BMEG 310 Final Report

Abstract

A bioinformatic analysis was performed on 352 liver hepatocellular carcinoma (HCC) cases from the Cancer Genome Atlas (TCGA) based on respective clinical, mutation, and RNA sequencing (RNA-seq) data. Through the use of unsupervised clustering, two different mutation profiles were obtained to determine the differences in biological pathway expression and patient outcomes between the two profiles. Survival analysis visualized a significant difference in prognosis between the two subgroups, with mutation profiles containing high levels of tumor protein P53 (TP53) mutations resulting in a significantly worse prognosis. Analysis of sex differences between and within mutation subgroups on patient outcomes was mostly inconclusive, largely due to the lack of female samples.

Subsequent differential expression analysis displayed no correlation between clusters found in mutation analysis and distinct expression profiles. In a practical study, this would have no clinical significance; however, we still chose to perform pathway analysis to show what would have been done with proper cluster mapping. Pathway analysis between mutation profile clusters showed that the group with high TP53 mutations had overexpressed cell cycle and DNA replication pathways and underexpression in metabolism-related pathways. Additional pathway analysis between sexes indicated that sex differences in gene expression could be contributing to these results. Further analysis in comparing gene expression differences between males and females of healthy liver samples should be done to confirm. This analysis provides opportunities to identify the gene-specific origins of HCC, predict HCC prognosis, and direct future therapeutic research tailored to sex-based differences.

Introduction

Liver hepatocellular carcinoma is the fourth leading cause of death worldwide and accounts for 90% of primary liver cancers [1]. Alongside a high mortality rate of 80% at advanced stages, HCC comes with debilitating symptoms, including nausea, enlarged liver, enlarged spleen, abdomen swelling, and fever [2]. 90% of HCC cases are correlated with various risk factors, most commonly involving chronic viral hepatitis, heavy alcoholism, and non-alcoholic fatty liver disease [1]. At the genomic level, previous studies have identified significantly mutated genes related to HCC occurrence, such as TP53 and catenin Beta 1 (CTNNB1) [3].

We aim to investigate if differences in highly mutated genes between HCC patients significantly affect survival outcomes and biological pathways. We also aim to discover if these mutation differences contribute to sex differences in survival and pathway expression. By doing so, certain genes will be identified as potential drug targets in the treatment of HCC. Initiation and progression of HCC is a multi-step process with partially understood underlying molecular events. [3] Therefore, this analysis allows for further defined liver HCC profiles, including how biological pathways patient outcomes are affected, along with a deeper understanding of sex-based differences. Physical symptoms typically do not appear until later stages of cancer [2], so an understanding of genomic correlation to HCC can aid in early diagnosis and treatment.

Methods

Data Wrangling

Clinical, mutation, and RNA-seq data obtained from (TCGA) were pre-processed to filter for only patient samples present in all three datasets with the use of for loops and the Dplyr package. This yielded 352 unique patients across all three datasets. RNA-seq was further filtered to only include one copy of each patient, with the first instance of each duplicate encountered by Dplyr removed.

Mutation Analysis Pipeline

Mutation analysis was then performed to begin exploring the variety of summaries, including (a) gene mutation occurrence (Figure 1), (b) gene mutation occurrence correlated to only high and moderate consequences based on the snpEff annotation document (Figure 2) [4], and (c) variant classification of all mutations (Figure 3).

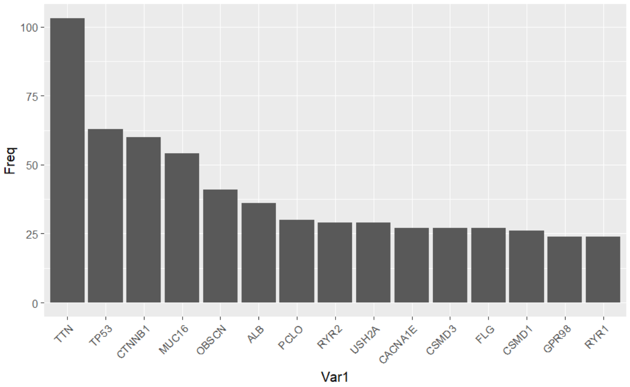

Fig 1. *Frequency of mutated genes across patients.*

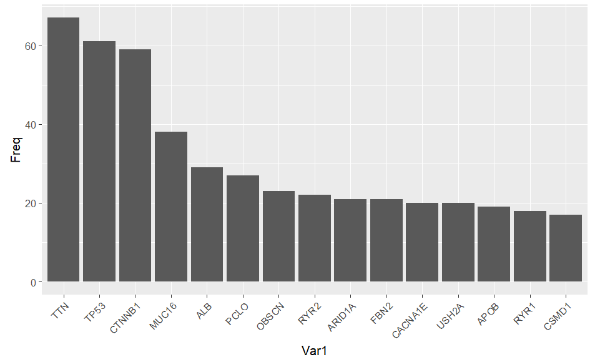

Fig 2. *Frequency of high/moderate impact mutated genes across patients. High and moderate classifications based upon snpEff annotation document [4].*

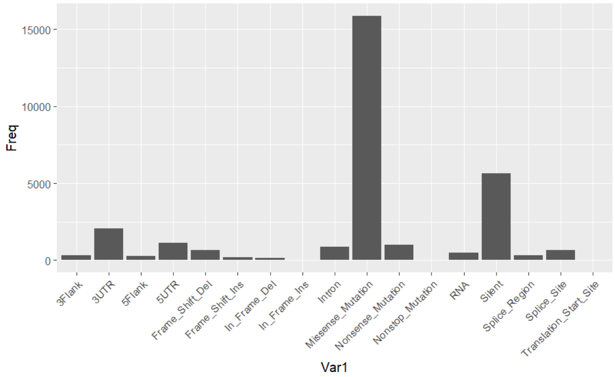

Fig 3. *Frequency of variant types based on classification.*

A data frame was created containing patients' variant classifications correlated to a gene. With a significant number of silent mutations (Figure 3), silent consequences were filtered out in the oncomat's transformation into a binary matrix (assigned value of 0) as these would have no outcome on patient phenotype. With all non-silent mutations converted to a value of 1, a heatmap analysis of the most frequently mutated genes present in more than 10% of patients was performed. The highest occurring mutated genes shared in over 35 patients (352 patients * 0.1) included titin (TTN), TP53, CTNNB1, mucin 16 (MUC16), obscurin (OBSCN), and albumin (ALB) (Figure 4). Unsupervised hierarchical clustering of the heatmap (k = 2) resulted in two subgroups based on mutation profile (Figure 4). Visually, the patients largely appeared to be clustered based on the presence of TP53 mutation. The use of unsupervised clustering allowed for the grouping of similar patients based on their mutation profile, which may correlate to prognosis in subsequent analysis.

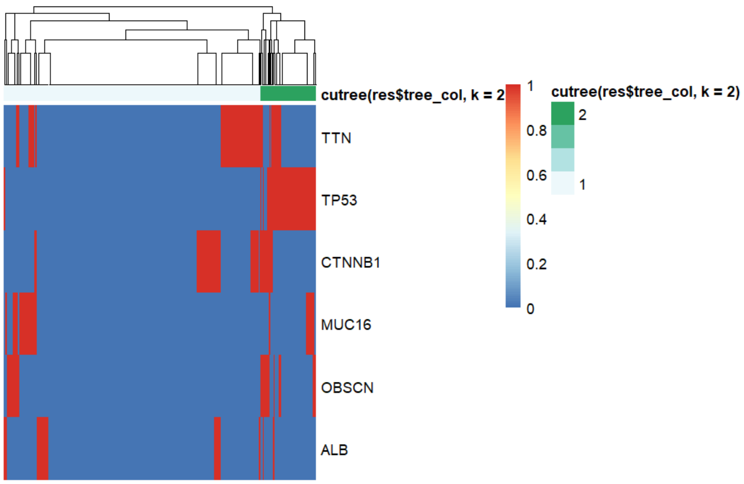

Fig 4. *Heatmap of mutated genes shared in over 10% of patients, ordered by frequency. Grouped by unsupervised hierarchical clustering (k=2).*

For further comparison, another variant classification data frame, binary matrix, and heatmap were generated of only high and moderate consequence gene mutations (Figure 2). However, this led to extremely asymmetrical clusters, with one subgroup containing only two patients which made further analysis difficult (Figure S8). With this small cluster size, it was determined that the clusters based on all non-silent mutations were better suited for subsequent analysis.

The cluster assignment for each patient was then appended to the clinical data frame to perform survival analysis via Kaplan-Meier survival curves. Patients in group 1 were labelled as a non-mutated TP53 type (non-TP53 type), whereas patients in group 2 were considered as a mutated TP53 type (TP53 type). Disease-specific survival (DSS) data, with only dead patients with tumors labelled as deceased, was used to generate survival curves of all patients. Several curves were analyzed, including survival between non-TP53 and TP53 subtypes, sex within subtypes, and sex across subtypes. The p-value was calculated to determine the significance of the survival comparison.

Differential Expression Pipeline

Differential expression of RNA-seq data was also performed to determine if differential expression analysis would generate similar clusters as the mutation analysis. The data matrix was filtered to remove genes with little to no expression across patients. Genes with a cumulative expression sum greater than one across all patients were kept. A principal component analysis (PCA) was applied to the filtered RNA-seq data to identify outliers. PC1 and PC2 were plotted due to having the highest cumulative variance among the PCs, and thresholds of PC1 < 220 and PC2 > -120 were set based on visual inspection (Figure 5). Overall, eight patients were removed from the dataset as a result of this analysis (Figure 6), resulting in 344 remaining patients for differential expression analysis (DESeq).

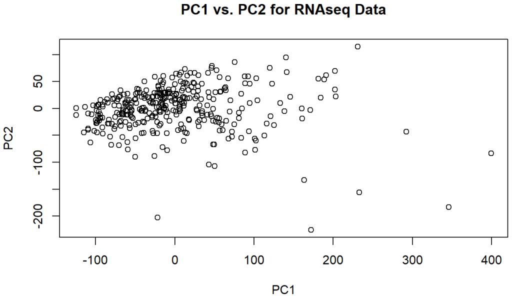

Fig 5. *PC1 and PC2 of RNAseq data including outliers.*

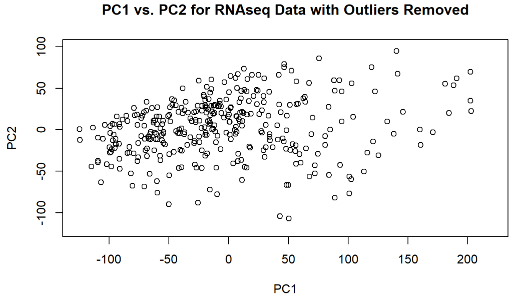

Fig 6. *PC1 and PC2 of RNAseq data with outliers removed.*

The fold change standard was set as non-TP53 (Cluster 1) vs. TP53 (Cluster 2). Any positive log2 fold change values would indicate upregulation in non-TP53; negative would mean downregulation in non-TP53.
A matrix containing the cluster labels for each patient and another matrix containing each patient's RNAseq data were input into the DESeq results pipeline. The results were filtered for a log2 fold change threshold of 1.5 (less than -1.5 or greater than 1.5); these were further filtered for an adjusted p-value of less than 0.01 as this yielded 284 genes, which was deemed a suitable number to input for pathway analysis.

A volcano plot was generated to visualize the significant upregulated and downregulated genes based on the thresholding parameters described above. (Figure 7) The plot showed some significant genes had a log2 fold change that was below the set threshold of less than -1.5 or greater than 1.5. Therefore, to remove these from heatmap visualization, the top 10 most upregulated and top 10 most downregulated significant (p < 0.01) genes were subset from the results.

Variance stabilizing transformation was performed on the DESeq dataset in preparation for generating a PC plot and heatmap, both of which determined whether the non-TP53 and TP53 clusters would map to clusters generated from differential expression analysis. The PC plot showed that when the patients were mapped to non-TP53 and TP53 (cluster 1 and cluster 2), the plot of the first two principal components did not produce any visualizable clusters. (Figure 8) Similarly, hierarchical clustering from the heatmap did not correlate the mutation clusters to those generated in the heatmap. (Figure 9)

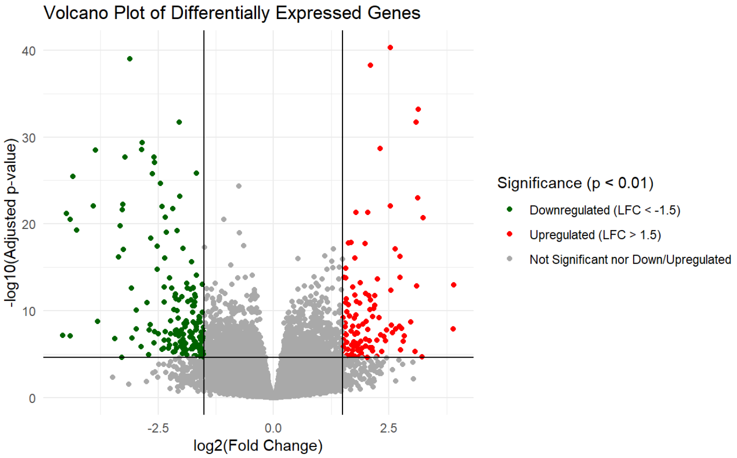

Fig 7. *Volcano plot of differentially expressed genes, thresholded at abs(LFC) > 1.5 and p-adjusted < 0.01. Visually, there appears to be an equal distribution of upregulated genes compared to downregulated genes.*

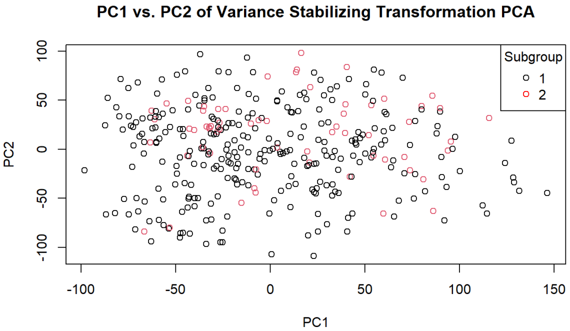

Fig 8. *Plot of the top 2 principal components of the variance stabilizing transformation (vst) of the differential expression data. As shown, the clusters generated by hierarchical clustering of differential expression analysis do not align with those from mutation analysis (1=non-TP53, 2=TP53)*

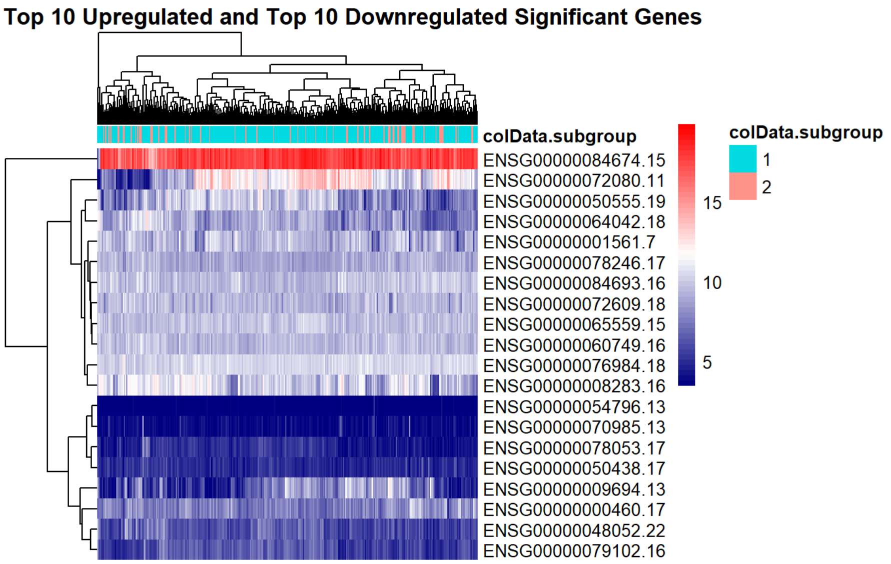

Fig 9. *Heatmap showing the mapping of clusters to the top 10 upregulated and top 10 downregulated significant genes (the x-axis is individual patients, not shown for clarity). As shown, the clusters generated by hierarchical clustering of differential expression analysis in both plots do not align with those from mutation analysis (1=non-TP53, 2=TP53)*

Pathway Analysis Pipeline

Because our mutation clusters displayed no correlation with differences in gene expression, pathway analysis based on TP53 clusters would not be clinically relevant. However, we still chose to perform pathway analysis to show what would have been done with proper cluster mapping.

After duplicate patient samples were removed, pathway analysis was executed with all remaining gene expression data. DESeq values and gene names were mapped to KEGG IDs. To map the gene names in the data structure to kegg genes, all values after the decimal point were removed (e.g. ENSG00000139618.14 would become ENSG00000139618). Initially, pathway analysis was executed using the filtered data in Figure 7; however, no differences in pathway expression were identified. In response to this, pathway analysis was done without filtering, leading to significant differences in pathway expression. The top 5 under and overexpressed pathways in non-TP53 vs. TP53 clusters were collected (Figure S6). As results were only achieved with low p-value and non-significant differences in gene expression, this reinforces our hypothesis that the results of pathway enrichment would be clinically insignificant.

In addition to pathway enrichment between the two clusters, gene expression differences in males and females were analyzed to determine if pathway analysis could provide more insight into potential survival differences. The gene expression was grouped by sex and was input into DESeq. When expression data was filtered with a p-adjusted score of less than 0.01 (indicating significance), no pathways were found. Analysis with no filtering showed that metabolism pathways were underexpressed and DNA replication was overexpressed in females with a significant p-value (See Figure S7). This pathway analysis further emphasizes the need for more female patient data to determine if similar results can be obtained with a higher significance to come to a more confident conclusion on sex differences in liver cancer patients.

Results

From clusters obtained from mutation analysis, Kaplan-Meier survival curves were generated to analyze differences between non-mutated and mutated TP53 types and/or sex.

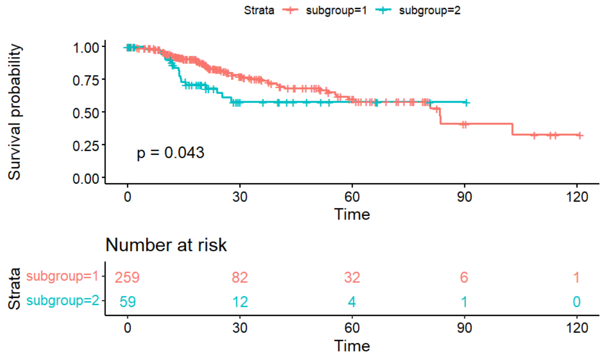

Fig 10. *Kaplan-Meier Curve and risk table between mutation clusters (subgroup 1 = non-TP53, subgroup 2 = TP53) based on DSS, p = 0.043. Comparable trend up to about 30 months. Time in months.*

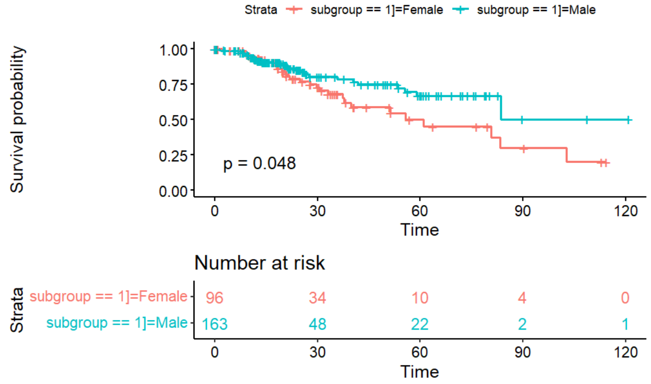

Fig 11. *Kaplan-Meier Curve and risk table between sex in the non-TP53 type (subgroup = 1) based on DSS, p = 0.048. Time in months.*

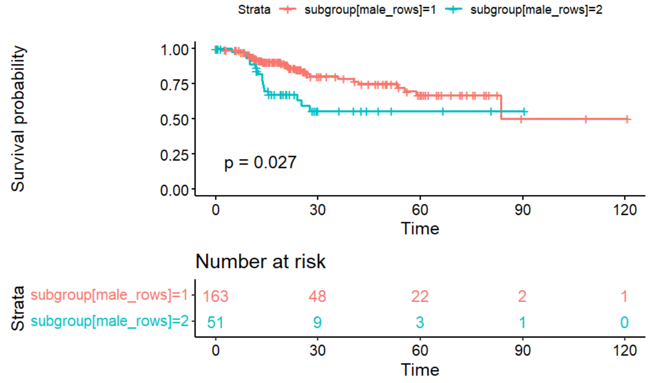

Fig 12. *Kaplan-Meier Curve and risk table between male patients in different mutation clusters (subgroup 1 = non-TP53, subgroup 2 = TP53) based on DSS, p = 0.027. Comparable trend up to about 30 months. Time in months.*

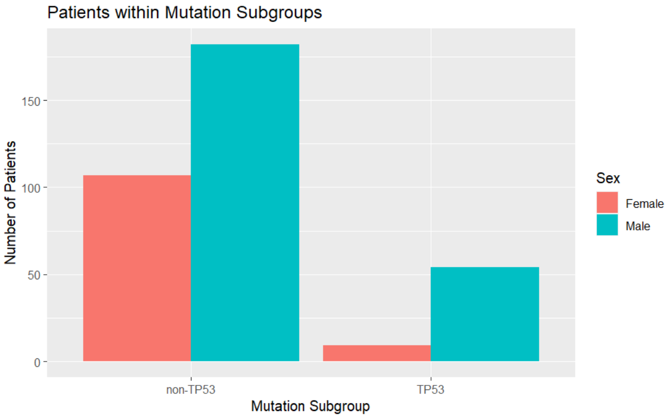

Figure 13. *Bar graph of female and male patients within non-TP53 and TP53 subtypes.*

In all survival analyses involving a mutated TP53 type, this subgroup had worse outcomes and quicker survival decline compared to non-mutated TP53 patients with significant p-values (Figures 10 and 12). Further investigation of survival outcomes based on sex between and within cluster groups was performed but resulted in no significance (Supplementary Figures S1-S5). Additionally, the proportion of females to males was asymmetrical in both clusters (Figure 13), which could have contributed to inconclusive results from sex differences.

As shown in Figures 8 and 9, differential expression analysis indicated that the clusters obtained from hierarchical clustering of significantly expressed genes did not align with non-TP53 and TP53 clusters obtained from the mutation analysis. Therefore, results from pathway analysis are not clinically relevant as data shows no correlation between clusters and gene expression. Over and underexpressed pathways between clusters and sexes can be found in supplementary material figures, S6 and S7. A further explanation of these results can be found in the discussion.

Discussion

In the survival analyses following hierarchical clustering of mutation data (Figures 10, 12, and S2), patients characterized into the TP53 type (group 2) exhibited worse survival outcomes than non-TP53 patients, regardless of sex. Along with being the second-most frequently mutated gene among HCC patients in the dataset, it appears abnormalities in TP53 correlate to HCC occurrence in literature. The significance of TP53 agrees with previous studies, as it is known for its tumor-suppressing functions that regulate cell division to inhibit uncontrolled proliferation [5]. Thus, mutations in the TP53 gene can lead to alterations or loss in its tumor-suppressing abilities, allowing HCC to form. The significance of mutated TP53 further reflects findings by Wheeler & Roberts [3], where TP53 was identified as a significantly mutated gene.

Because the gene expression in the liver is known to be especially dimorphic between males and females [6], we wanted to further analyze the survival difference between sexes. Figure 10 shows our initial survival curve which shows significant differences in survival probability between TP53 type and non-TP53 type, which aligns with results from Wheeler & Roberts [3]. Figure 12, which compares only male-identified patients in each cluster, displays the same results. In graphs where the survival from female-identified patients is segregated, there appears to be a drop in survival probability for the non-TP53 mutated group; however, there are too few patients to make a significant conclusion. Other survival plots were plotted and can be found in the supplemental figures (Figure S1-S5). However, these were not able to provide additional insight as the p-value was not below the significance threshold of 0.05. Difficulty in obtaining significant p-values may have been due to the lack of female samples (Figure 13). Future studies could include more female patients to determine if the TP53 mutation influences the survival of female patients, however collecting this data would be an issue 4 to 1 of liver cancer patients are male [6].

Kim et. al showed [7] that 7 genes (BAP1, CTNNB1, FOXA1, GSTO1, GSTP1, IL6, and SRPK1) showed sex-biased function in liver cancer. Out of these 7, CTNNB1 was the only one present in the mutation clusters, specifically, the non-TP53 cluster. High expression of CTNNB1 in females was associated with a worse survival prognosis compared to low expression (p = 0.026), and CTNNB1 expression did not appear to cause the same effects in males (p = 0.14). Therefore, the poorer prognosis of non-TP53 females could be attributed to higher levels of CTNNB1 expression; however, a comparison of non-TP53 to TP53 females and non-TP53 to TP53 males would be needed to confirm.

Following discussions with Dr. Elizabeth Rideout, the survival differences of TP53 males vs. TP53 females were deemed to be inconclusive due to the disproportionate ratio of females to males in the TP53 subtype (8 females to 51 males) (Figure 13). A similar decision was made regarding survival differences of females between non-TP53 and TP53 types. To determine if the difference in survival is because of sex differences, more data on female patients is required. Alternatively, pathway analysis and differential gene expression analysis may provide further insight when compared to genes with significant expression differences between males and females.

According to pathway analysis, cell cycle, DNA replication, and homologous recombination were significantly more expressed in the non-TP53 group than in TP53, with p values of 5.01E-6, 1.55E-4, and 2.8E-3, respectively. This is the opposite of what was expected as the non-TP53 group had a higher survival probability, but the analysis shows that these pathways are characteristic of more aggressive cancer (Figure S6).

Three metabolic pathways are in the top five underexpressed in non-TP53: fatty acid metabolism, drug metabolism, and retinol metabolism, with respective p-values of 2.95E-4, 1.34E-4, and 6.41E-4 (Figure S7). As the liver's main function is metabolism, these underexpressed pathways may indicate that liver function is decreased in the non-TP53 mutated group. These results are unintuitive from what we saw in our survival plots; however, as our clusters from mutation analysis and differential expression analysis are not correlated, we expected our results to be contradictory to existing literature. From the supplementary figures S6 and S7, the log2 fold change of genes is never greater than 1 or less than -1, meaning that the difference in expression is not significant. We expected this because our clusters did not have distinct expression profiles.

The difference in metabolism pathway expression could also result from the disproportionate amount of females in each group. It is known that females tend to have a slower metabolism than males [8], so the higher proportion of females in the non-TP53 group could be the result of sex differences and are not disease-related. Had our pathway analysis been significant, further studies could have been done to compare sex-biased gene expression in healthy and malignant tumor samples.

Conclusion

Our pipeline showed that mutations in TP53 could be an indicator of worse outcomes for patients with liver cancer; however, our data displays no correlation between gene expression and pathways that could be causing this survival difference. We recommend collecting more data on females to determine if survival differences between sexes exist. Afterward, differences in gene expression between the sexes should be further analyzed to determine the effects of liver cancer on metabolic pathways and survival for both genders. A potential future direction could be generating our clusters from differential expression data instead of mutation data, to determine if these clusters will result in different pathway expressions

Contributions

Laura Siemens (43603836) - Pathway analysis, pathway analysis pipeline in methods, pathway analysis and gender comparison discussion.
Joyce Xi (48582605) - Data wrangling, differential expression analysis, research in sex differences.
Caitlin Ambrose (76312008) - Mutation analysis (heatmap/clustering), survival analysis, TP53 discussion.

All contributed equally to report writing.

References

[1] J. H. Oh and D. W. Jun, “The latest global burden of liver cancer: A past and present threat,” Clinical and molecular hepatology, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10121295/ (accessed Dec. 7, 2023).

[2] “Signs and symptoms of liver cancer,” American Cancer Society, https://www.cancer.org/cancer/types/liver-cancer/detection-diagnosis-staging/signs-symptoms.html#:~:text=Some%20liver%20tumors%20make%20hormones,can%20cause%20fatigue%20or%20fainting (accessed Dec. 7, 2023).

[3] D. A. Wheeler and L. R. Roberts, “Comprehensive and integrative genomic characterization of hepatocellular carcinoma,” NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5680778/ (accessed Dec. 7, 2023).

[4] P. Cingolani, F. Cunningham, W. McLaren, and K. Wang, Variant annotations in VCF format, https://snpeff.sourceforge.net/VCFannotationformat_v1.0.pdf (accessed Dec. 7, 2023).

[5] “TP53 gene,” MedlinePlus, https://medlineplus.gov/genetics/gene/tp53/#:~:text=Normal%20Function,or%20in%20an%20uncontrolled%20way (accessed Dec. 7, 2023).

[6] W.-L. Liou et al., “Gender survival differences in hepatocellular carcinoma: Is it all due to adherence to surveillance? A study of 1716 patients over three decades,” NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10230112/#:~:text=Hepatocellular%20carcinoma%20(HCC)%20is%20the,of%20cancer%E2%80%90related%20mortality%20worldwide.&amp;text=The%20incidence%20of%20HCC%20is,ratio%20of%204%20to%201 (accessed Dec. 7, 2023).

[7] S. Y. Kim et al., “Sex-biased molecular signature for overall survival of liver cancer patients,” NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7585639/ (accessed Dec. 7, 2023).

[8] D. J. Waxman and M. G. Holloway, “Sex differences in the expression of hepatic drug metabolizing enzymes,” NCBI, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2713118/ (accessed Dec. 7, 2023).

Supplementary Figures

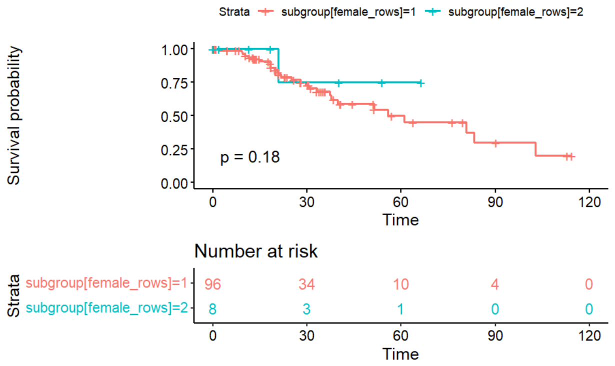

Figure S1. Kaplan-Meier Curve and risk table between female patients in different mutation clusters (TP53 and non-TP53 type) based on DSS, p = 0.18. Time in months.

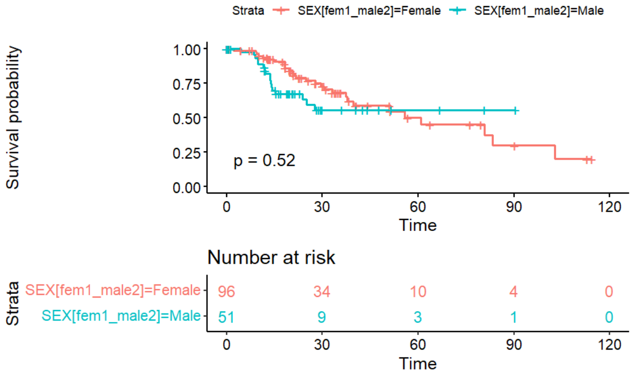

Figure S2. Kaplan-Meier Curve and risk table between non-TP53 type females and TP53 type males based on DSS, p = 0.52. Time in months.

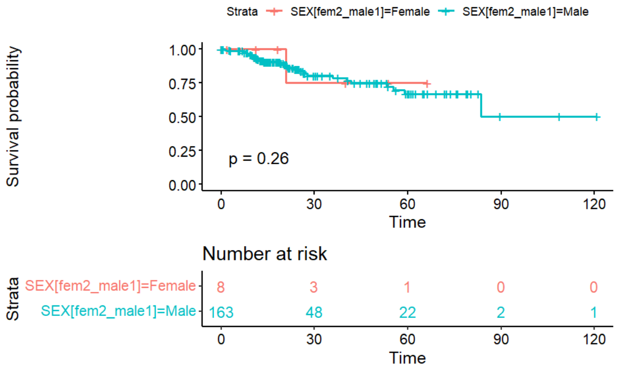

Figure S3. Kaplan-Meier Curve and risk table between TP53 type females and non-TP53 type males based on DSS, p = 0.26. Time in months.

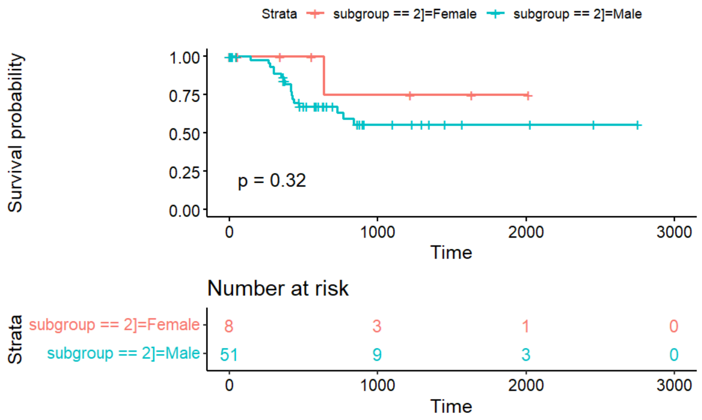

Figure S4. Kaplan-Meier Curve and risk table between TP53 type females and TP53 type males based on DSS, p = 0.32. Time in months.

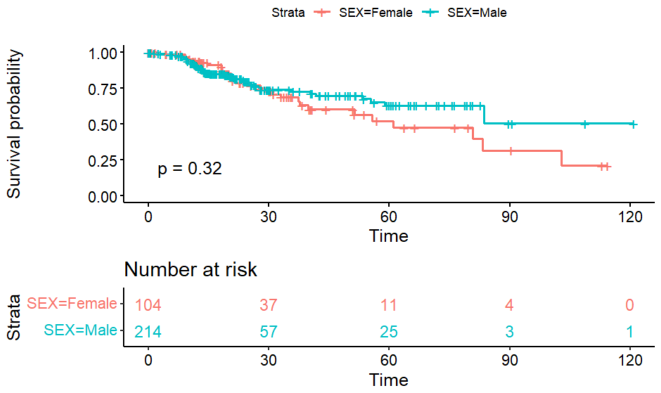

Figure S5. Kaplan-Meier Curve and risk table between all females and all males based on DSS, p = 0.52. Time in months.

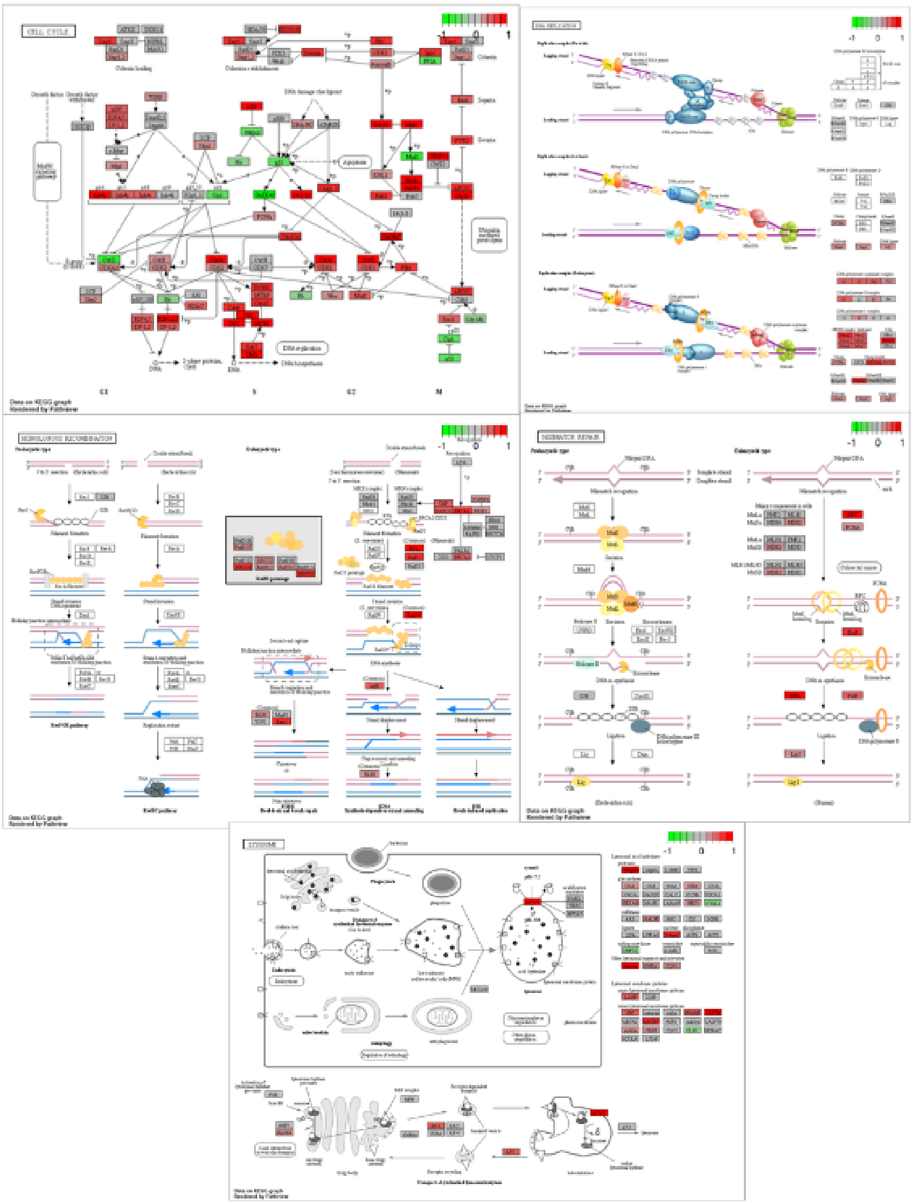

Figure S6. Top 5 significant overexpressed pathways in TP53 group when compared to non-TP53 group. From top left to bottom right, with the most upregulated at the top, the pathways are: cell cycle, DNA replication, homologous recombination, mismatch repair, and lysosome.

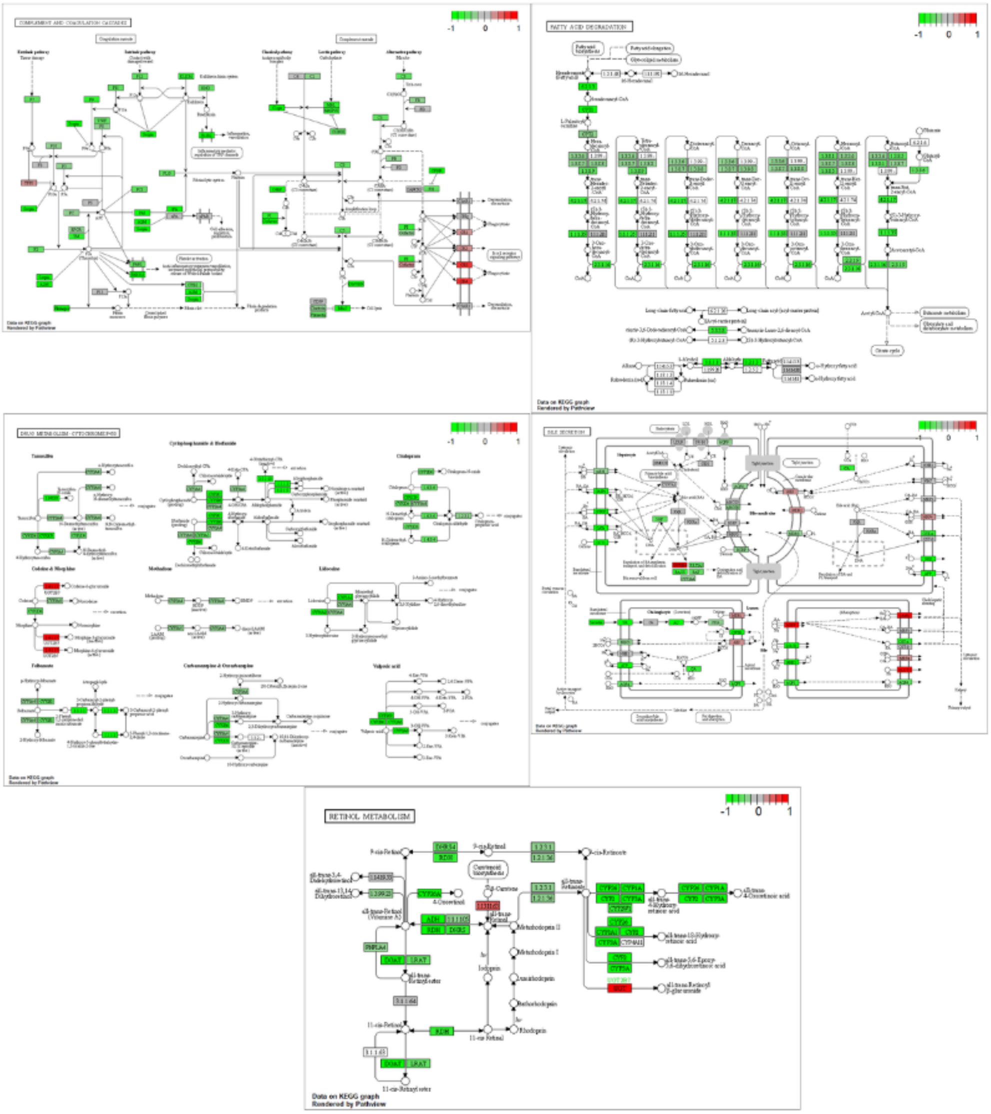

Figure S7. Top 5 significant underexpressed pathways in TP53 group when compared to non-TP53 group. From top left to bottom right, with the most downregulated at the top, the pathways are: Complement and coagulation cascades, fatty acid metabolism, drug metabolism, bile secretion, and retinal metabolism

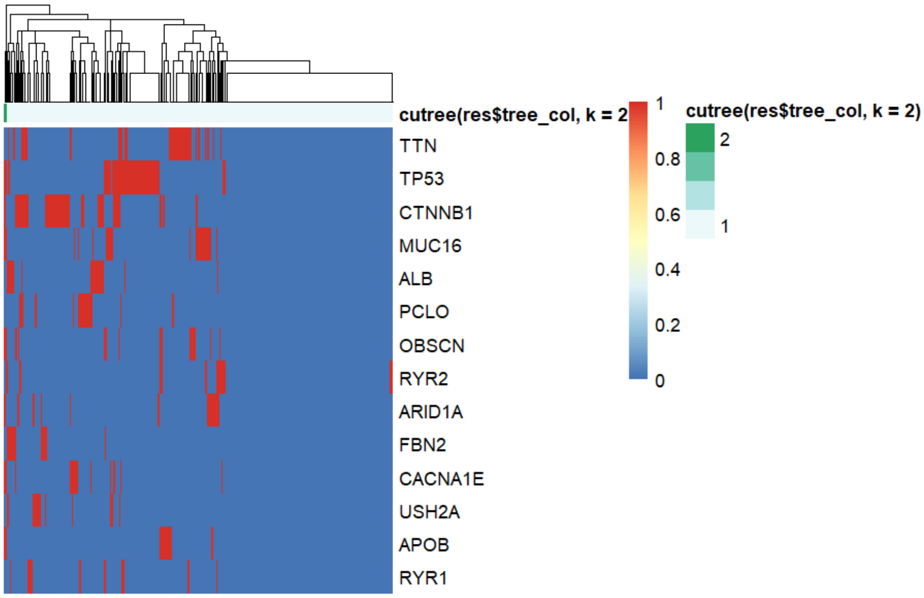

Figure S8. Heatmap and hierarchical clustering (k=2) filtered for only high/moderate consequence variants. Resulted in extremely asymmetrical clusters, with one group containing only 2 patients.

Underexpressed pathways in females

```
                                                p.val
hsa00071 Fatty acid metabolism              2.950824e-05
hsa00982 Drug metabolism - cytochrome P450 1.345140e-04
hsa00830 Retinol metabolism                 6.409657e-04
hsa00350 Tyrosine metabolism                1.027774e-03
hsa00380 Tryptophan metabolism              2.152422e-03
hsa04020 Calcium signaling pathway          3.684076e-03
```

Overexpressed pathways in females

```
                                                             p.val
hsa03030 DNA replication                                  0.0001553537
hsa00030 Pentose phosphate pathway                        0.1130155900
hsa00520 Amino sugar and nucleotide sugar metabolism      0.1619326770
hsa03020 RNA polymerase                                   0.1943904768
hsa00604 Glycosphingolipid biosynthesis - ganglio series 0.2355260014
hsa04914 Progesterone-mediated oocyte maturation          0.2525086616
```

Table T1: The over and underexpressed pathways and their p values when comparing sex.