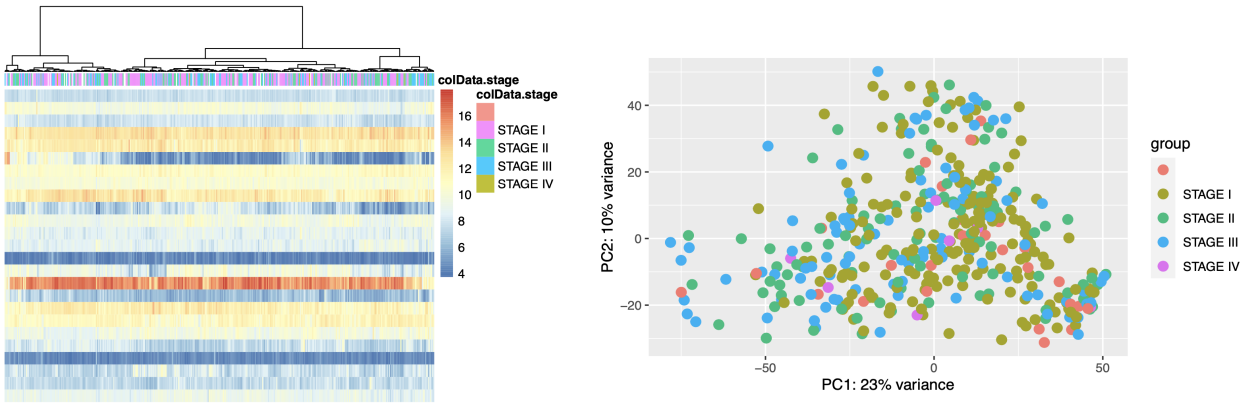
The frequencies of patient sex, age, race, and cancer stage in cluster 3 (Figure 5) was then examined. As there was no significant disproportion, there was no need to do a confusion matrix on each one.

Because cluster 3 was not significantly composed of a certain type of cancer patient, an examination was conducted to determine whether there was a common mutation that differentiated cluster 3 from the others. A confusion matrix was created for each of the top five most common mutations found in individuals belonging to cluster 3. The mutation with the most accurate confusion matrix was tested with survival analysis to see if it impacted patient survival.

To look for similarities between the patient clusters and mutation data, DE analysis was performed on the RNA sequence data between patients with top 50 mutated genes and those without. Pathway analysis and survival analysis was conducted on the DE genes using the “GAGE” package and “CRAN” package, respectfully. Finally, the PCA plots, pathways and survival analysis were compared between the expression and mutation analysis in an attempt to find key genes or pathways that may be significant.

Results

Our primary results show that there were no significant differences in gene expression between tumor stages as seen in the heatmap and PCA plot in Figure 1a and 1b. There was an additional attempt to find a significant difference in gene expression between the sexes, though none was found. (Appendix A).



(a) (b)

Figure 1. *Patient-gene heatmap (a) and PCA plot (b) on differentially expressed (DE) genes between tumor groups. The rna sequence was filtered with Padj < 0.05 to get significant genes and log2FoldChange >1 and <-1 to get differentially expressed genes.*

Thus, we attempted to approach our analysis from a different angle by finding patient clusters based on similar gene expression. Hierarchical clustering on the gene expression levels using the "Ward" methodology successfully produced roughly even clusters as seen in Figure 2. This implies that there exist patient groupings based on gene expression.

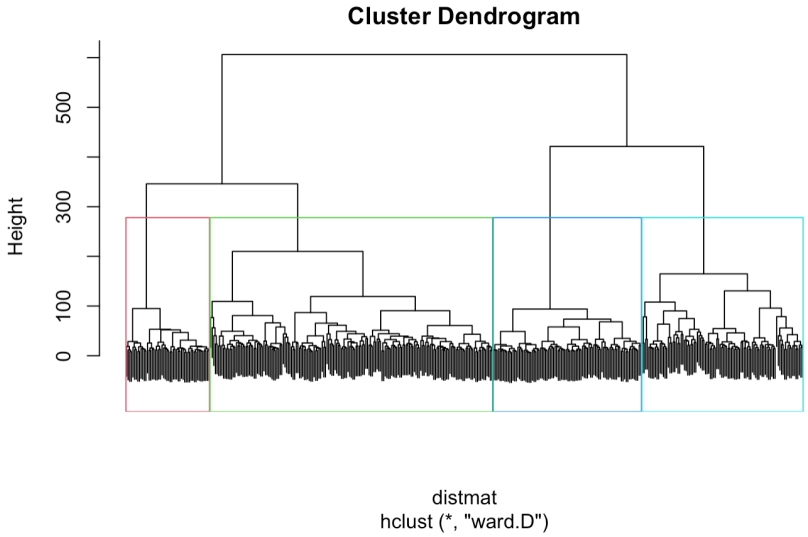
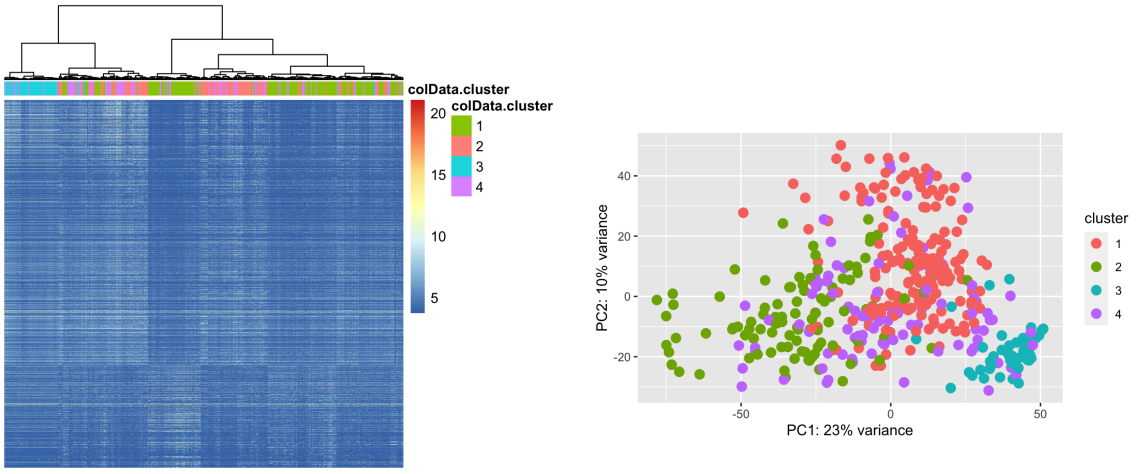


Figure 2. *Hierarchical clustering on gene expression levels with cluster = 4*

Differential analysis showed that out of the four clusters defined by hierarchical clustering, only one of them was significant as seen by the clear separation of the blue cluster in Figure 3a and 3b. The remaining three clusters were not as well defined on the first two principal components, so the difference between them is not nearly as significant (Figure 3b). This means that the gene expression in cluster 3 is significantly different from the rest of the clusters.



(a) (b)

Figure 3. *Patient-gene heatmap (a) and PCA plot (b) on differentially expressed (DE) genes between patient clusters. The rna sequence was filtered with Padj < 0.05 to get significant genes and log2FoldChange >1 and <-1 to get differentially expressed genes.*

From the DDseq data, we were able to identify the most down-regulated pathway to be Steroid Hormone Biosynthesis (Appendix C.1) and the most upregulated pathway to be Intestinal Immune Network for IGA Production (Appendix C.2). This gives us a general image of what pathways determined the formation of these clusters.

From our survival analysis on the patient clusters (Figure 4) , we can see that patients in cluster 3 have a significantly lower survival rate compared to patients in the other clusters, particularly compared to cluster 1 and 4.

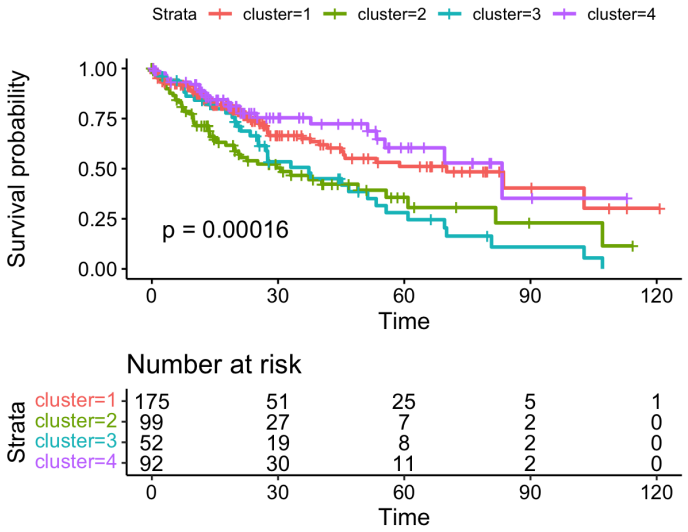


Figure 4. *Survival analysis on patient clusters. This survival analysis has a p-value < 0.05 meaning that the null hypothesis was rejected and that there are significant differences in survival rate between the groups.*

We wanted to see if any of the clinical factors affected cluster 3. Before attempting confusion matrices, we quickly looked at the patient demographics for cluster 3. As far as we could tell, there was no distinct cancer stage

group that belonged to that cluster. Since none of the factors seemed promising (figure 5), we moved on to check if there was a specific mutation associated with that cluster.

	Cluster 3	Overall
Type 1	18	173
Type 2	12	87
Type 3	12	85
Type 4	1	6

Figure 5. *Comparison of Frequencies of Cancer stages in Cluster 3 and Patient Data. Overall, the ratios look very similar, indicating that no stage is more prone to be in cluster 3.*

Looking at the top 5 most expressed genes in cluster 3 gave us the following: TTN, MUC16, TP53, CTNNB1, and ANKHD1. Creating confusion matrices based on the presence of these mutations within and without cluster 3 showed some results (Appendix B). Overall, the accuracy was mediocre, and the recall and precision was low, indicating that these genes by themselves were not good differentiating factors of cluster 3. The best accuracy measure was ANKHD1: the confusion matrix yielded an accuracy of ~85%.

Our Survival Analysis found that there was no significant correlation with regards to survival between patients with mutation in the ANKHD1 gene and those without as seen in Figure 6.

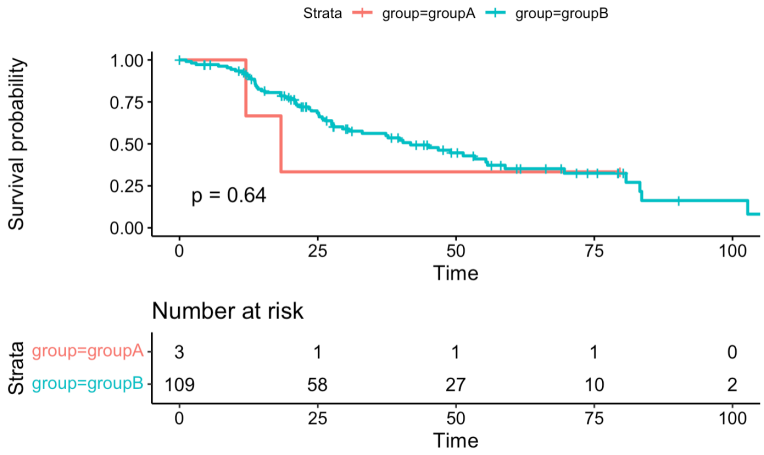
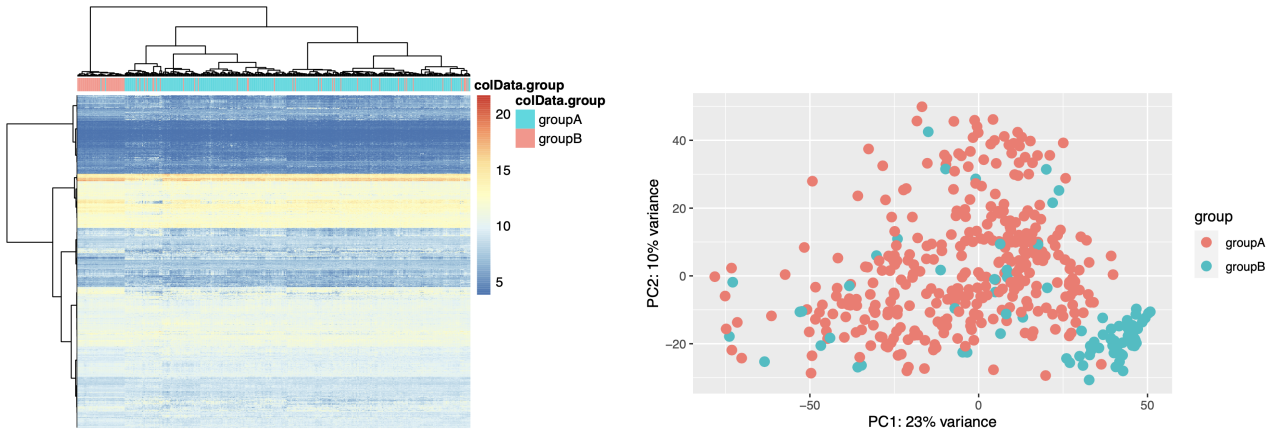


Figure 6. *Survival analysis on mutation groups where groupA included patients with mutations in the ANKHD1 gene and Group B included those without. This survival analysis has a pvalue > 0.05 meaning that the null hypothesis was not rejected and therefore we cannot draw conclusions regarding the correlation of survival with mutations in ANKHD1*

As this doesn't match with the survival analysis for cluster 3, this indicates that ANKHD1 is not at all an indicator of whether an individual belongs to this cluster. This means that there likely isn't any one gene that differentiates cluster 3 individuals.

Surprisingly, we noticed that based on visualization alone, the size, shape, and location of cluster 3 seemed to match with Group B from a previously-done mutation analysis. (Figure 7) Individuals of Group B were diagnosed with liver cancer, but did not have any of the top 50 mutations that were expected to occur.



(a) (b)

Figure 7. *Patient-gene heatmap (a) and PCA plot (b) on differentially expressed (DE) genes between groups with vs without specific mutated genes. The rna sequence was filtered with Padj < 0.05 to get significant genes and log2FoldChange >1 and <-1 to get differentially expressed genes. groupA included patients with mutations in one of the top 50 mutated genes and groupB included those without.*

This led us to conduct a survival analysis between the mutation data groups to see if the survival rates of Group B were similar to that of cluster 3 (Figure 8)

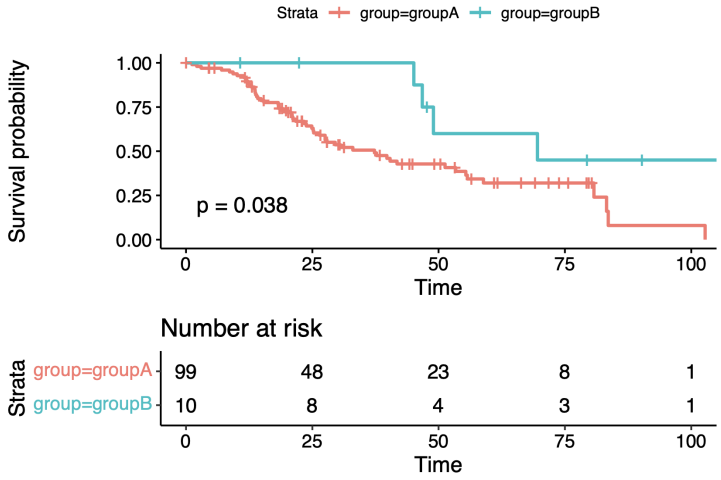


Figure 8. *Survival analysis on mutation groups where groupA included patients with mutations in one of the top 50 mutated genes and Group B included those without. This survival analysis has a p-value < 0.05 meaning that the null hypothesis was rejected and that there are significant differences in survival rate between the groups.*

From our results, it seems that cluster 3 and Group B may be in some way related, as their clustering is very similar. However, this does not mean that cluster 3 is group B as shown by the discrepancy in their survival analysis results.

Discussion

From our analysis, we found no significant differences in gene expression based on clinical patient features such as sex and tumor stage. However, studies have shown that males are more likely to get liver cancer compared to females [7]. This suggests that sex and other clinical patient features may affect the risk of cancer but it may be due to external behavioral or environmental factors rather than due to gene expression differences. Furthermore, studies have shown that Liver Hepatocellular Carcinoma is most commonly caused by damage or scarring of the liver from external behaviors such as excessive alcohol consumption which has been noted to be more prevalent in males [8]. Thus, this may explain the lack of differences found between gene expression across different patient groups in our research.

What is more surprising, was that the cancer stages appear to not be well defined by their mutations. We expected that as the cancer progressed, it would mutate more frequently, as it grows exponentially. In theory, this would mean that stage 4 cancers have mutations that stage 1 cancers do not. This however was not the case, and there did not appear to be many differences in mutations between the various cancer stages. Upon further thought, this was likely due to cancer mutations being random. So there might not be a single mutation that defines a stage 4 cancer, and even if there was, there would be many many more that are different between two samples. This would make it incredibly difficult to find.

When verifying our formed clusters with DDSeq, cluster 3 was the most defined cluster, however, we can see from the heatmap that it wasn't vastly different from the other clusters. This indicates that there might be only one or two gene expressions that separate it from the rest. It is also apparent that even those gene expressions do not differentiate it strongly, as the first and second principal components in figure 3b are not very high. (23% and 10% variance respectively). This indicates that despite the relatively clear clustering, it might be difficult to accurately pinpoint the particular genes that caused the separation, as the difference might be very slight.

Another hint that the difference between clusters is relatively small (and by extension will be difficult to find) is the survival analysis results. Although they are significant with a p value of 0.00016, the difference between cluster 3 and cluster 2 is not very large, and since we were considering clusters 1, 2, and 4 to be the same by extension the difference between the large cluster and cluster 3 will not be very big. It is also interesting to note that despite clusters 1,2, and 4 having relatively poor separation, there is a rather stark difference in survival rates between clusters 2 and the others.

Given how the DDseq clustering separates clusters 1 and 2 rather evenly, and difference in survival rates between clusters 1 and 2, a repeat analysis utilizing this methodology should consider using 6 clusters. This might result in more prominent and even-sized clusters, making it easier to find significant genes.

When comparing the patient clusters from expression analysis to the mutation groups from mutation analysis, we observed some overlap between cluster 3 and the group without the top 50 mutated genes. However, we were unable to draw any clear conclusions with regards to this due to discrepancy in their survival analysis results.

Lastly, while the pathways did not give us data about the top mutated genes, the processes mentioned in the pathways are clearly related to the function of the liver. Furthermore, it is interesting to note that one of the most down-regulated pathways for the mutation cluster based on gene expression is in complement and coagulation cascades and one of the most upregulated pathways is in cell cycle. Both of these pathways have been shown to be significantly enriched in literature[9]. Biologically, this makes sense as cancer is essentially unchecked cell growth and those upregulation in the cell cycle would accelerate division rates. It has been shown that people with cancer are more at risk for blood clots[10] and thus it makes sense that mutations affecting downregulating the coagulation cascade may cause that.

Challenges/Limitations

There were several limiting factors in the clinical patient dataset that restricted the options for analysis. Although the dataset had many patients, there was not much diversity in demographic. Male patients outnumbered female patients 2:1, and the majority of patients were either of white or asian ethnicity. Additionally, patient height was not included in the dataset, making patient weight far less reliable to use as a factor as BMI could not be computed. The dataset was also predominantly made of individuals in stage 1 or 2 of liver cancer, with far fewer being in stage 3 or 4. Since the dataset was composed of many other independent datasets, there were many blanks in the patient data, which provided further restrictions. This provided a challenge in initially deciding what factor would be the best to target. We eventually chose to determine genetic differences between the various stages of liver cancer, as we reasoned that the later stages would be strongly differentiated from earlier ones by several mutations.

Our initial workflow plans were cut short when we found that using the cancer stage as a condition for "DDseq" did not yield clean clusters as we had hoped. This was somewhat expected though, as the problem has historically not been easy to solve. We decided on a workaround secondary approach that involved clustering the samples based on genetic expression, verifying the clusters in "DDseq," and testing if certain types of patients were more prone to be in certain clusters.

However, even after taking the top 500 variable genes, our genetic expression clustering proved to be too noisy. We overcame this by using PCA to reduce the noise, as well as the computation time. The work plan was changed again when "DDseq" could only verify one of our clusters. Analysis was continued by comparing the one cluster to the rest of the dataset in order to observe what had differentiated it so strongly compared to the rest, and we reconfigured our initial goal.

Another issue encountered was that the GAGE package simply refused to work only with the dataset of genes where p adjusted < 0.05. For some unknown reason it kept setting all the gene expression values to NA and returning the default list of pathways, featuring no up or down regulated genes. We had no choice but to change the accepted False Positive rate to 6% in order to get it working. Overall, this was not a huge difference, but was very confusing.

A limitation that we could not bypass was that the strongly differentiated cluster was very small. Because we were comparing it to the rest of the dataset, it was difficult to determine if the most frequent mutations in that cluster were actually significant. Most of them had mediocre accuracy except for ANKHD1, which had a mediocre 85%. However, the recall and precision were overall abysmal, so the results were not likely to be significant. Just to be sure, we ran survival analysis between patients who had a ANKHD1 mutation and those that didn't.

Overall, the general limitations in current approaches regarding cancer bioinformatics were applicable to this analysis. For example, it has been shown that a variety of mutations are required for cancer to occur, as opposed to just one or two [11]. Thus our analysis was unable to pinpoint any one gene that causes differentiation, and could only verify a general pathway where said genes may be located.

Conclusion

In this study, we were unable to answer our original research question. Our data analysis showed no significant difference in gene expression between different stages of cancer, nor between any other patient metrics. Further research, perhaps with a larger and more flushed out dataset, or more sophisticated methods, is required in order to achieve a proper conclusion.

In terms of what this study actually yielded, we were able to find a cluster of patients that had a lower rate of survival when compared to the rest of the dataset. Due to the clusters small size, we were unable to properly determine if its individuals had a predominant trait or mutation. However, it is likely that the genes that determine members of this cluster are part of the Steroid Hormone Biosynthesis and the Intestinal Immune network for Iga production pathways.

Further studies are needed to validate the existence of this cluster, as well as the possible factors that define it.

Contributions (as agreed on by all members)

All members contributed equally to the report. Adi was the main contributor for the presentation. The main mutation analysis was completed by Tiffany. The main expression analysis was completed by Ryan. Adi helped with both as well as data processing. Tiffany worked on clusters A and B analysis while Ryan worked on cluster 4 analysis.

References

[1] "Key statistics about liver cancer," *ACS Journals*, 12-Jan-2022. [Online]. Available: https://www.cancer.org/cancer/liver-cancer/about/what-is-key-statistics.html. [Accessed: 09-Dec-2022].

[2] R. Dhanasekaran, S. Bandoh, and L. R. Roberts, "Molecular pathogenesis of hepatocellular carcinoma and impact of therapeutic advances," *F1000Research*, vol. 5, p. 879, 2016.

[3] "Liver: Anatomy and functions," *Liver: Anatomy and Functions*, 19-Nov-2019. [Online]. Available: https://www.hopkinsmedicine.org/health/conditions-and-diseases/liver-anatomy-and-functions#:~:text=Functions%20of%20the%20liver&text=All%20the%20blood%20leaving%20the,body%20or%20that%20are%20nontoxic. [Accessed: 09-Dec-2022].

[4] "Liver cancer," *Mayo Clinic*, 18-May-2021. [Online]. Available: https://www.mayoclinic.org/diseases-conditions/hepatocellular-carcinoma/cdc-20354552. [Accessed: 09-Dec-2022].

[5] "Hepatocellular carcinoma (HCC): Causes, symptoms, treatments & prognosis," *Hepatocellular Carcinoma (HCC)*. [Online]. Available: https://my.clevelandclinic.org/health/diseases/21709-hepatocellular-carcinoma-hcc. [Accessed: 09-Dec-2022].

[6] J. D. Yang, "Detect or not to detect very early stage hepatocellular carcinoma? the western perspective," *Clinical and Molecular Hepatology*, vol. 25, no. 4, pp. 335–343, 2019.

[7] "Liver Cancer Risk Factors," *ACS Journals*, 12-Jan-2022. [Online]. Available: https://www.cancer.org/cancer/liver-cancer/causes-risks-prevention/risk-factors.html. [Accessed: 09-Dec-2022].

[8] "Liver cancer," *NHS Inform*, 07-Dec-2022. [Online]. Available: https://www.nhsinform.scot/illnesses-and-conditions/cancer/cancer-types-in-adults/liver-cancer. [Accessed: 10-Dec-2022].

[9] M. Wu, Z. Liu, A. Zhang, and N. Li, "Identification of key genes and pathways in hepatocellular carcinoma," *Medicine*, vol. 98, no. 5, Feb. 2019.

[10] "Clotting problems," *Cancer.Net*, Aug-2019. [Online]. Available: https://www.cancer.net/coping-with-cancer/physical-emotional-and-social-effects-cancer/managing-physical-side-effects/clotting-problems#:~:text=People%20with%20cancer%20and%20those,called%20clotting%20or%20coagulation%20factors. [Accessed: 09-Dec-2022].

[11] N. Beerenwinkel, T. Antal, D. Dingli, A. Traulsen, K. W. Kinzler, V. E. Velculescu, B. Vogelstein, and M. A. Nowak, "Genetic progression and the waiting time to cancer," *PLoS Computational Biology*, vol. 3, no. 11, 2007.

Appendices

Appendix A. DE analysis based on Sex

Figure A.1. *PCA plot of differentially expressed (DE) genes between patient clusters.*

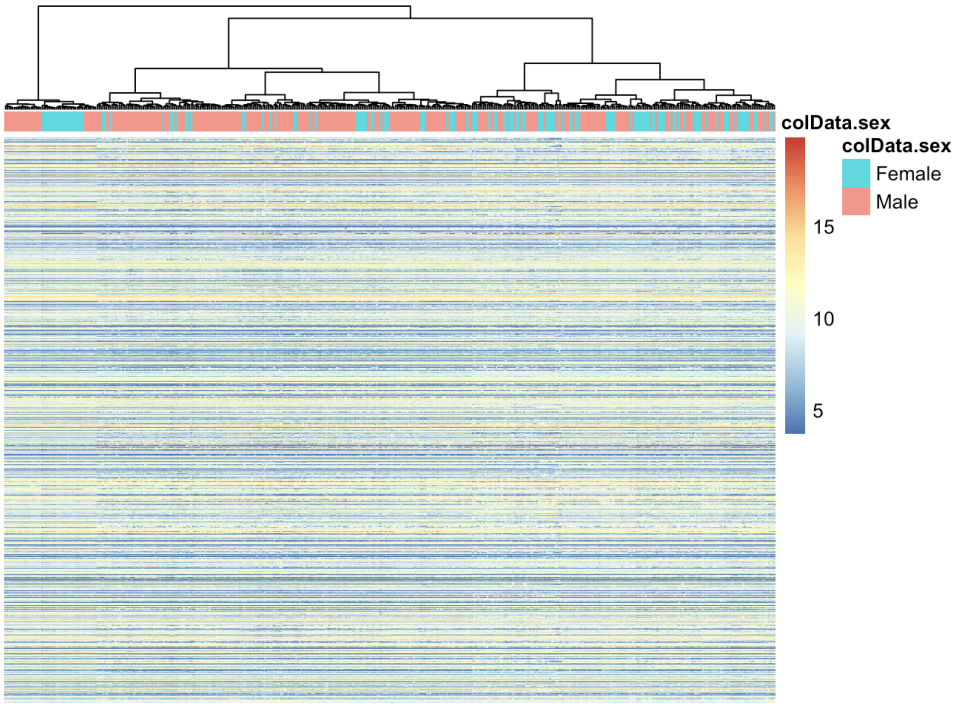


Figure A.2. *Patient-gene heatmap on differentially expressed (DE) genes between patient clusters. The rna sequence was filtered with Padj < 0.05 to get significant genes and log2FoldChange >1 and <-1 to get differentially expressed genes.*

Appendix B: Confusion Matrices for Top Five Most Common Mutated Genes In Cluster 3

"TTN"

prediction \ real	FALSE	TRUE	Accuracy =0.6062
FALSE	239	128	Recall = 0.1049
TRUE	37	15	Precision = 0.05906

"MUC16"

prediction \ real	FALSE	TRUE	Accuracy = 0.6993
FALSE	284	83	Recall = 0.09783
TRUE	43	9	Precision = 0.030717

"TP53"

prediction \ real	FALSE	TRUE	Accuracy = 0.6372
FALSE	256	111	Recall = 0.09016
TRUE	43	11	Precision = 0.0412

"CTNNB1"

prediction \ real	FALSE	TRUE	Accuracy = 0.6659
FALSE	270	97	Recall = 0.0849
TRUE	43	9	Precision = 0.03226

"ANKHD1"

prediction \ real	FALSE	TRUE	Accuracy = 0.8592
FALSE	355	12	Recall = 0.09615
TRUE	47	5	Precision = 0.01389

Appendix C: Most Significant Pathway for Patient Clusters based on Gene expression

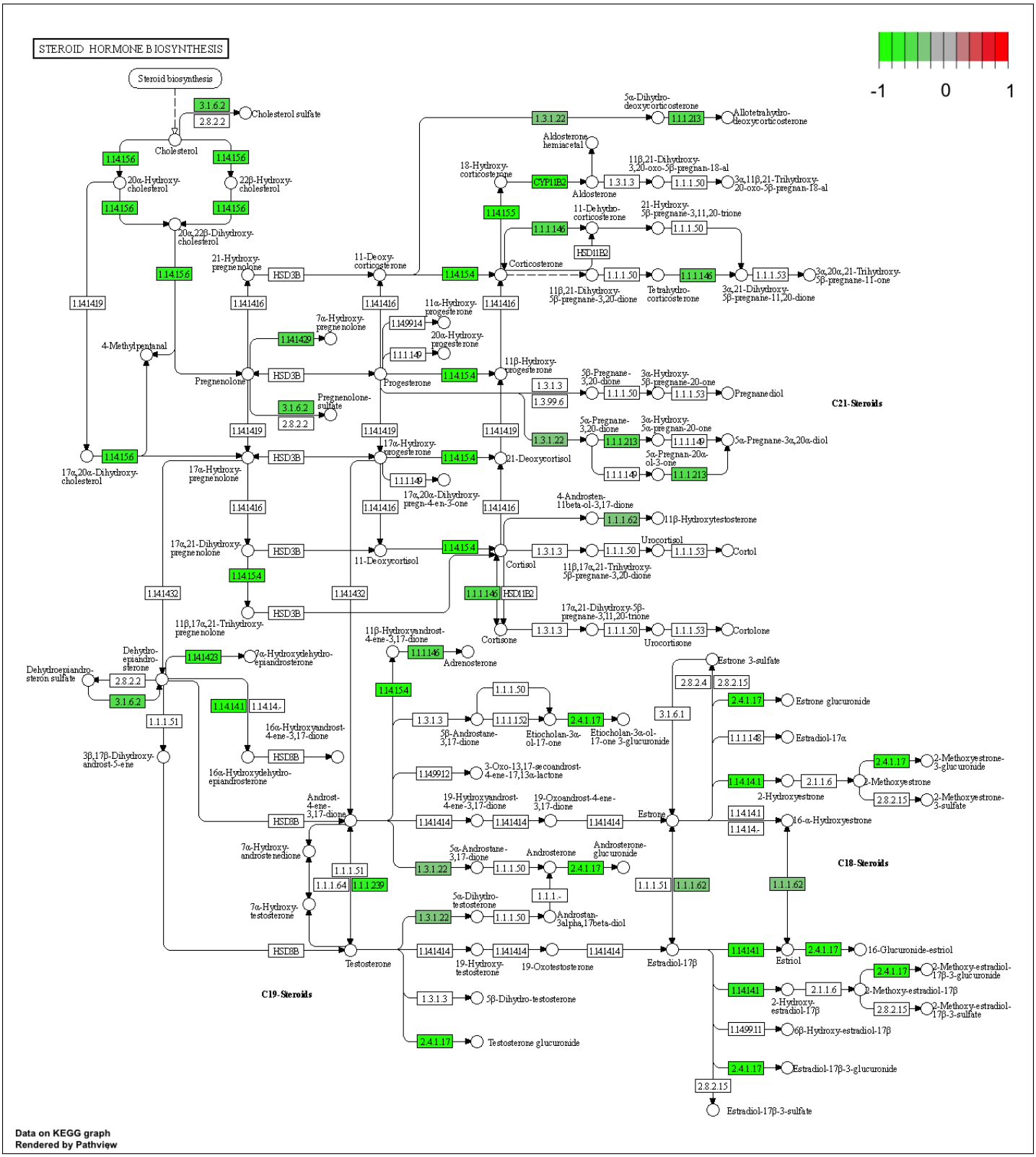


Figure C.1. *Most down-regulated pathway for mutation clusters*

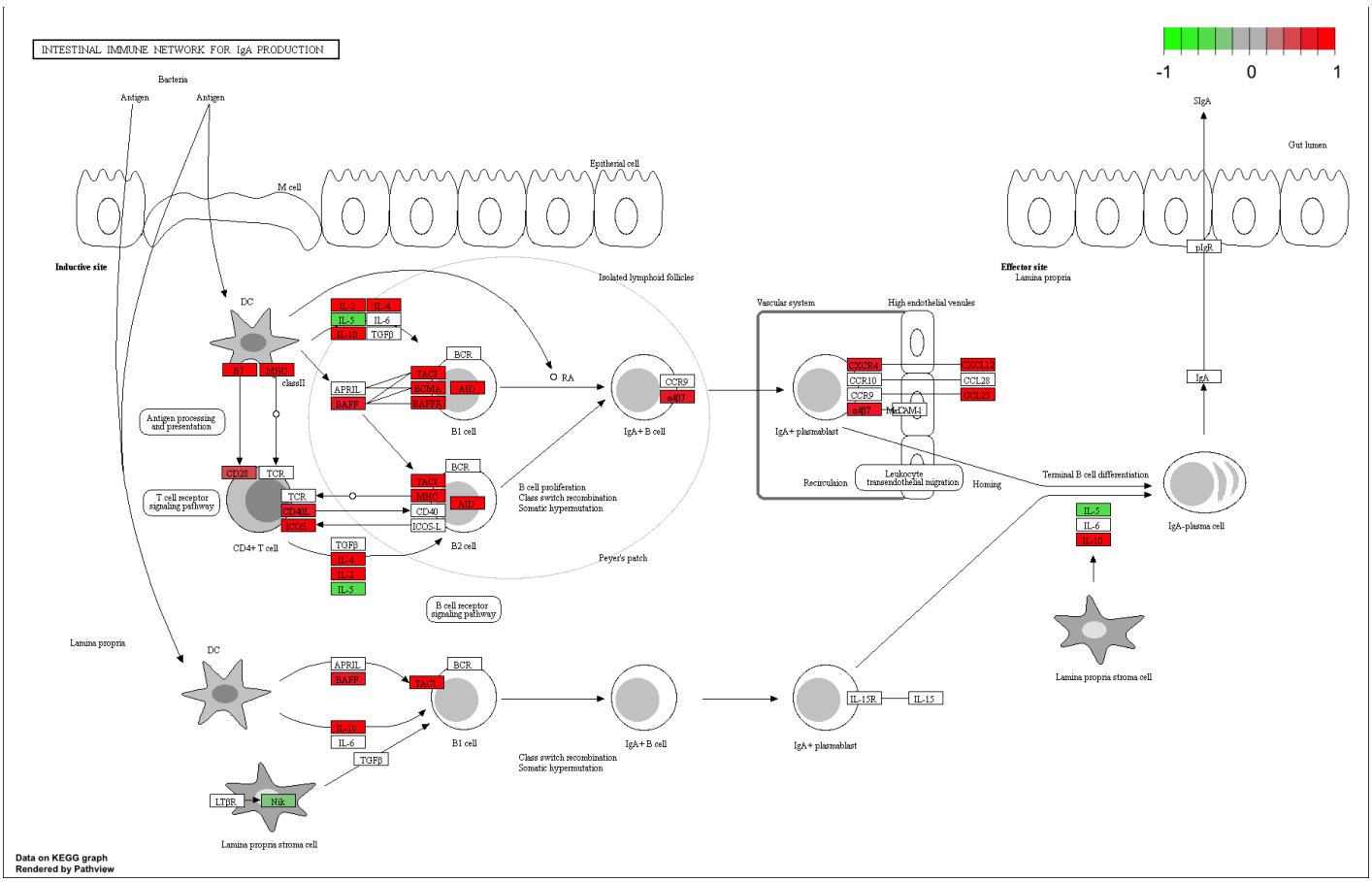


Figure C.2. *Most up-regulated pathway for patient clusters*

Appendix D: Most Significant Pathway for Mutation Clusters based on Gene expression

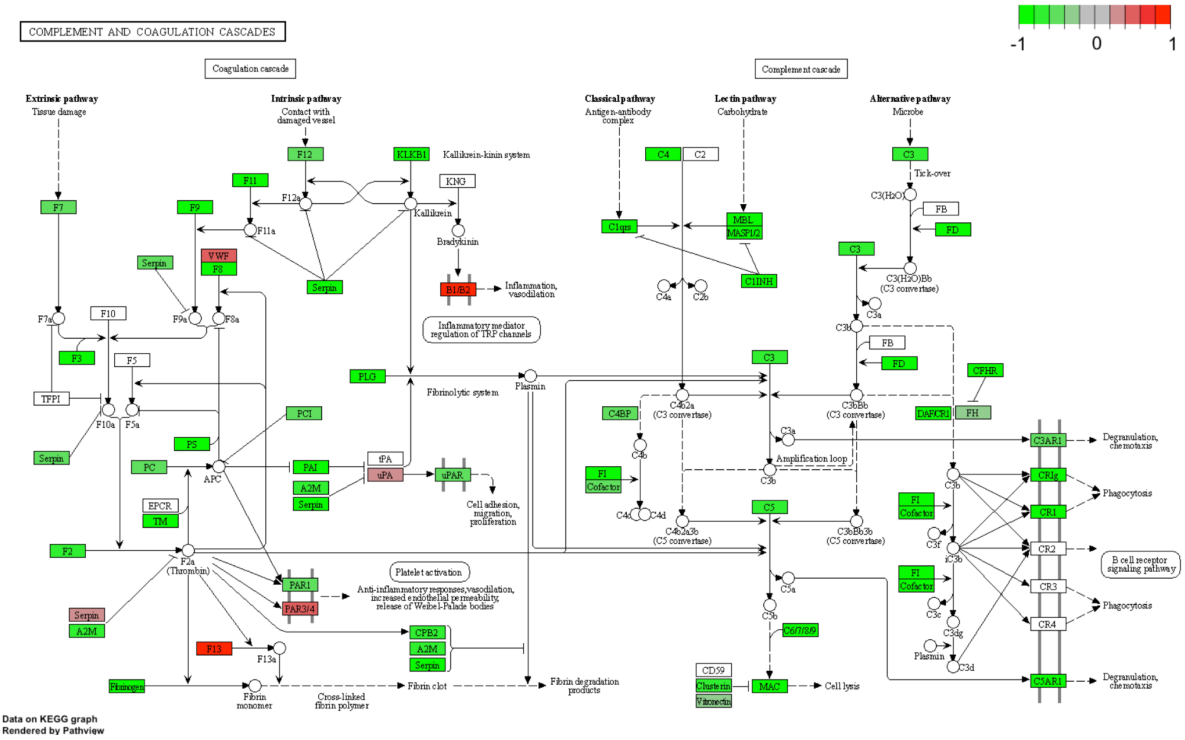


Figure D.1. *Most down-regulated pathway for mutation clusters*

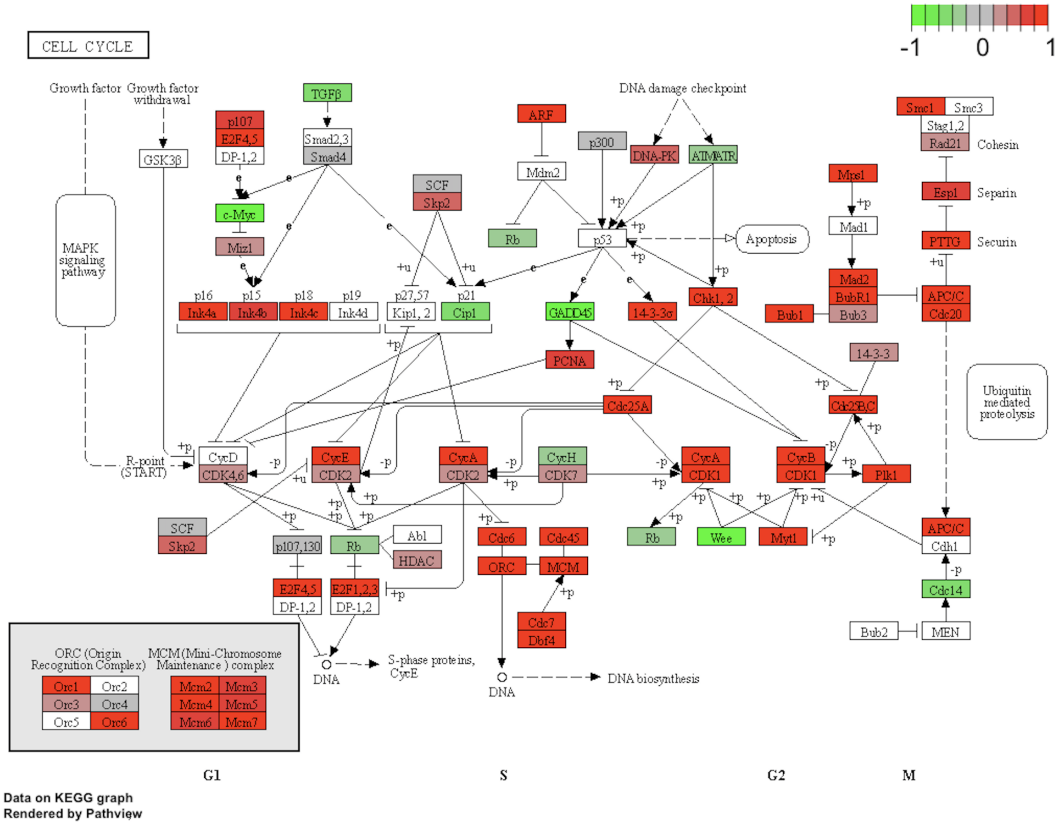


Figure D.2. *Most down-regulated pathway for mutation clusters*