Title: Sex-biased gene expression and pathway activation in hepatitis-associated HCC.

By: Annika Jorgensen

May 2023

Table of Contents

Abstract

Hepatocellular Carcinoma (HCC) is one of the main types of liver cancer accounting for 75% of cases and is the second deadliest cancer worldwide. Chronic Hepatitis B (HBV) and Hepatitis C (HCV) remain one of the most important global risk factors and account for 80% of all HCC cases. HCC also exhibits sex-differences with significantly higher incidence and worse prognosis in males. The mechanistic basis of these sex-differences is poorly understood. To identify genes and pathways that are sex-differentially expressed in viral-mediated HCC, we performed differential expression analysis on tumor vs. tumor adjacent samples that were stratified based on sex, viral etiology, and both. The differentially expressed genes were then used in a pathway enrichment analysis to identify potential pathways of interest. We found differentially expressed genes in both sexes and both etiologies. 65 genes were unique to females and 184 genes unique to males. 381 genes are unique to HBV and 195 genes are unique to HCV. We also found pathways that were significantly enriched by the differentially expressed genes. Ten pathways unique to the female tumor tumor-adjacent comparison and a majority of those pathways were a part of the cell cycle. Four enriched pathways unique to male tumor tumor-adjacent and three of them were a part of the immune system. There were no pathways unique to either etiology, but seven pathways shared by both etiologies. Two were a part of the cell cycle and one involved lipid metabolism.These differentially expressed genes and significant pathways are potential targets for individualized therapeutics and diagnostics for HCC.

Background

Liver cancer is the second deadliest cancer worldwide, comprising 8.3% of all cancer related deaths (Chhikara and Parang 2023). Hepatocellular Carcinoma (HCC) is one of the main types of liver cancer accounting for 75% of all cases (Altekruse et al. 2011). Hepatocellular Carcinoma (HCC) is increasing in prevalence in many countries including the United States (El-Serag 2012). Hepatitis B virus and Hepatitis C virus remain the most important global risk factors for HCC, where approximately 80% of cases are mediated by one of the viruses (McGlynn et al. 2021), (El-Serag 2012).

HCC has known sex differences in incidence and prognosis and these differences are not fully understood. Males have a two to four times higher incidence of HCC compared to females in most countries (McGlynn et al. 2021). Males are also two times more likely to die from chronic liver disease and cirrhosis than are females (Guy and Peters 2013). Sex differences in HCC remain even after adjusting for risk factors including age, race, smoking, alcohol intake and others (Yuan et al. 2016). In order to further explain these differences, a greater understanding of the fundamental mechanisms driving this sex disparity is needed. Understanding these mechanisms would illuminate potential targets for individualized therapies and treatment of HCC.

Previous studies have examined the gene expression differences between males and females that may contribute to the sex bias in HCC incidence and prognosis. One study by Yuan et al. observed gene regulation and somatic mutations in HCC (2016). In this analysis a comparison of tumors was done and sexual differences in normal tumor-adjacent tissues were not accounted for. Comparing tumor tissue to normal tumor-adjacent tissue is important because the comparison accounts for inherent sex differences. Another study by Natri et.al, has shown sex-specific differences in gene expression and sex-specific regulatory functions across tumor and tumor-adjacent liver tissue (2019). However, this study did not account for viral etiology or the differences in viral-mediated HCC.

The creation of individualized cancer diagnostics and therapeutics for males and females diagnosed with viral-mediated HCC requires an understanding of the biological mechanisms underlying the sex differences. We analyzed whole transcriptome data obtained from the International Cancer Genome Consortium LIRI-JP dataset and performed tumor versus tumor-adjacent differential expression analysis and pathway enrichment analysis (Zhang et al. 2019). We made comparisons stratified by sex, etiology and a combination of both. We reveal that viral-mediated HCC exhibits sex-specific differences in gene expression. We also identified sex-differnces in the pathways that are enriched in the differentially expressed genes. These results are expected to guide future development of sex-specific diagnostics and therapeutics for males and females with viral-mediated HCC.

Methods

Samples

Whole transcriptome data (RNA-seq) from 260 donors representing tumor and tumor-adjacent viral-mediated Hepatocellular Carcinoma samples were obtained from the International Cancer Genome Consortium LIRI-JP dataset (controlled access permission to Dr. Ken Buetow) (Zhang et al. 2019). Tumor-adjacent samples were taken from adjacent healthy liver tissue. The distribution of samples by tumor tumor-adjacent, sex, and etiology can be seen in Table 1 (Table 1).

Alignment and Quantification

RNA sequencing FASTA files underwent quality control using FASTQC (Babraham bioinformatics - FastQC A qu...)Data was trimmed using Trimmomatic with parameters of 2 seed mismatches, palindrome clip threshold 30, simple clip threshold 10, leading quality value 3, trailing quality value 3, sliding window size 4, minimum window quality 30, and minimum read length of 50 (Bolger et al. 2014). Read mapping was performed with Hisat2. To overcome mismapping of short sequencing reads due to sequence homology on X and Y chromosomes the reads were mapped to a sex-specific reference genome. The specific human genome was GRCh38.p7 from Gencode (GENCODE - Human Release 25). Read count quantification was performed with *Subread featureCounts* (Liao et al. 2014). The sample with sample ID "RK023" was removed from the dataset due to low quality.

Filtering and Processing

FPKM (fragments per kilobase of exon per million fragments mapped) and TMM (Trimmed Mean of M-values) were obtained using *EdgeR* (Robinson et al. 2010). The dataset retained genes with a mean FPKM value of ≥ 0.5 and read count of > 6 in at least 10 samples across all samples under investigation.

Differential Expression

Differential Expression analyses were performed for all tumor tumor-adjacent data, tumor tumor-adjacent data stratified by etiology, tumor tumor-adjacent data stratified by sex, and tumor tumor-adjacent data stratified by etiology. The design matrix created to fit the linear model had lesion type as a predictor variable. Library type was added as a covariate to the model to account for differences in the strandedness of the samples (Supplementary Figure 1). Sex was added as a covariate to the overall tumor tumor-adjacent analysis, and analysis stratified by etiology. Raw counts per million reads (CPM) were log2 normalized and adjusted for quality using the *voomWithQualityWeights* function in the limma R package(Supplementary Figure 2,4,6,8), and passed into the limma pipeline (Supplementary Figure 3,5,7,9) (Law et al. 2014). The *duplicateCorrelation* function was used to compute the correlation between tumor

tumor-adjacent samples on the same patient (Ritchie et al. 2015). The correlation was also included in the limma pipeline. Genes were assumed to be differentially expressed if they had log fold change (logFC) ≥ 2 and p value of < 0.05. Empirical Bayes smoothing increased the power of the analysis. Volcano plots visually represented the significant differentially expressed genes.

Pathway Enrichment Analysis

Genes deemed significantly differentially expressed by the DE analysis were compiled and analyzed using Reactome over-representation analysis (Fabregat et al. 2018). Reactome identified pathways that were significantly enriched given the genes in the list. Pathways were included in the pathway tables if they had a p-value of less than 0.05. Upset plots were generated using the UpSetR package (Conway et al. 2017) to visually represent the enriched pathways unique and shared between sexes and etiologies.

Results

Sex drives differential gene expression in viral-mediated hepatocellular carcinoma

We performed a differential expression analysis between all tumor and tumor-adjacent samples and identified 509 genes downregulated in tumors and 158 genes upregulated in tumors (Figure 2). To examine the characteristics responsible for the differentially expressed genes, we performed a multidimensional scaling (MDS) analysis on the top 50 differentially expressed genes. The first principal component accounts for 37% of the variation in the samples, and can be attributed to tumor tumor-adjacent differences. The second principal component accounts for 12% of the variation in the samples and can be attributed to male vs. female sex. The large degree of variation attributable to sex demonstrates that sex may drive differential gene expression.

To examine genes and pathways influenced by sex variation, we completed a differential expression analysis on female tumor tumor-adjacent and male tumor tumor-adjacent samples. We found 430 genes were downregulated in tumors and 157 were upregulated in tumors in females and 542 genes were downregulated in tumors and 164 upregulated in tumors in males (Figure 3 and 4). A majority (67.7%) of the genes are shared between males and females, however, 65 genes (8.4%) are unique to female tumor tumor-adjacent samples and 184 genes (23.9%) are unique in male tumor tumor-adjacent samples.To account for small female sample size, the p value threshold for female differentially expressed genes was relaxed from 0.05 to 0.1 to see if more genes were differentially expressed. No additional genes were identified.

Multidimensional scaling analysis does not reveal a strong effect of etiology on sample variation

To investigate the effect of etiology on differential gene expression, a MDS analysis was performed on the top 50 differentially expressed genes. The gene lists were subsetted by sex and tumor vs. tumor-adjacent samples (Figure 5). The MDS plots did not show a strong effect of etiology and did not explain the variation in the samples. We completed differential expression analysis on liver tissue samples infected with Hepatitis B virus (HBV) and samples infected with Hepatitis C virus (HCV) (Figure 6 and 7). In HBV tumor and tumor-adjacent samples, we found 586 genes downregulated in tumors and 238 genes upregulated in tumors. In HCV tumor and tumor-adjacent samples, we found 514 genes are downregulated in tumors and 124 genes are upregulated in tumors. Majority of the genes are shared between HBV and HCV tumor vs. tumor-adjacent comparisons, where 443 (43.5%) genes are shared, 381 (37.4%) genes are unique to HBV tumor tumor-adjacent, and 195 (19.1%) genes are unique to HCV tumor tumor-adjacent. While the MDS plot shows that etiology has little effect on sample variation, the differential expression analysis showed that there are genes that are unique to each etiology that warrant further investigation.

Majority of differentially expressed genes separated by both sex and etiology are shared by both males and females.

A differential expression analysis was completed on samples subsetted by both sex and etiology. In the male HBV samples, we found 612 genes downregulated in tumors and 232 genes upregulated in tumors, and in male HCV samples we found 578 genes downregulated in tumors and 140 genes upregulated in tumors (Figure 10 and 11). In the female HBV samples, we found 543 genes downregulated and 305 genes upregulated, and in the female HCV samples we found 435 genes downregulated and 136 upregulated (Figure 8 and 9). Almost half of the differentially expressed genes are shared between male and female HBV where 554 (48.8%) genes are shared, 290 (25.5%) are unique to male HBV, and 292 (25.7%) are unique to female HBV. Over half of differentially expressed genes are also shared between male and female HCV where 481 (59.5%) genes are shared, 237 (29.3%) genes are unique to male HCV, and 90 (11.1%) genes are unique to female HCV. Further investigation is needed to see the effect of etiology and small sample size.

Pathway analysis reveals enrichment of cell cycle pathways in female differentially expressed genes and immune pathways in male differentially expressed genes.

To identify pathways enriched in males and females, differentially expressed genes in female tumor vs. tumor-adjacent samples and male tumor tumor-adjacent samples were analyzed with Reactome. Pathways with a p value of less than 0.05 were considered enriched. Ten enriched pathways were unique to female tumor-adjacent cells (Figure 12). Five out of the ten pathways were involved with the cell cycle.

Differentially expressed genes in male tumor vs. tumor-adjacent and overall tumor vs tumor-adjacent comparisons were submitted to Reactome where a p value of less than 0.05 were considered enriched. There were a small number of pathways that were enriched where two pathways were found unique to male tumor vs. tumor-adjacent and two pathways were shared between overall tumor vs. tumor-adjacent comparison (Figure 12). Three out of these four pathways were a part of the immune system. The other pathway was involved with the transport of small molecules (Table 4).

Differentially expressed genes in Hepatitis B tumor vs. tumor-adjacent tissue and Hepatitis C tumor vs. tumor-adjacent tissue have no pathways unique to a specific etiology.

To see enriched pathways that are associated with the etiology of the cancer, differentially expressed genes from HBV tumor vs. tumor-adjacent and HCV tumor vs. tumor-adjacent

comparisons were submitted to Reactome. Pathways were considered enriched if they had p value of less than 0.05. There were no pathways that were unique to HBV or HCV and 7 pathways shared between HBV and HCV etiologies (Figure 13). Two of the seven pathways were associated with the cell cycle and one involved lipid metabolism (Table 5).

Discussion

Drivers of differential gene expression in viral-mediated HCC

Sex-biased expression in HCC and other cancers has been established. For example Natri et al. previously reported differential gene expression and regulatory networks in HCC between males and females (Natri et al. 2019). However, this analysis did not examine viral-mediated tissue and did not include viral etiology as a variable in their comparisons. The results presented here include comparisons of sex, viral etiology, and tumor tumor-adjacent tissues. These comparisons account for inherent sex differences and viral etiological differences presented,as well as detect sex or viral etiology specific differentiated genes that are dysregulated in HCC.

We found sex to be a driver of differential gene expression. The differences between our tumor vs. tumor-adjacent comparison points to gene dysregulation in tumors which is consistent with our understanding of carcinogenesis (Figure 2). The sex specific comparisons reveal a number of genes expressed in a sex-biased way (Figure 3-4). This result is consistent with our understanding of HCC as a sex-biased cancer and consistent with other reports of sex specific gene dysregulation in HCC (McGlynn et al. 2021), (Natri et al. 2019).

A couple genes of note are H19 and EEF1A2. H19 is not significant in the comparison of female tumor to tumor-adjacent tissues and is upregulated in tumors for males. H19 is an oncogenic long coding RNA which is paternally imprinted and maternally expressed. H19 is expressed in the embryo, repressed at birth and then expressed in some cancers (Ghafouri-Fard et al. 2020). H19 has shown potential effects as an oncogene in HCC in mouse models (Gamaev et al. 2021), (Yoshimizu et al. 2008). EEF1A2 is also not significant in female tumor tumor-adjacent and is downregulated in male tumors. EEF1A2, which may be critical to the development of ovarian cancer, is a gene located on the 20th chromosome and encodes two isoforms of the alpha subunit of the elong factor-1 complex. The alpha 1 isoform is expressed in the brain, placenta, lung, liver, kidney, and pancreas and alpha 2 is expressed in brain, heart and skeletal muscle. (EEF1A2 eukaryotic translation elongat...). One study has shown that EEF1A2 mRNA and protein levels have been found to be significantly higher in HCC cancer tissue and that EEF1A2 is a potential oncogene (Qiu et al. 2016). H19 and EEF1A2 illustrate the broader potential of sex-biased differentially expressed genes found in viral-mediated HCC and are potential targets for individualized medicine.

The results presented here show how etiology drives differential gene expression. The differential expression analysis revealed many unique genes in each viral etiology (Figure 6-7). The unique genes suggest the possibility of significant differentiated genes that are etiology specific. However, the MDS analysis found that etiology does not have a strong effect on sample variation (Figure 5). This analysis suggests that viral etiology is not a significant factor in differential gene expression. The number of genes unique to each etiology seems to contradict the results from the MDS plot. The number of HBV samples is considerably smaller than the number of HCV samples so relaxing the p-value may cause a considerable change in

differentiated genes. Additionally, the MDS analysis was only done with the top 50 differentiated genes and two principal components , so increasing the number of genes drastically to 500 genes and increasing the number of principal components may also help explain the contradiction. Further investigation into each possibility are future directions for this project.

To see how etiology and sex drive differential gene expression together we completed differential expression analysis on samples that were subsetted by both etiology and sex. A majority of the differentially expressed genes are shared, however, the number of unique genes proposes a potential sex- and etiology biased difference in HCC carcinogenesis (Figures 8-11). In order to validate these results further investigation is needed into the effect of etiology on differentiated genes and sample size difference. We relaxed the p-value threshold for differentially expressed genes expressed in the female tumor vs. tumor-adjacent comparison from 0.05 to 0.1 to account for small female sample size. The number of unique female HBV increased from 292 to 305 and the number of shared genes between female and male HBV increased from 554 to 563. The number of male and female HCV differentially expressed genes did not change. Further investigation is needed to adjust for sample sizes to better assess the overlap of the differentially expressed genes, and this investigation is a future direction of this project.

Sex and Viral etiology biased pathways in viral-mediated hepatocellular carcinoma

The pathway enrichment analysis done on the differentially expressed genes from the overall tumor tumor-adjacent comparison, the male and the female tumor tumor-adjacent comparison found enriched pathways that were unique to each sex (Figure 12, Table 3 and 4). Males had three immune system pathways enriched with two of the enriched pathways related to cytokine signaling. Cytokines have been found to be closely associated with HCC with relevance in advanced stages of HCC, diagnostics, and therapeutics. The IL-10 cytokine signaling pathway was enriched in males and its role in HCC is not clearly understood. Studies have linked high levels of IL-10 protein to lower survival rates (Rico Montanari et al. 2021). Females had 5 enriched pathways associated with the cell cycle. Four of the five pathways were associated with the mitotic cell cycle. Mitotic cell cycle pathways have shown to be enriched HCC(Yan et al. 2017), (Natri et al. 2019). The cytokine signaling and mitotic cell cycle pathways are enriched pathways and are potential targets for sex specific diagnostic and therapeutic efforts.

The pathway enrichment analysis done on the differentially expressed genes from the overall tumor tumor-adjacent comparison, the HBV and the HCV tumor-adjacent comparison found only enriched pathways that were shared by both HBV and HCV (Figure 13, Table 5). Phosphorylation of Emi1 has been shown to be phosphorylated by CDKs (Moshe et al. 2011). Additionally, HCV has been shown to stimulate the G1/S pathway which in turn activates the CDKs. Another pathway that was enriched was Synthesis of (16-20)-hydroxyeicosatetraenoic acids which is a pathway associated with lipid metabolism. HCV dysregulates host lipid metabolism which causes liver fat accumulation (Vescovo et al. 2016). The pathways enriched in our samples have known associations with both HCC and HCV, providing further evidence of

etiology driven gene expression. Further investigation is needed to confirm the results and ensure the results are not a product of the difference in HBV and HCV sample sizes. Future directions of the project include the confirmation of viral etiology enriched pathway results and a sex and etiology pathway analysis which will further elucidate sex and etiology-biased pathways.

Conclusions

The goal of this analysis was to determine sex differences in differentially expressed genes and pathways in viral mediated hepatocellular carcinoma. We aimed to do this by identifying differentially expressed genes through differential expression analysis and identifying enriched pathways from the identified differentially expressed genes. We identified differentially expressed genes that were sex specific and viral etiology sex. We also identified genes that were box sex and viral-etiology specific. We identified enriched cell cycle and cytokine signaling pathways that were sex specific and had known associations with HCC. We also identified pathways that had known associations with HCV and HCC. Further investigations are needed to validate the sex and viral specific gene expression on tumors, its effects in HCC and other viral-mediated sex-specific cancers across diverse populations.

Figures

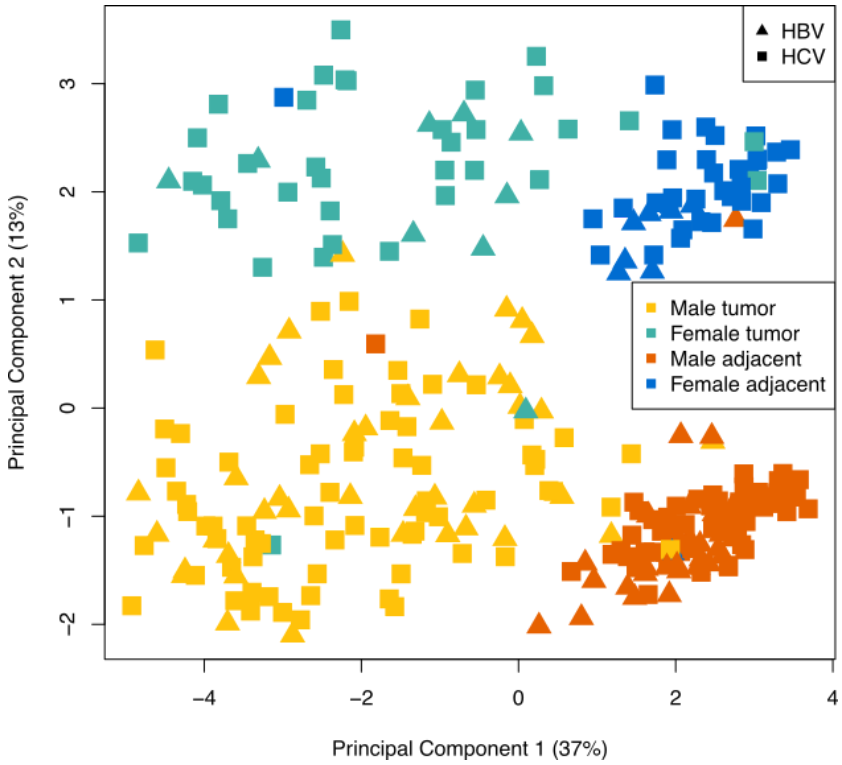


Figure 1. Multidimensional scaling analysis demonstrates the impact of sex on driving differential gene expression

The MDS plot shows the results of a multidimensional scaling (MDS) analysis of the top 50 differentially expressed genes selected for all pairwise comparisons. The MDS plot is colored by sex and tissue type where yellow is male tumor samples, green is female tumor samples, orange is male tumor-adjacent samples, and blue is female tumor-adjacent samples. The triangle shapes represent liver tissue samples infected with Hepatitis B and the square shapes represent liver tissues infected with Hepatitis C. Principal component 1 accounts for 37% of the difference in data and is attributable to tumor vs tumor-adjacent tissue and principal component 2 accounts for 12% of the data and is attributable to male vs female sex.

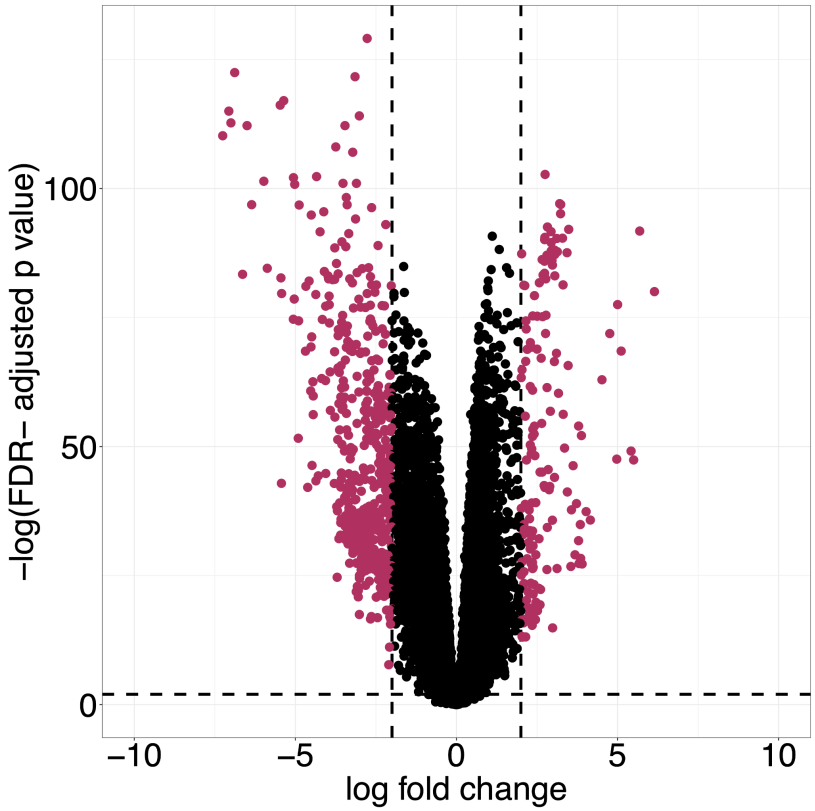


Figure 2. Differentially expressed genes identified in overall tumor vs. tumor-adjacent comparison.

509 genes were downregulated in tumors and 158 genes were upregulated in tumors. The dashed lines indicate the levels of significance with the vertical lines representing an absolute logFC ≥ 2 and the horizontal line representing a p-value of 0.05. Maroon dots indicate genes with an absolute value of logFC ≥ 2 and an adjusted p-value < 0.05 and are considered differentially expressed.

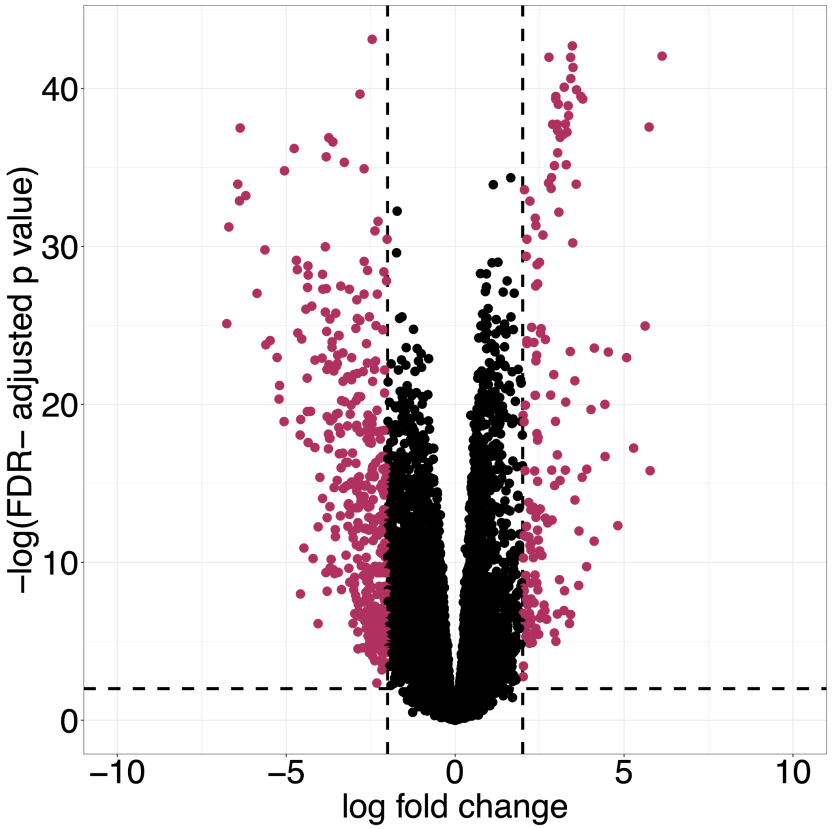


Figure 3. Sex-biased differentially expressed genes identified in female tumor vs. tumor-adjacent comparison

Here we visually represent the differentially expressed genes in female tumor tumor-adjacent data. 430 genes were down regulated in tumors and 157 were upregulated in tumors. The dashed lines indicate the levels of significance with the vertical lines representing an absolute logFC ≥ 2 and the horizontal line representing a p-value of 0.05. Maroon dots indicate genes with an absolute value of logFC ≥ 2 and an adjusted p-value < 0.05 and are considered differentially expressed.

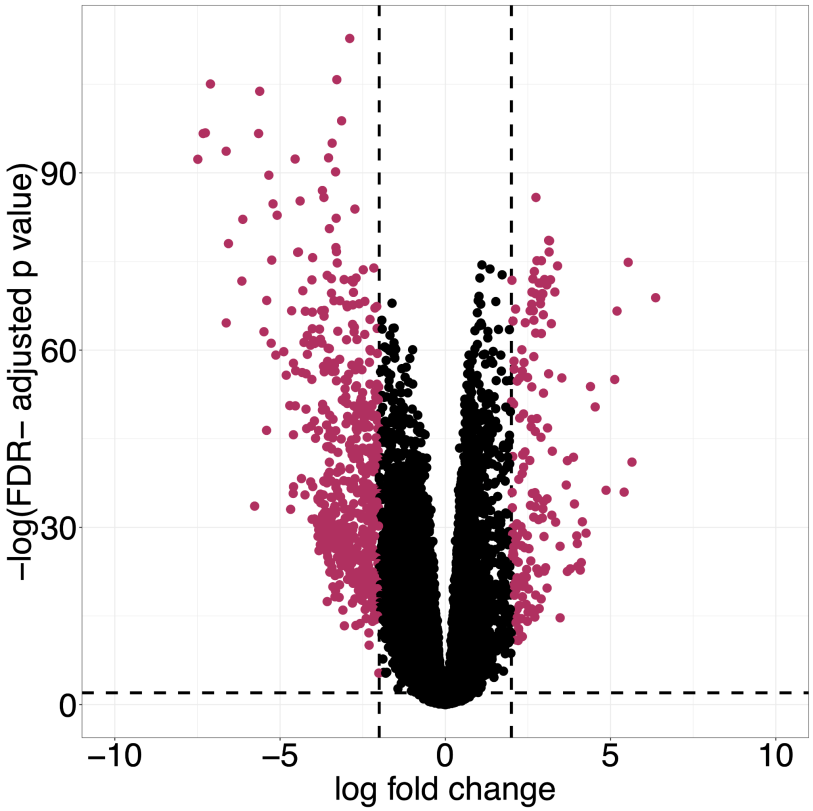
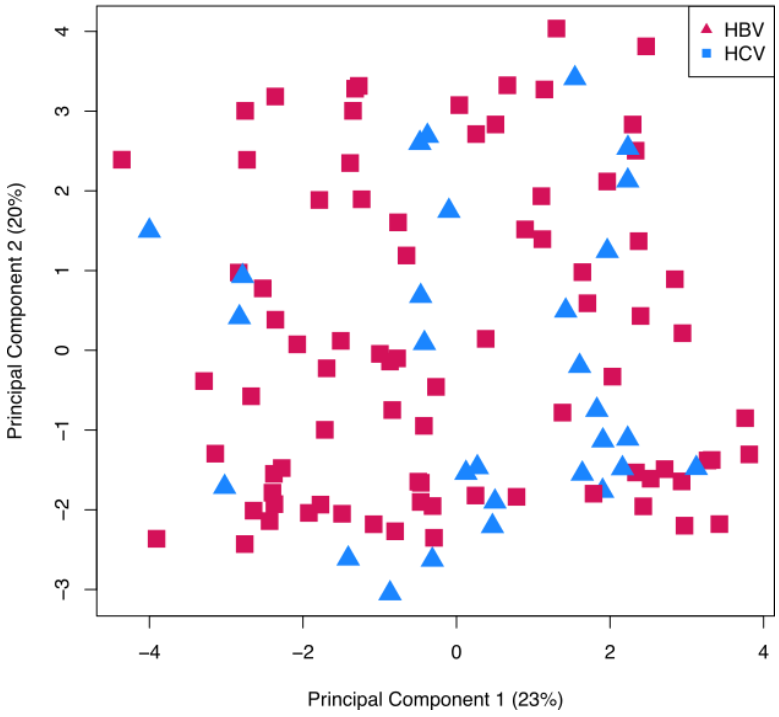


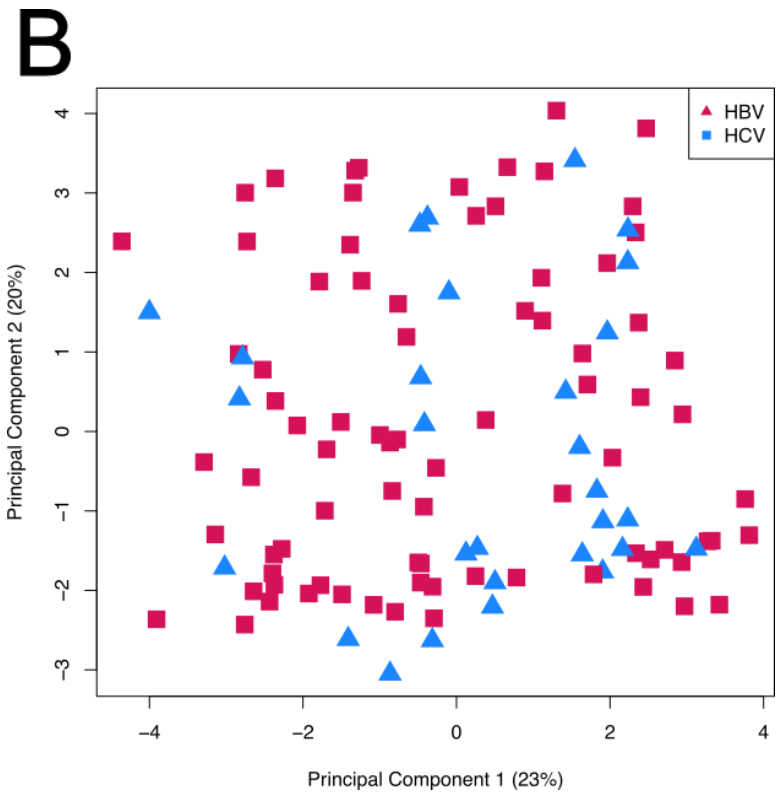
Figure 4. Sex-biased differentially expressed genes identified in male tumor tumor-adjacent comparison.

Here we visually represent the differentially expressed genes in male tumor tumor-adjacent data. 542 genes were downregulated in tumors and 164 upregulated in tumors The dashed lines indicate the levels of significance with the vertical lines representing an absolute logFC ≥ 2 and the horizontal line representing a p-value of 0.05. Maroon dots indicate genes with an absolute value of logFC ≥ 2 and a adjusted p-value < 0.05 and are considered differentially expressed.

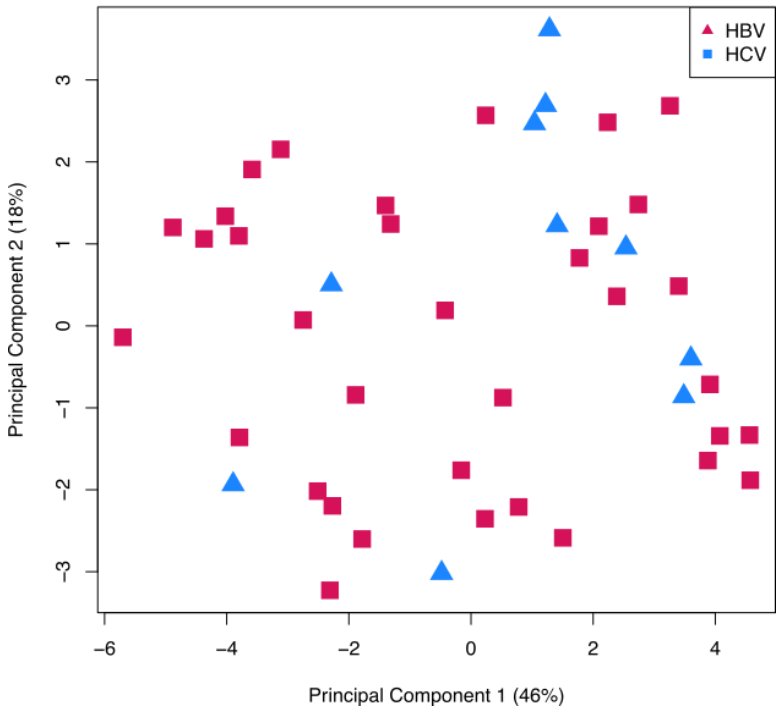
A


HBV
HCV
Principal Component 2 (20%)
Principal Component 1 (23%)
−4
−2
0
2
4
−3
−2
−1
0
1
2
3
4

B


HBV
HCV
Principal Component 2 (20%)
Principal Component 1 (23%)
−4
−2
0
2
4
−3
−2
−1
0
1
2
3
4

C


HBV
HCV
Principal Component 2 (18%)
Principal Component 1 (46%)
−6
−4
−2
0
2
4
−3
−2
−1
0
1
2
3

D

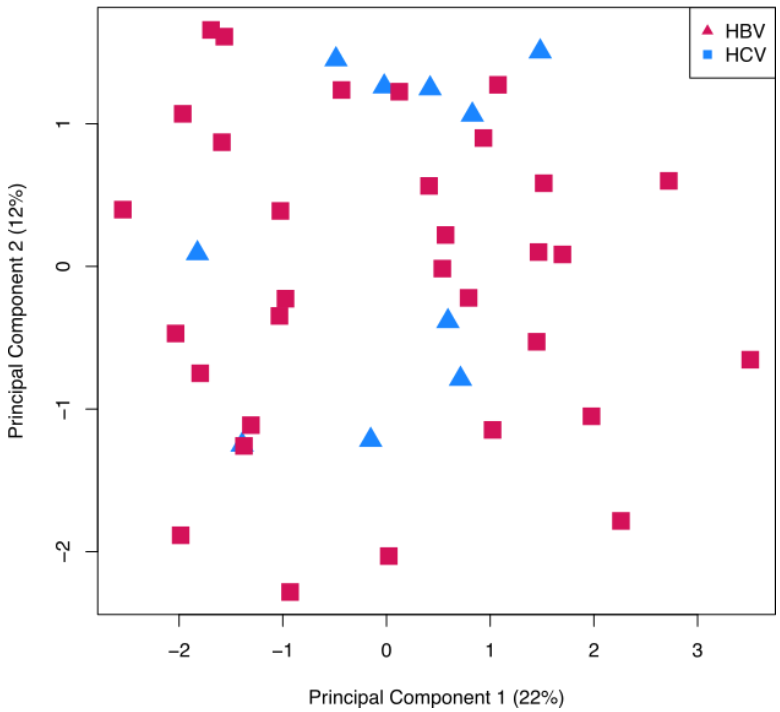

HBV
HCV
Principal Component 2 (12%)
Principal Component 1 (22%)
−2
−1
0
1
2
3
−2
−1
0
1

Figure 5. Male and Female tumor and tumor-adjacent MDS plots do not show a strong effect of etiology on sample variation.

Multidimensional scaling (MDS) plots on the top 50 differentially expressed genes within each subset. Pink triangles represent liver tissue samples with Hepatitis B virus and blue squares represent liver tissue samples with Hepatitis C virus. (A) Male tumor samples, (B) male tumor-adjacent samples, (C) female tumor samples, (D) female tumor-adjacent samples.

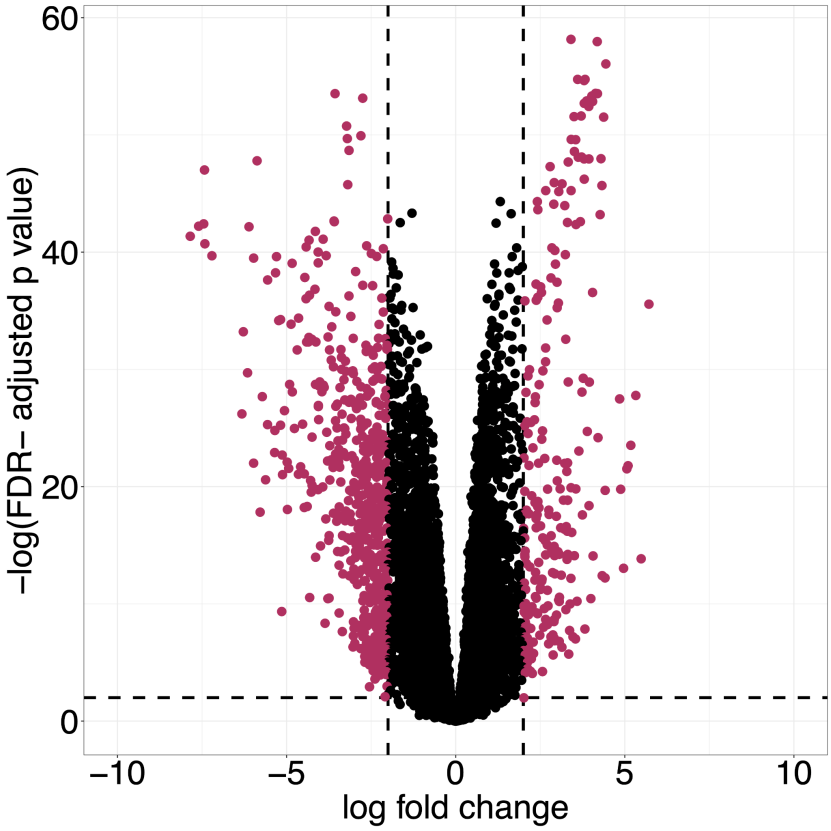


Figure 6. Viral- etiology specific genes identified in HBV tumor tumor-adjacent comparison

Here we visually represent the differentially expressed genes in HBV tumor tumor-adjacent data. 586 genes were downregulated in tumors and 238 upregulated in tumors The dashed lines indicate the levels of significance with the vertical lines representing an absolute logFC ≥ 2 and the horizontal line representing a p-value of 0.05. Maroon dots indicate genes with an

absolute value of logFC ≥ 2 and an adjusted p-value < 0.05 and are considered differentially expressed.

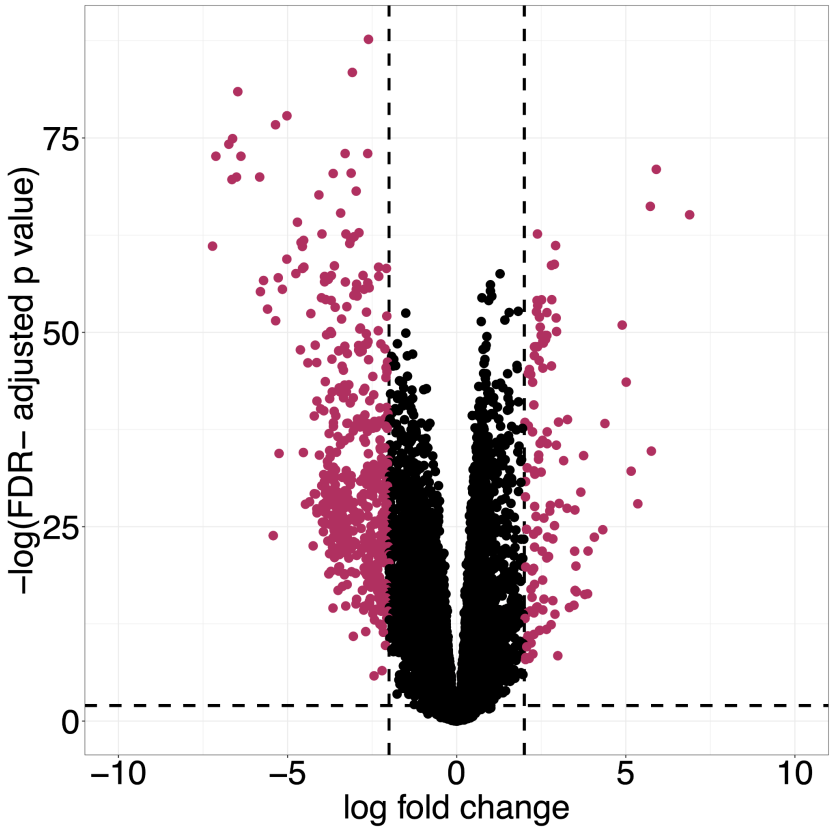


Figure 7.Viral- etiology specific genes identified in HCV tumor tumor-adjacent comparison

Here we visually represent the differentially expressed genes in HCV tumor tumor-adjacent data. 514 genes were downregulated in tumors and 124 upregulated in tumors The dashed lines indicate the levels of significance with the vertical lines representing an absolute logFC ≥ 2 and the horizontal line representing a p-value of 0.05. Maroon dots indicate genes with an absolute value of logFC ≥ 2 and an adjusted p-value < 0.05 and are considered differentially expressed.

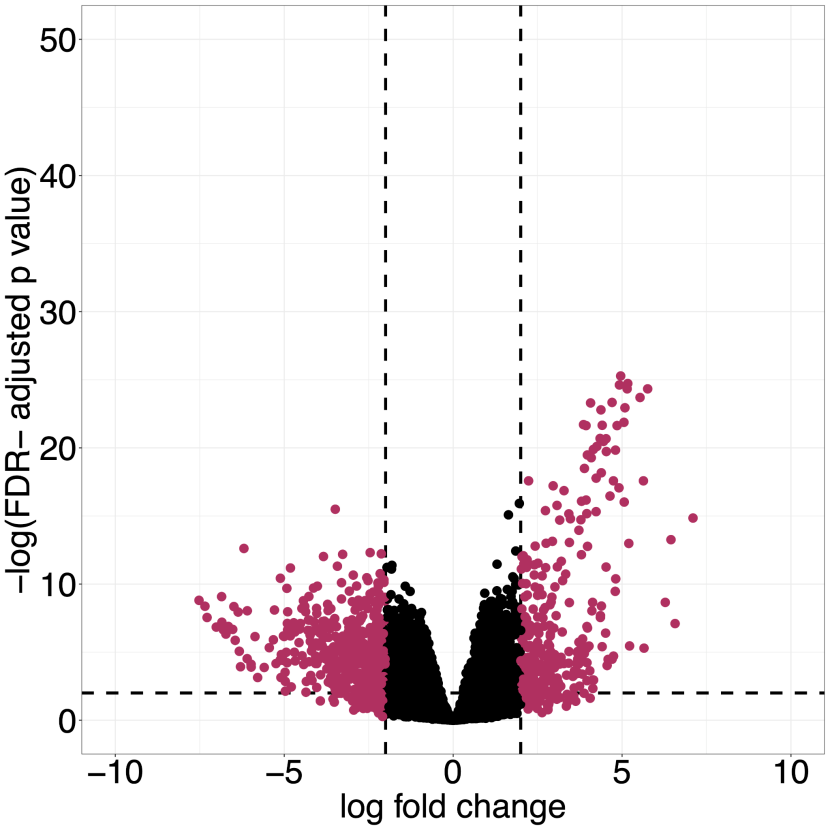


Figure 8. Sex and viral etiology specific genes identified in Female HBV tumor tumor-adjacent comparison.

Here we visually represent the differentially expressed genes in female HBV tumor tumor-adjacent data. 543 genes were downregulated in tumors and 305 upregulated in tumors. The dashed lines indicate the levels of significance with the vertical lines representing an absolute logFC ≥ 2 and the horizontal line representing a p-value of 0.05. Maroon dots indicate genes with an absolute value of logFC ≥ 2 and an adjusted p-value < 0.05 and are considered differentially expressed.

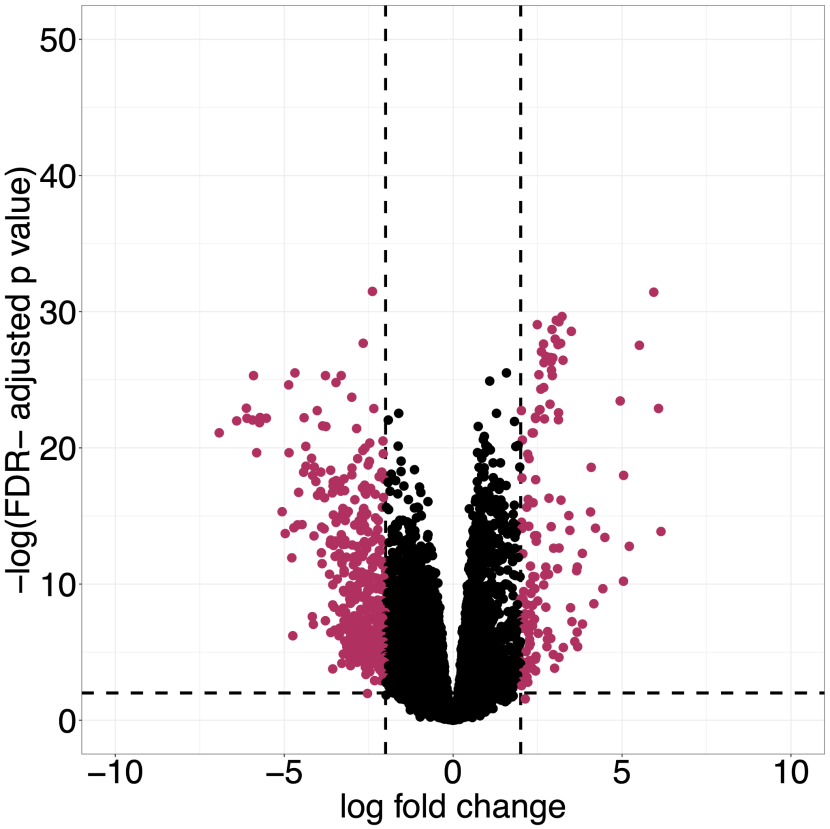


Figure 9. Sex and viral etiology specific genes identified in Female HCV tumor tumor-adjacent comparison.

Here we visually represent the differentially expressed genes in female HCV tumor tumor-adjacent data. 435 genes were downregulated in tumors and 136 upregulated in tumors The dashed lines indicate the levels of significance with the vertical lines representing an absolute logFC ≥ 2 and the horizontal line representing a p-value of 0.05. Maroon dots indicate genes with an absolute value of logFC ≥ 2 and an adjusted p-value < 0.05 and are considered differentially expressed.

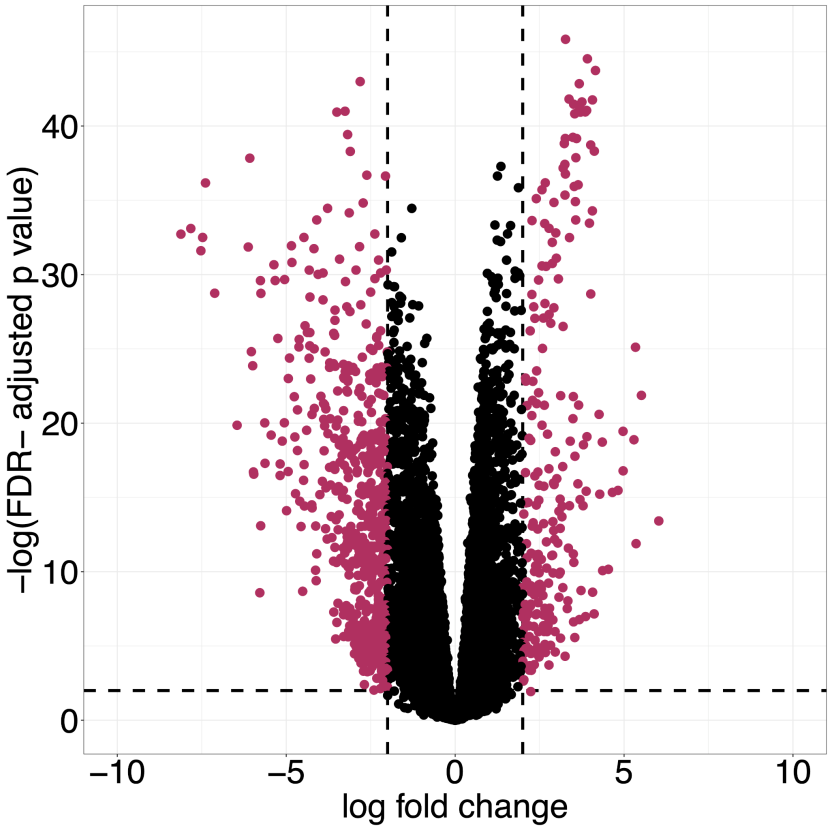


Figure 10. Sex and viral etiology specific genes identified in Male HBV tumor tumor-adjacent comparison.

Here we visually represent the differentially expressed genes in male HBV tumor tumor-adjacent data. 612 genes were downregulated in tumors and 232 upregulated in tumors. The dashed lines indicate the levels of significance with the vertical lines representing an absolute logFC ≥ 2 and the horizontal line representing a p-value of 0.05. Maroon dots indicate genes with an absolute value of logFC ≥ 2 and an adjusted p-value < 0.05 and are considered differentially expressed.

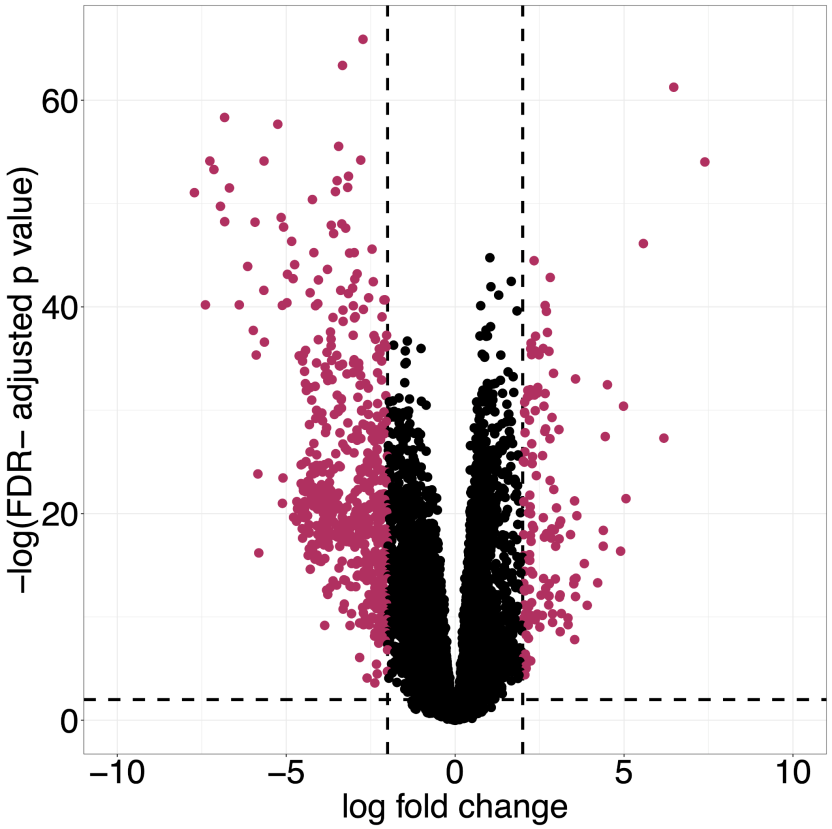


Figure 11. Sex and viral etiology specific genes identified in Male HCV tumor tumor-adjacent comparison.

Here we visually represent the differentially expressed genes in male tumor tumor-adjacent data. 578 genes were downregulated in tumors and 140 upregulated in tumors. The dashed lines indicate the levels of significance with the vertical lines representing an absolute logFC ≥ 2 and the horizontal line representing a p-value of 0.05. Maroon dots indicate genes with an absolute value of logFC ≥ 2 and an adjusted p-value < 0.05 and are considered differentially expressed.

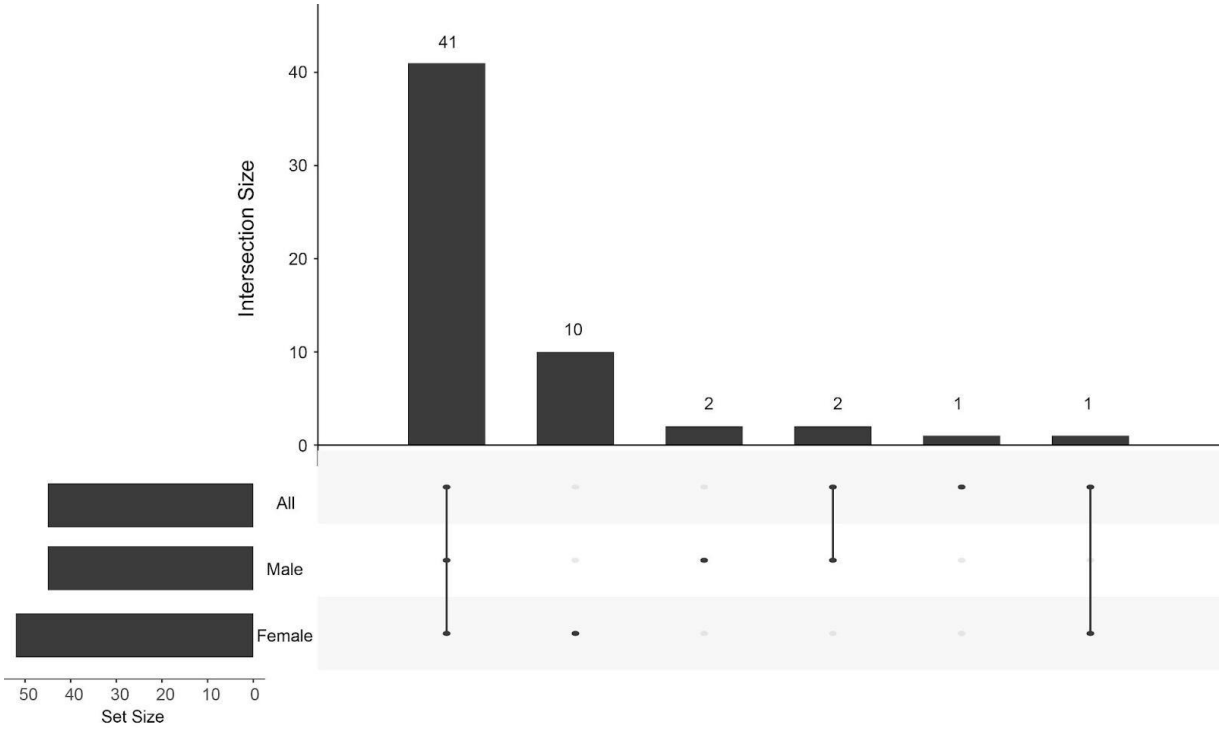


Figure 12. Enriched pathways identified in male and female differentially expressed genes

Here we visualize the pathway enrichment analysis results done using Reactome on the overall tumor vs. tumor-adjacent, male tumor vs. tumor-adjacent and female tumor vs. tumor-adjacent datasets through an upset plot. The vertical bars represent the number of pathways. The horizontal bars represent “Set Size” which is the number of enriched pathways in each comparison. The horizontal bars represent the number of samples in the dataset, where “Male” corresponds to the male tumor vs. tumor-adjacent dataset, “Female” corresponds to the female tumor vs. tumor-adjacent dataset and “all” corresponds to the overall tumor vs. tumor-adjacent data. The overall tumor vs. tumor-adjacent dataset does not include sex as a covariate. The dot means the dataset is included and the lines represent an intersection. The upset plot shows 41 pathways shared between all datasets, ten are unique to females, and two are unique to males. Two are shared between the males and overall tumor tumor-adjacent, one is unique to overall tumor tumor-adjacent, and one is shared between overall tumor tumor-adjacent, and females.

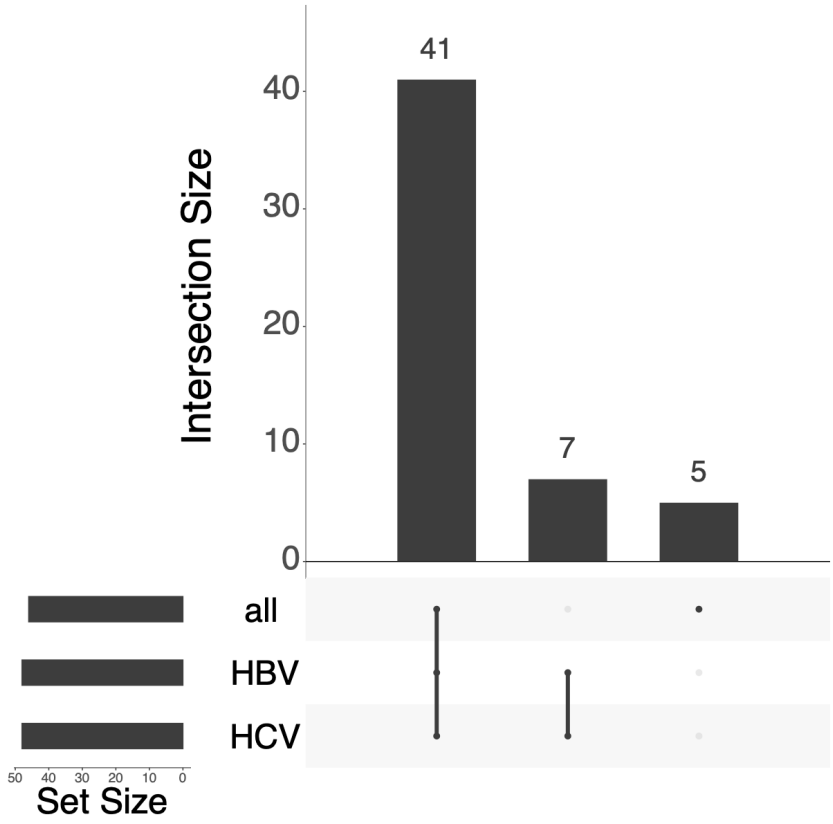


Figure 13. Enriched pathways identified in etiology specific differentially expressed genes.

Here we visualize the pathway enrichment analysis results done using Reactome on the overall tumor vs. tumor-adjacent, liver tissue infected with Hepatitis B (HBV) tumor vs. tumor-adjacent and liver tissue infected with Hepatitis C (HCV) tumor vs. tumor-adjacent datasets through an upset plot. The vertical bars with the axis “Intersection Size” represent the number of pathways. The horizontal bars with the axis label “Set Size” represent the number of samples in the dataset, where “HBV” corresponds to the HBV tumor vs. tumor-adjacent dataset, “HCV” corresponds to the HCV tumor vs. tumor-adjacent dataset and “all” corresponds to the overall tumor vs. tumor-adjacent data. The overall tumor vs. tumor-adjacent dataset includes sex as a covariate. The dot means the dataset is included and the lines represent an intersection. The upset plot shows 41 pathways shared between all datasets, seven are shared between HBV and HCV datasets, and five are unique to overall tumor tumor-adjacent dataset.

Tables

Table 1. Sample distribution of RNA-seq data.

The total number of samples includes 261 male samples and 93 female samples. The word tumor means a sample was taken from an HCC liver tumor. Tumor-adjacent means a sample was taken from healthy liver tissue. HBV means the liver tissue was infected with Hepatitis B. HCV means the liver tissue was infected with Hepatitis C . The row title “HBV & HCV” means the liver tissue was infected with both HBV and HCV. The “No hepatitis” and “HBV & HCV” samples were not used in analyses.

	Male Tumor	**Male Tumor-adjacent**	**Female Tumor**	**Female Tumor-Adjacent**
No hepatitis	25	25	3	3
HBV	33	40	8	9
HCV	59	72	34	36
HBV & HCV	4	4	0	0
Total	**121**	**140**	**8**	**9**

Table 2. Pathway table shared by all three tumor tumor-adjacent comparisons

These 41 pathways are shared by overall tumor tumor-adjacent, male tumor tumor-adjacent and female tumor tumor-adjacent differentially expressed genes.

Immune system*			
Innate immune system			
Complement cascade		Classical antibody-mediated complement activation	
		Initial triggering of complement	
		Regulation of complement cascade	
		Creation of C4 and C2 activators	
Fc epsilon receptor (FCERI) signaling		FCERI mediated MAPK activation	
		FCERI mediated Ca2+ mobilization	
		Role of LAT2/NTAL/LAB on calcium mobilization	
		FCERI mediated NF-kB activation	
Fcgamma receptor (FCGR) dependent phagocytosis		FCGR activation	
		Role of phospholipids in phagocytosis	
		Regulation of actin dynamics for phagocytic cup formation	
Adaptive immune system			
Signaling by the B Cell Receptor (BCR)		CD22 mediated BCR regulation	
		Antigen activates B Cell Receptor (BCR) leading to generation of second messengers	
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell			
Vesicle-mediated transport			
Binding and Uptake of Ligands by Scavenger Receptors		Scavenging of heme from plasma	
Hemostasis*			
Cell surface interactions at the vascular wall			
Cellular response to stimuli*			
Response to metal ions		Metallothioneins bind metals	
Oncogene Induced Senescence			
Disease*			
Infectious Disease*			
Leishmania infection	Leishmania parasite growth and survival	Anti-inflammatory response favouring Leishmania parasite infection	FCGRA-mediated IL 10 synthesis

	Parasite infection	Leishmania phagocytosis	FCGR3A-mediated phagocytosis
SARS-CoV Infections		Potential therapeutics for SARS	
Cell Cycle*			
Cell Cycle, Mitotic*			
Mitotic G1 phase and G1/S transition		G0 and Early G1	Transcription of E2F targets under negative control by p107 (RBL1) and p130 (RBL2) in complex with HDAC1
		G1/S Transition*	G1/S-Specific Transcription
Mitotic G2-G2/M phases*		G2/M Transition*	Polo-like kinase mediated events

* Added for grouping purposes not significant

Table 3. Multiple cell cycle pathways enriched in only female tumor tumor-adjacent differentially expressed genes.

Cell Cycle			
Cell Cycle Mitotic			
Regulation of mitotic cell cycle*	APC/C-mediated degradation of cell cycle proteins*	Regulation of APC/C activators between G1/S and early anaphase*	Phosphorylation and Emi1
M Phase*	Mitotic Prophase*	Nuclear Envelope Breakdown*	Activation of NIMA Kinases NEK9, NEK6, NEK7
	Mitotic Prometaphase*	Resolution of Sister Chromatid Cohesion	
Drug ADME*			
Aspirin ADME			
Signal Transduction*			
Signaling by GPCR*			
GPCR ligand binding*	Class A/1 Rhodopsin-like receptor*	Amine ligand-binding receptor*	Adrenoceptors
Metabolism*			
Metabolism of ligands*			
Fatty acid metabolism*	Arachidonic acid metabolism*	Synthesis of (16-20)-hydroxyeicosatetraenoic acids (HETE)	
Biological Oxidations*			
Phase I - Functionalization of compounds*	Cytochrome P450 - arranged by substrate type*	Xenobiotics	CYP2E1 reactions

* Added for grouping purposes not significant

Table 4. Pathways enriched in only male tumor tumor-adjacent differentially expressed genes and shared with overall tumor tumor-adjacent comparison.

Transport of small molecules*		
Plasma lipoprotein assembly, remodeling, and clearance*		
Plasma lipoprotein remodeling		
Immune System*		
Cytokine Signaling in Immune system		
Signaling by Interleukins*	Interleukin-1 family signaling*	Interleukin-33 signaling
	Interleukin-10 signaling**	
Innate Immune System		
Complement Cascade*	Terminal pathway of complement**	

* Added for grouping purposes not significant

** Pathways shared by overall tumor tumor-adjacent genes and male tumor tumor-adjacent genes

Table 5. Pathways enriched in HBV and HCV tumor tumor-adjacent differentially expressed genes.

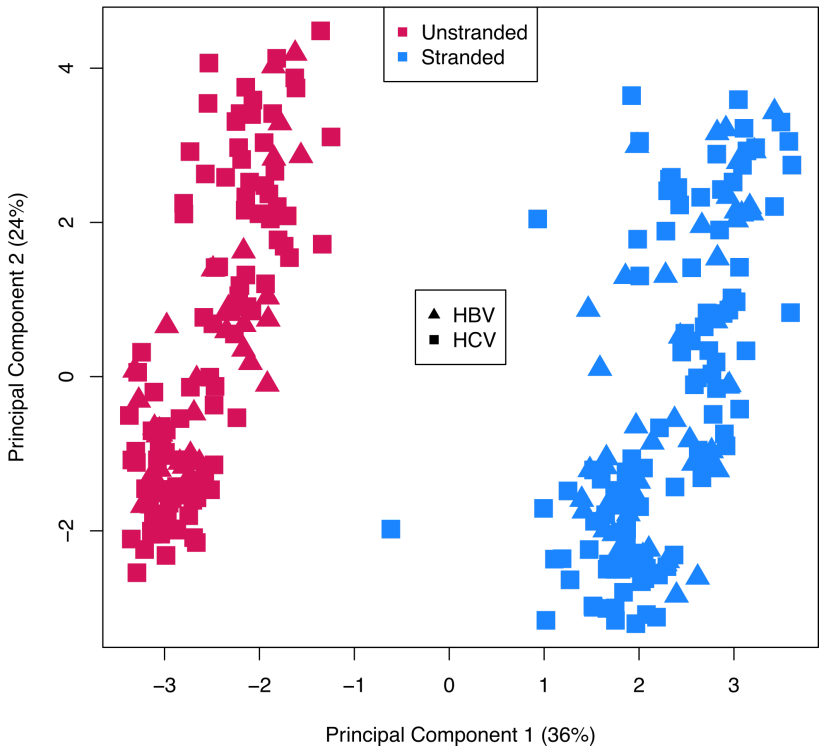
Drug ADME			
Aspirin ADME			
Metabolism*			
Metabolism of lipids*			
Fatty acid metabolism*	Arachidonic acid metabolism*	Synthesis of (16-20)-hydroxyeicosatetraenoic acids (HETE)	
Cell Cycle*			
Cell Cycle, Mitotic			
Regulation of mitotic cell cycle*	APC/C-mediated degradation of cell cycle proteins*	Regulation of APC/C activators between G1/S and early anaphase*	Phosphorylation and Emi1
Cell Cycle Checkpoints*			
G2/M Checkpoints*		G2/M DNA replication checkpoints	
Transport of small molecules*			
Plasma lipoprotein assembly, remodeling, and clearance*			
Plasma lipoprotein remodeling			

* Added for grouping purposes not significant

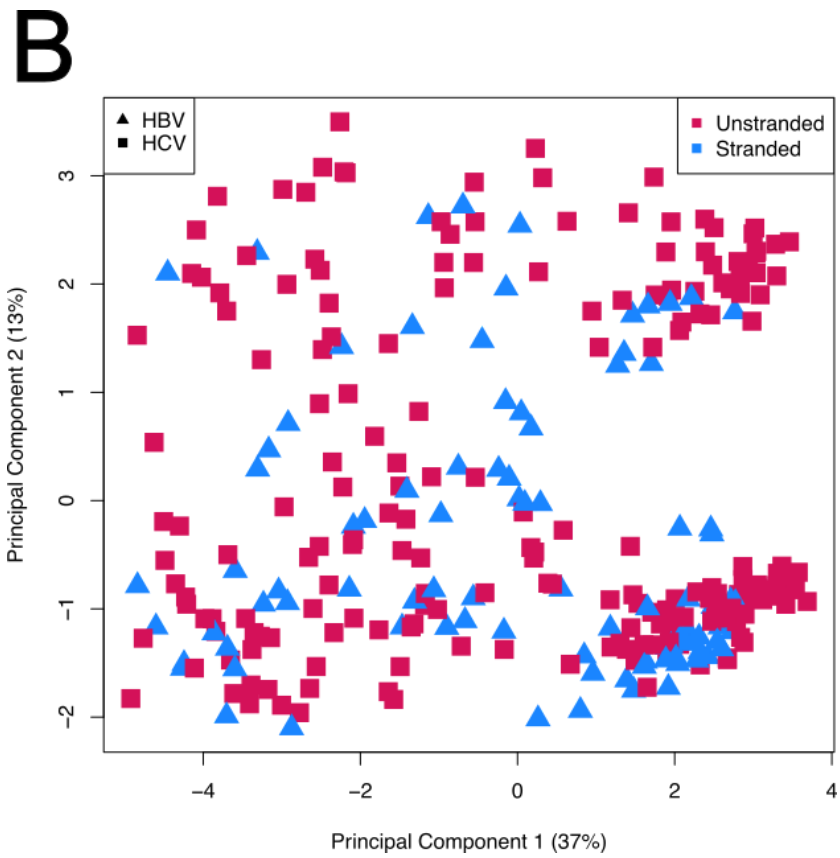
Supplementary Material

Supplementary Figures

A

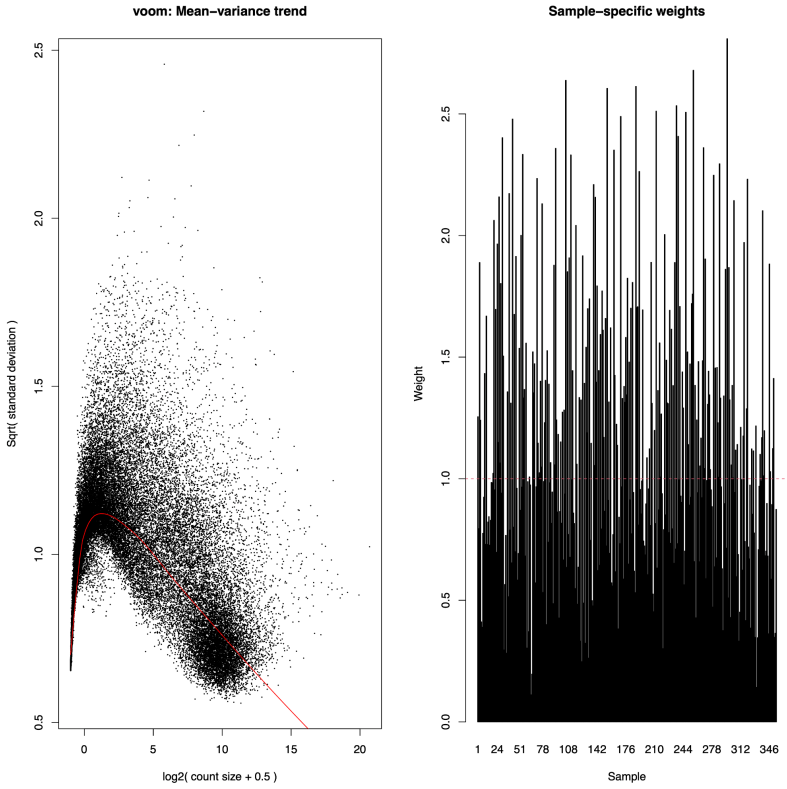

Unstranded
Stranded
HBV
HCV
Principal Component 2 (24%)
Principal Component 1 (36%)
4
2
0
−2
−3
−2
−1
0
1
2
3

B


HBV
HCV
Unstranded
Stranded
Principal Component 2 (13%)
Principal Component 1 (37%)
3
2
1
0
−1
−2
−4
−2
0
2
4

Supplementary Figure 1. Library strandedness batch effect among samples

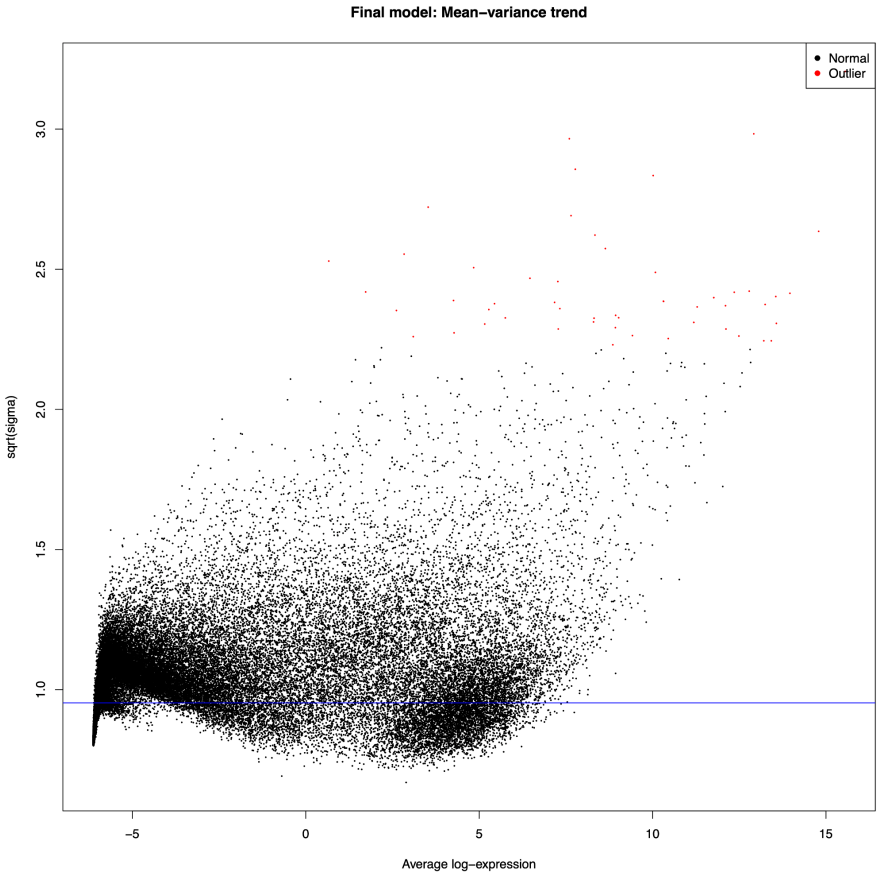
A batch effect noted in our sample was library type since the dataset had both stranded and unstranded RNA in the samples. A multidimensional scaling (MDS) analysis of the top 50 differentially expressed genes was performed to confirm this batch effect. MDS plot A shows the distribution of the overall tumor tumor-adj. samples without library strandness accounted for. Principal component 1 in plot A is attributable to library type and accounts 44% of the difference. MDS plot B shows the distribution of the samples with the library strandness accounted for. Both plots are colored to represent unstranded and stranded library types with pink representing unstranded and blue representing stranded. The triangle shapes represent liver tissue samples infected with Hepatitis B and the square shapes represent liver tissue samples infected with Hepatitis C



Supplementary Figure 2. Differential expression analysis on the overall tumor tumor-adjacent samples using limma/voom R package.

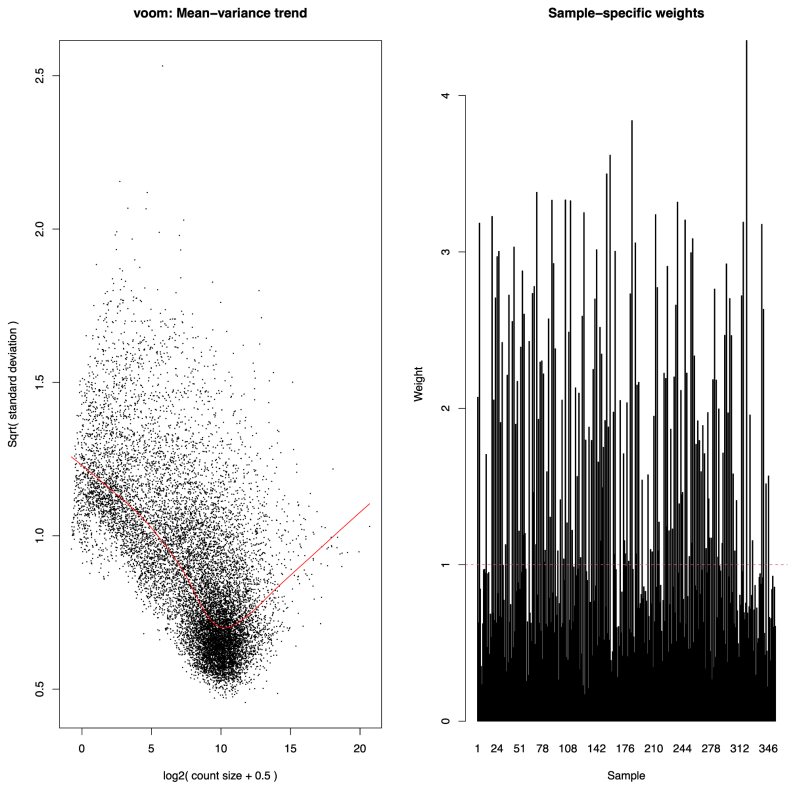
The design matrix for this analysis had lesion type as a predictor variable and library type as a covariate. Parameters for the differential expression analysis include log fold change (logFC) ≥ 2 and p value of < 0.05. Left graph is showing the distribution of the log2 normalized counts by their residuals which is the square root of the standard deviation. The red line is showing the average expression values which are used to obtain the sample weights. The right graph is

using the weight values given to each sample. These values along with the log2 counts per million are passed into the limma model.



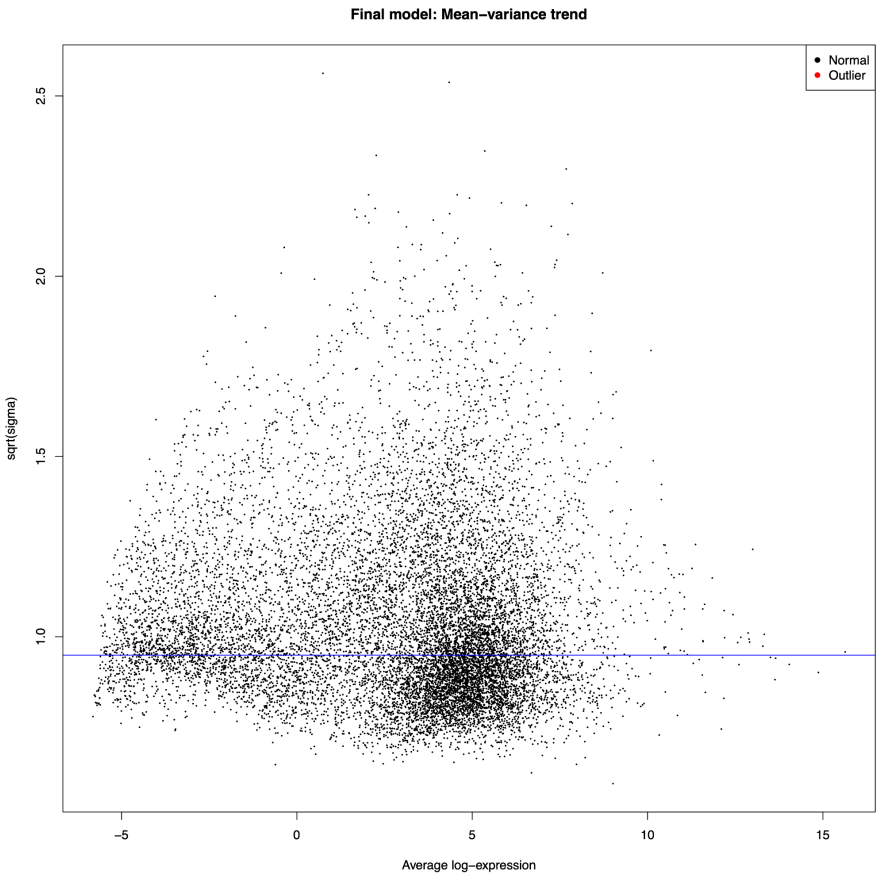
Supplementary Figure 3. Final linear model for differential expression analysis on overall tumor tumor-adjacent sample.

The graph is showing the plot of average log expression values by the square root of the standard deviation. The blue line represents the fitted curve generated by the limma pipeline. The red dots represent average log expression values that are outliers.



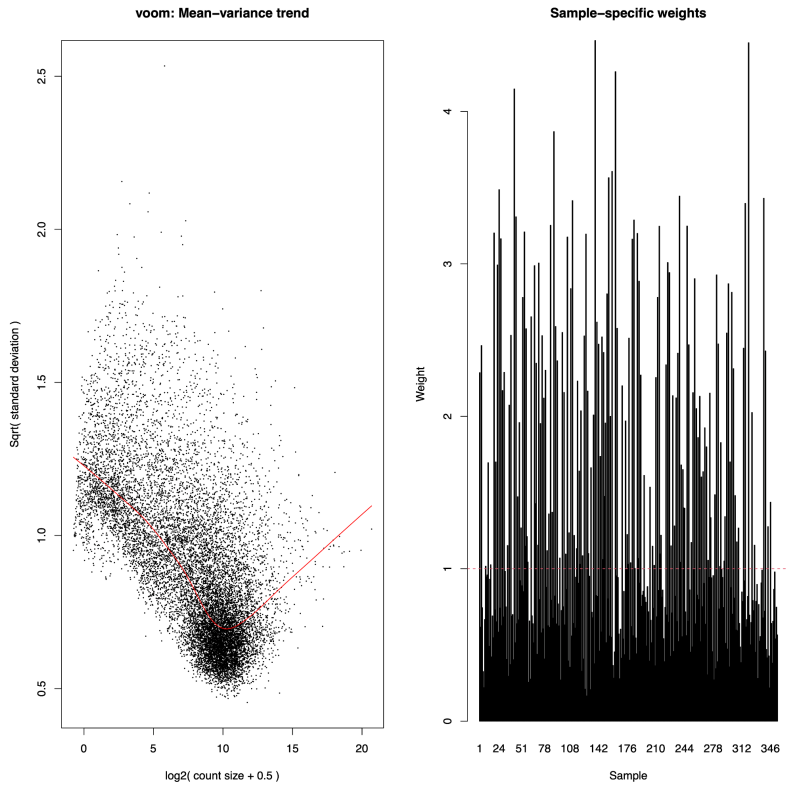
Supplementary Figure 4. Differential expression analysis on the tumor tumor-adjacent samples differentiated by sex using limma/voom R package.

The first design matrix for this analysis design matrix has lesion type as a predictor variable with library as a covariate The second design matrix has lesion type differentiated by sex as a predictor variable and library type as a covariate. Parameters for the differential expression analysis include log fold change (logFC) ≥ 2 and p value of < 0.05. Left graph is showing the distribution of the log2 normalized counts by their residuals which is the square root of the standard deviation. The red line is showing the average expression values which are used to obtain the sample weights. The right graph is using the weight values given to each sample. These values along with the log2 counts per million are passed into the limma model.



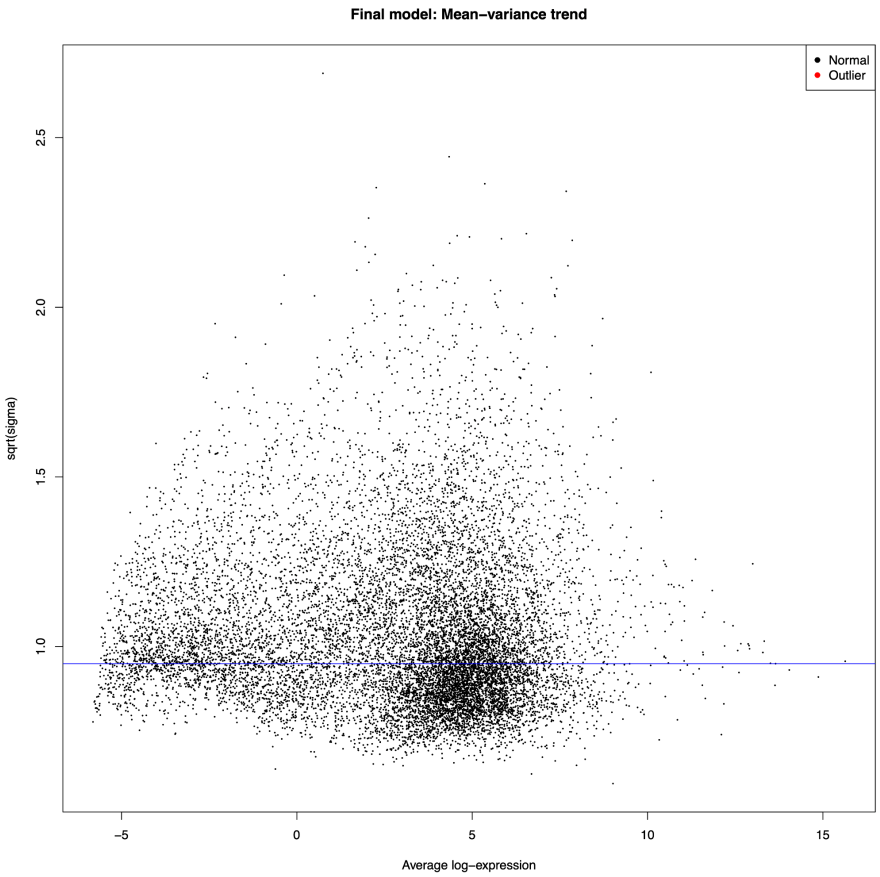
Supplementary Figure 5. Final linear model for differential expression analysis on tumor tumor-adjacent samples differentiated by sex.

The graph is showing the plot of average log expression values by the square root of the standard deviation. The blue line represents the fitted curve generated by the limma pipeline. There are no outliers in this final model.



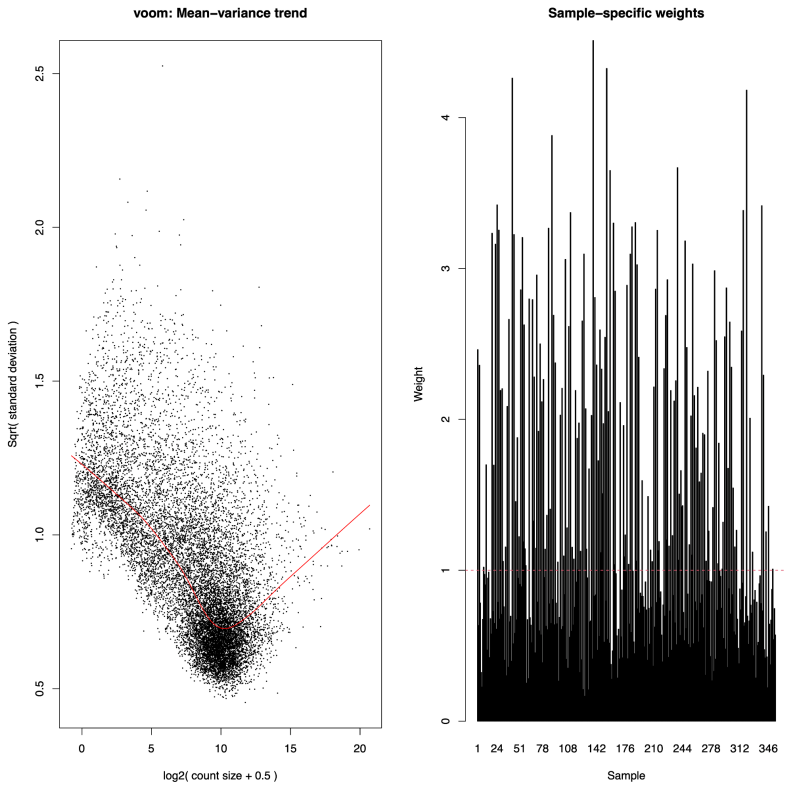
Supplementary Figure 6. Differential expression analysis on the tumor tumor-adjacent samples differentiated by sex and etiology using limma/voom R package.

The first design matrix for this analysis has lesion type as a predictor variable and library type and sex as covariates. The second design matrix for this analysis has lesion type differentiated by sex and etiology as a predictor variable and library type as a covariate. Parameters for the differential expression analysis include log fold change (logFC) ≥ 2 and p value of < 0.05. Left graph is showing the distribution of the log2 normalized counts by their residuals which is the square root of the standard deviation. The red line is showing the average expression values which are used to obtain the sample weights. The right graph is using the weight values given to each sample. These values along with the log2 counts per million are passed into the limma model.



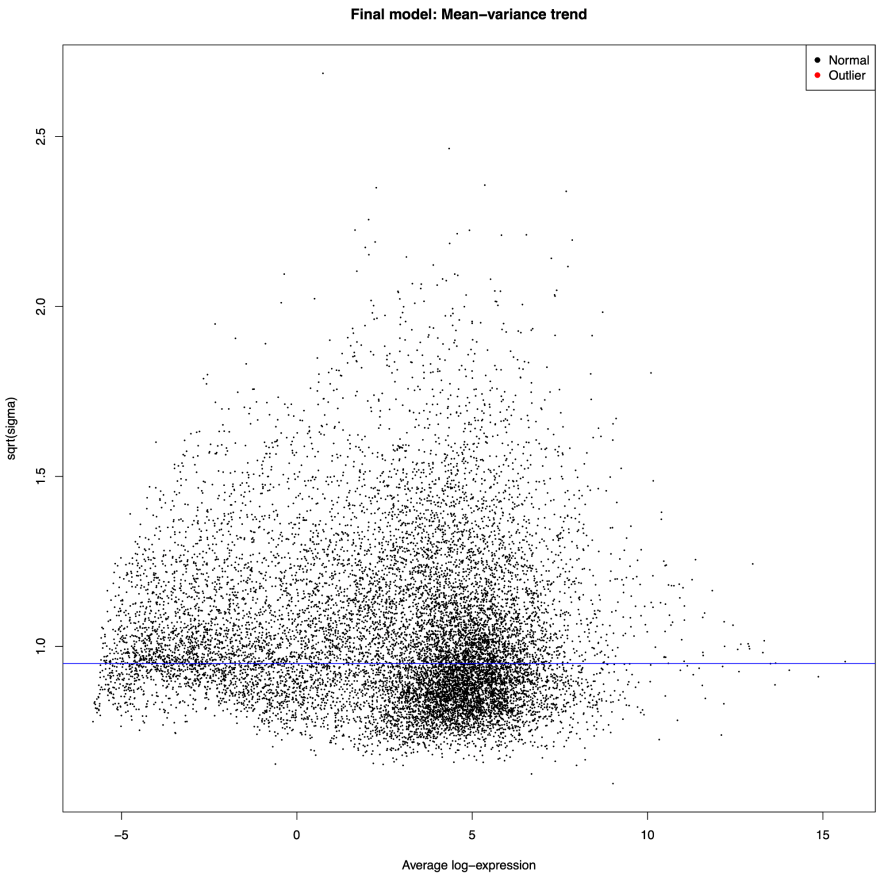
Supplementary Figures 7. Final linear model for differential expression analysis on tumor tumor-adjacent samples differentiated by sex and etiology

The graph is showing the plot of average log expression values by the square root of the standard deviation. The blue line represents the fitted curve generated by the limma pipeline. There are no outliers in this final model.



Supplementary Figures 8. Differential expression analysis on the tumor tumor-adjacent samples differentiated by etiology using limma/voom R package.

The first design matrix for this analysis has lesion type as a predictor variable and library type and sex as covariates. The second design matrix for this analysis has lesion type differentiated by sex and etiology as a predictor variable and library type and sex as covariates. Parameters for the differential expression analysis include log fold change (logFC) ≥ 2 and p value of < 0.05. Left graph is showing the distribution of the log2 normalized counts by their residuals which is the square root of the standard deviation. The red line is showing the average expression values which are used to obtain the sample weights. The right graph is using the weight values given to each sample. These values along with the log2 counts per million are passed into the limma model.



Supplementary Figures 9. Final linear model for differential expression analysis on tumor tumor-adjacent samples differentiated by etiology.

The graph is showing the plot of average log expression values by the square root of the standard deviation. The blue line represents the fitted curve generated by the limma pipeline. There are no outliers in this final model.

Supplementary Tables

Supplementary Notes

Acknowledgements

I would like to thank the members of Dr. Wilson's Sex Chromosome Lab for their useful discussion and contribution to this body of work. I would like to thank my thesis committee, Elizabeth Borden, my advisor Dr. Melissa Wilson, and Dr. Kenneth Buetow, for their mentorship and guidance on this project. I would like to thank my friends and family members for their support throughout this process.

References

Altekruse, Sean F., Susan S. Devesa, Lois A. Dickie, Katherine A. McGlynn, and David E. Kleiner. 2011. “Histological Classification of Liver and Intrahepatic Bile Duct Cancers in SEER Registries.” *Journal of Registry Management* 38 (4): 201–5.

“Babraham Bioinformatics - FastQC A Quality Control Tool for High Throughput Sequence Data.” n.d. Accessed March 17, 2023. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bolger, Anthony M., Marc Lohse, and Bjoern Usadel. 2014. “Trimmomatic: A Flexible Trimmer for Illumina Sequence Data.” *Bioinformatics* 30 (15): 2114–20.

Chhikara, Bhupender S., and Keykavous Parang. 2023. “Global Cancer Statistics 2022: The Trends Projection Analysis.” *Chemical Biology Letters* 10 (1): 451–451.

Conway, Jake R., Alexander Lex, and Nils Gehlenborg. 2017. “UpSetR: An R Package for the Visualization of Intersecting Sets and Their Properties.” *Bioinformatics* 33 (18): 2938–40.

“EEF1A2 Eukaryotic Translation Elongation Factor 1 Alpha 2 [Homo Sapiens (human)] - Gene - NCBI.” n.d. Accessed March 16, 2023. https://www.ncbi.nlm.nih.gov/gene/1917.

El-Serag, Hashem B. 2012. “Epidemiology of Viral Hepatitis and Hepatocellular Carcinoma.” *Gastroenterology* 142 (6): 1264–73.e1.

Fabregat, Antonio, Konstantinos Sidiropoulos, Phani Garapati, Marc Gillespie, Kerstin Hausmann, Robin Haw, Bijay Jassal, et al. 2016. “The Reactome Pathway Knowledgebase.” *Nucleic Acids Research* 44 (D1): D481–87.

Gamaev, Lika, Lina Mizrahi, Tomer Friehmann, Nofar Rosenberg, Orit Pappo, Devorah Olam, Evelyne Zeira, et al. 2021. “The pro-Oncogenic Effect of the lncRNA H19 in the Development of Chronic Inflammation-Mediated Hepatocellular Carcinoma.” *Oncogene* 40 (1): 127–39.

“GENCODE - Human Release 25.” n.d. Accessed March 18, 2023. https://www.gencodegenes.org/human/release_25.html.

Ghafouri-Fard, Soudeh, Mohammadhosein Esmaeili, and Mohammad Taheri. 2020. “H19 lncRNA: Roles in Tumorigenesis.” *Biomedicine & Pharmacotherapy = Biomedecine & Pharmacotherapie* 123 (March): 109774.

Guy, Jennifer, and Marion G. Peters. 2013. “Liver Disease in Women: The Influence of Gender on Epidemiology, Natural History, and Patient Outcomes.” *Gastroenterology & Hepatology* 9 (10): 633–39.

Law, Charity W., Yunshun Chen, Wei Shi, and Gordon K. Smyth. 2014. “Voom: Precision Weights Unlock Linear Model Analysis Tools for RNA-Seq Read Counts.” *Genome Biology* 15 (2): R29.

Liao, Yang, Gordon K. Smyth, and Wei Shi. 2014. “featureCounts: An Efficient General Purpose Program for Assigning Sequence Reads to Genomic Features.” *Bioinformatics* 30 (7): 923–30.

McGlynn, Katherine A., Jessica L. Petrick, and Hashem B. El-Serag. 2021. “Epidemiology of Hepatocellular Carcinoma.” *Hepatology* 73 Suppl 1 (Suppl 1): 4–13.

Moshe, Yakir, Ortal Bar-On, Dvora Ganoth, and Avram Hershko. 2011. “Regulation of the Action of Early Mitotic Inhibitor 1 on the Anaphase-Promoting Complex/cyclosome by Cyclin-Dependent Kinases.” *The Journal of Biological Chemistry* 286 (19): 16647–57.

Natri, Heini M., Melissa A. Wilson, and Kenneth H. Buetow. 2019. “Distinct Molecular Etiologies of Male and Female Hepatocellular Carcinoma.” *BMC Cancer* 19 (1): 951.

Qiu, Fu-Nan, Yi Huang, Dun-Yan Chen, Feng Li, Yan-An Wu, Wen-Bing Wu, and Xiao-Li Huang. 2016. “Eukaryotic Elongation Factor-1α 2 Knockdown Inhibits Hepatocarcinogenesis by Suppressing PI3K/Akt/NF-κB Signaling.” *World Journal of Gastroenterology: WJG* 22 (16): 4226–37.

Rico Montanari, N., Anugwom, C. M., Boonstra, A., & Debes, J. D. (2021). The role of cytokines in the different stages of hepatocellular carcinoma. *Cancers*, *13*(19), 4876.

Ritchie, Matthew E., Belinda Phipson, Di Wu, Yifang Hu, Charity W. Law, Wei Shi, and Gordon K. Smyth. 2015. “Limma Powers Differential Expression Analyses for RNA-Sequencing and Microarray Studies.” *Nucleic Acids Research* 43 (7): e47.

Robinson, Mark D., Davis J. McCarthy, and Gordon K. Smyth. 2010. “edgeR: A Bioconductor Package for Differential Expression Analysis of Digital Gene Expression Data.” *Bioinformatics* 26 (1): 139–40.

Vescovo, T., G. Refolo, G. Vitagliano, G. M. Fimia, and M. Piacentini. 2016. “Molecular Mechanisms of Hepatitis C Virus-Induced Hepatocellular Carcinoma.” *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases* 22 (10): 853–61.

Yan, Hongxian, Zhaohui Li, Quan Shen, Qian Wang, Jianguo Tian, Qingfeng Jiang, and Linbo Gao. 2017. “Aberrant Expression of Cell Cycle and Material Metabolism Related Genes Contributes to Hepatocellular Carcinoma Occurrence.” *Pathology, Research and Practice* 213 (4): 316–21.

Yoshimizu, Tomomi, Audrey Miroglio, Marie-Anne Ripoche, Anne Gabory, Maria Vernucci, Andrea Riccio, Sabine Colnot, et al. 2008. “The H19 Locus Acts in Vivo as a Tumor Suppressor.” *Proceedings of the National Academy of Sciences of the United States of America* 105 (34): 12417–22.

Yuan, Yuan, Lingxiang Liu, Hu Chen, Yumeng Wang, Yanxun Xu, Huzhang Mao, Jun Li, et al. 2016. “Comprehensive Characterization of Molecular Differences in Cancer between Male and Female Patients.” *Cancer Cell* 29 (5): 711–22.

Zhang, Junjun, Rosita Bajari, Dusan Andric, Francois Gerthoffert, Alexandru Lepsa, Hardeep Nahal-Bose, Lincoln D. Stein, and Vincent Ferretti. 2019. “The International Cancer Genome Consortium Data Portal.” *Nature Biotechnology* 37 (4): 367–69.