Jonathan Le, Minwoo Cho, Echo Tang
4/8/2022
QBIO-490

Characterizing Molecular Differences Between Male and Female Colorectal Cancer Patients

Jonathan Le, Minwoo Cho, Echo Tang

Introduction

Colorectal cancer (CRC) is the second most common cause of cancer death and is responsible for over 52,980 deaths per year in the United States (SEER, 2021). Multiple studies have reported sex-related differences in colorectal cancer survival, showing that female CRC patients had significantly higher overall survival rates than male patients (Yang et al., 2017). Additionally, previous papers have found sex-based differences in gene and protein expression levels and mutation rates in genes like *EGFR* and *TP53*, with men tending to have worse prognosis and tumor growth (Chang et al., 2015 and Yang et al. 2017). Given the association between gene expression, mutation, and sex-differentiated survival in other cancers, it is plausible that such associations are present in CRC as well.

Our research project seeks to further investigate the molecular differences between male and female CRC patients in order to determine what factors may contribute to the different survival rates between the two populations. To conduct our analysis, we used publicly available data from both The Cancer Genome Atlas (TCGA) and the Clinical Proteomic Tumor Analysis Consortium (CPTAC). TCGA is a cancer genomics program that has collected molecular data of over 20,000 primary cancer datasets across 33 different cancer types, including genomics, epigenomics, transcriptomic data. CPTAC contains large-scale proteomic data in relation to clinical data found in TCGA; coupled with TCGA, it allows for a comprehensive investigation of

CRC at a multi-omic level. By analyzing the characteristics of CRC at a multi-omic level, we hope to gain a more holistic perspective of the disease.

Our findings indicated that there exists a significant difference between overall male and female patient survival, and many genes were found to be differentially expressed and correlated among the two sexes. In particular, we have identified *FUCA2*, *CDKN2A*, and *VSIG1* as possible druggable targets for female patients specifically, and *BPIFB1*, *THBS1*, *PCDH1*, and *CDKN2A* for male patients. Along with sex-differentiated survival data, the identification of sex-specific effects of these respective genes can lead to more targeted and nuanced therapeutics in the future.

Methods

This study implemented packages from R and Python for data analysis at a multi-omic level. CRC clinical and RNAseq data was accessed from TCGA using the R package TCGAbiolinks with the accession code "COAD," and proteomics data was accessed from CPTAC using the Python package 'cptac'.

To visualize the mutation characteristics of male and female patients, a MAF analysis was conducted using the 'maftools' R package. Male and female patients were separated into subsets, and a co-oncoplot was generated for the top 5 most mutated genes for each sex. To visualize the differences in gene expression levels of specific genes in males and females, a box plot was created for each of the top 5 mutated genes.

Transcriptomic analysis of RNA expression was then performed through DESeq2 analysis from the 'DESeq2' package in R. DESeq2 analysis was conducted to determine the most upregulated genes in both male and female patients. DESeq2 analysis was also conducted to determine the most differentially expressed genes between cancer tissue and normal tissue in

both male and female patients. Additionally, data from the DESeq2 analysis was also analyzed to determine if the most commonly mutated genes resulted in differential RNA expression between male and female patients.

The proteomic data from CPTAC for five of the ten most differentially expressed genes in both female and male patients obtained from the DESeq2 analysis was then used in tandem with the TCGA data to quantify the relationship between proteomics and transcriptomics data. The 'Matplotlib' and 'Seaborn' Python packages were used to quantify this relationship for those genes to make heatmaps of Spearman correlation coefficients.

Results

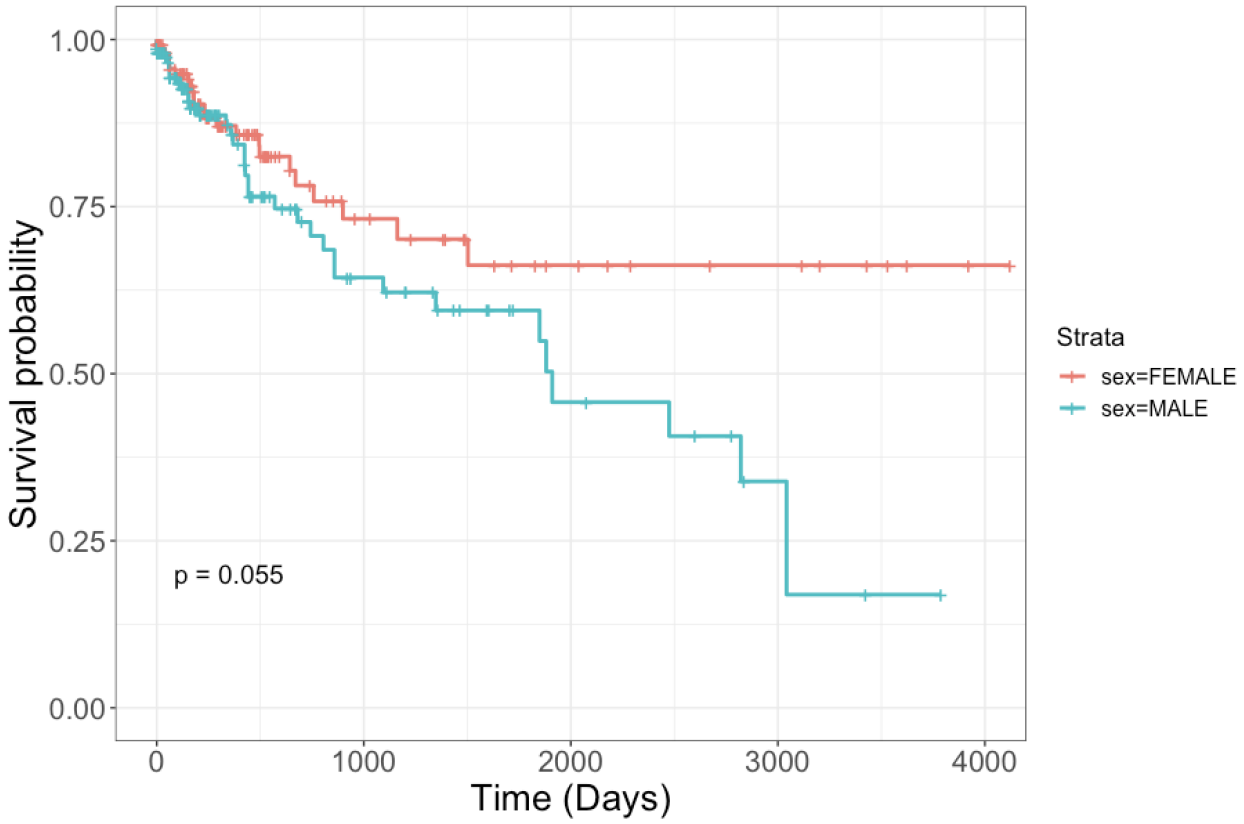

Figure 1. Survival plot comparing survival over time for male and female CRC patients. Female patients show better overall survival over time than male patients.

Figure 1 demonstrates that female CRC patients have better overall survival over time than male patients. The p-value of 0.055 is low, but on the borderline of significance with an alpha level of 0.05.

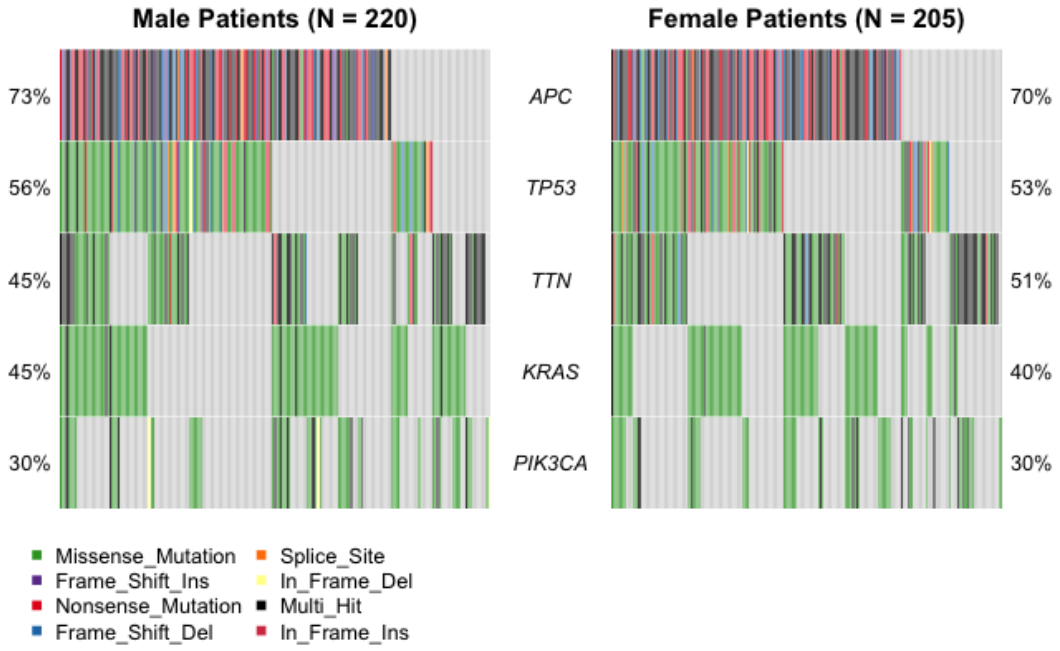

Figure 2. The mutation percentages and types of mutations are similar between sexes for each gene. Co-oncoplot depicting the top 5 most mutated genes in male and female patients with colorectal cancer. The different colors represent the different types of mutations present.

We first conducted a MAF analysis to determine if this survival difference was due to mutation differences. The five most mutated genes for both sexes were *APC*, *TP53*, *TTN*, *KRAS* and *PIK3CA*. These give genes were the same in both sexes, and there were only slight differences present in mutation percentages. There was a 3% difference between the most mutated gene *APC* with the males showing 73% of those genes mutated compared to 70% in the females. The mutation percentage difference between sexes were minimal with the biggest difference occurring in *TTN* with a 6% difference higher in females (Figure 2).

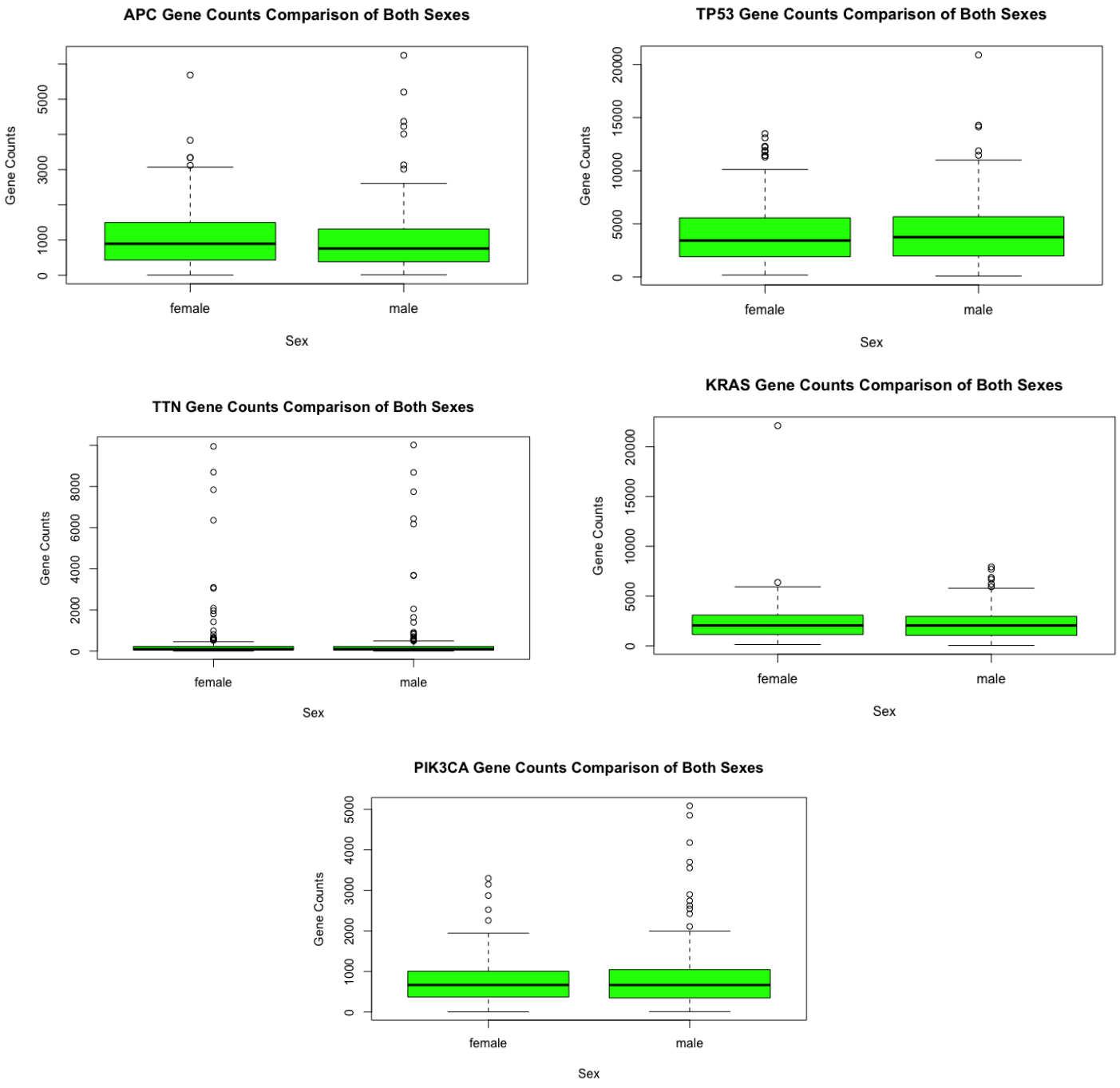

Figure 3. No significant differences in gene count averages were observed throughout the 5 most mutated genes. Boxplot of the top 5 most mutated gene counts (*APC, TP53, TTN, KRAS, PIK3CA*) in relation to sex.

Gene	log2FoldChange	padj
APC	-0.1361067	0.364089
TP53	-0.0631684	0.747449
TTN	0.1193335	0.662543
KRAS	-0.0860401	0.462877
PIK3CA	0.1028847	0.449482

Table 1. DESeq2 analysis of the top 5 most mutated genes in male and female CRC patients. Based on this table, none of the 5 genes show a $\log_2$ fold change of at least 1 or -1, which means that these genes are not significantly differentially expressed between male and female CRC patients. Additionally, the padj values are all greater than 0.05, which indicates non-significance in differential expression.

Differential expression analysis of the five most mutated genes in male and female CRC patients found in Figure 2 showed that there was no significant difference in expression of these genes between male and female patients. Each of the genes had a log2FoldChange between -0.13 and 0.11, which indicates that each of these genes are expressed at similar levels in both male and female CRC patients (Table 1). Additionally, the padj values are all greater than 0.05, which indicates non-significance in differential expression at a 0.05 significance level. Figure 4 shows that the top 5 most overexpressed genes in female CRC patients are *ZFY*, *AC010086.3*, *EIF1AY*, *TMSB4Y*, and *TXLNGY*, while the top 5 most overexpressed genes in male patients are *XIST*, *APOA4*, *MAGEB2*, *Z84478.1*, and *MS4A10*. However, most of these genes are sex-linked, and not directly linked to

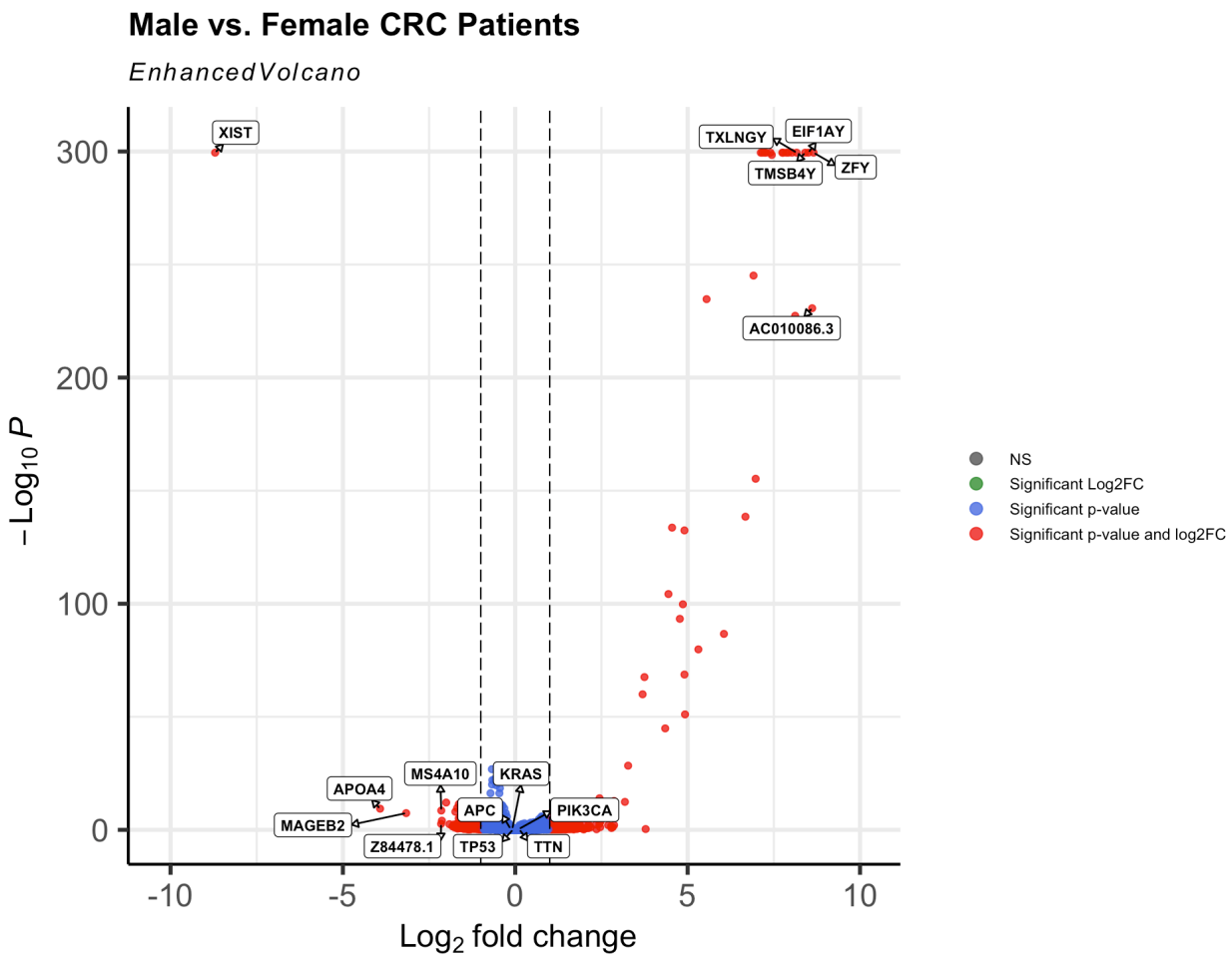

Figure 4. Volcano plot showing the differential expression of genes between male and female patients. The genes on the right side of the figure are overexpressed in male CRC patients while the genes on the left are those overexpressed in female patients. The five most differentially expressed genes for both male and female patients are labeled, as well as the five most mutated genes identified in Figure 1. We only considered genes with a significant difference in expression as those with a Log2FoldChange greater than or equal to 1 or less than or equal to -1, which would represent at least a 2x difference in expression. The p-value threshold is set at 0.05.

	Male Patients		Female Patients	
Gene Name	**log2FoldChange**	**padj**	**log2FoldChange**	**padj**
APC	-0.854426	4.24E-05	-0.887375	2.08E-05
TP53	0.632119	4.45E-03	0.704852	2.15E-03
TTN	0.317041	4.43E-01	0.556318	1.36E-01
KRAS	-0.645084	1.29E-05	-0.601406	7.11E-05
PIK3CA	-0.108256	6.34E-01	-0.10705	5.77E-01

Table 2. DESeq2 analysis of the top five most mutated genes between normal tissue and cancer tissue in male and female CRC patients. A positive log2FoldChange represents higher expression in cancer tissue, while a negative log2FoldChange is higher expression in normal tissue. There does not appear to be a significant difference between the expression of these genes in normal vs cancer tissue between male and female patients.

	Male Patients		Female Patients	
Gene Name	**log2FoldChange**	**padj**	**log2FoldChange**	**padj**
SLC30A2	3.360023	1.31E-09	0.410356	4.97E-01
RNA5SP225	3.567865	2.35E-01	0.828148	8.60E-01
BPIFB1	3.629322	6.19E-04	0.710147	5.36E-01
CASC8	3.902552	4.37E-22	0.974988	2.25E-02
RNA5-8SP5	4.331192	5.16E-01	0.583018	9.11E-01

Table 3. DEseq2 analysis showing the top 5 genes that are overexpressed in cancer tissue compared to normal tissue in male CRC patients but not female patients. A positive log2FoldChange represents higher expression in cancer tissue, while a negative log2FoldChange is higher expression in normal tissue. Each of these genes have a log2FoldChange greater than 1 in male patients, but less than 1 in female patients.

	Male Patients		Female Patients	
Gene Name	**log2FoldChange**	**padj**	**log2FoldChange**	**padj**
NDUFB4P1	0.636343	7.29E-01	3.435513	4.65E-02
AL121759.1	0.744646	3.46E-01	3.457415	5.52E-05
MTCYBP3	0.991466	2.56E-01	3.480849	6.31E-03
AC092666.1	0.455752	5.22E-01	3.578039	5.91E-05
VSIG1	0.604403	2.43E-01	2.811835	3.38E-08

Table 4. DEseq2 analysis showing the top 5 genes that are overexpressed in cancer tissue compared to normal tissue in female CRC patients but not male patients. A positive log2FoldChange represents higher expression in cancer tissue, while a negative log2FoldChange is higher expression in normal tissue. Each of these genes have a log2FoldChange greater than 1 in female patients, but less than 1 in male patients.

	Male Patients		Female Patients	
Gene Name	**log2FoldChange**	**padj**	**log2FoldChange**	**padj**
PCDH11Y	-3.876945	8.91E-08	0.708959	6.62E-01

EVX2	-2.992034	1.95E-07	1.902321	7.98E-03
CPB1	-2.985887	1.03E-04	-0.823741	3.39E-01
THBS4	-2.683584	2.84E-07	-0.529879	3.71E-01
OR51E2	-2.458816	7.32E-13	-0.911789	2.33E-02

Table 5. DEseq2 analysis showing the top 5 genes that are underexpressed in cancer tissue compared to normal tissue in male CRC patients but not female patients. A positive log2FoldChange represents higher expression in cancer tissue, while a negative log2FoldChange is higher expression in normal tissue. Each of these genes have a log2FoldChange less than 1 in male patients, but greater than 1 in female patients.

	Male Patients		**Female Patients**	
Gene Name	**log2FoldChange**	**padj**	**log2FoldChange**	**padj**
TREH	-0.858492	2.15E-02	-3.213631	1.90E-17
XPNPEP2	-0.661773	2.20E-01	-3.108729	5.79E-10
FADS6	-0.860738	2.50E-01	-3.015768	9.79E-06
BTNL2	-0.975765	2.10E-01	-2.572947	5.22E-04
CYP2C8	-0.670567	8.68E-02	-2.532882	5.07E-12

Table 6. DEseq2 analysis showing the top 5 genes that are underexpressed in cancer tissue compared to normal tissue in female CRC patients but not male patients. A positive log2FoldChange represents higher expression in cancer tissue, while a negative log2FoldChange is higher expression in normal tissue. Each of these genes have a log2FoldChange less than -1 in female patients, but greater than -1 in male patients.

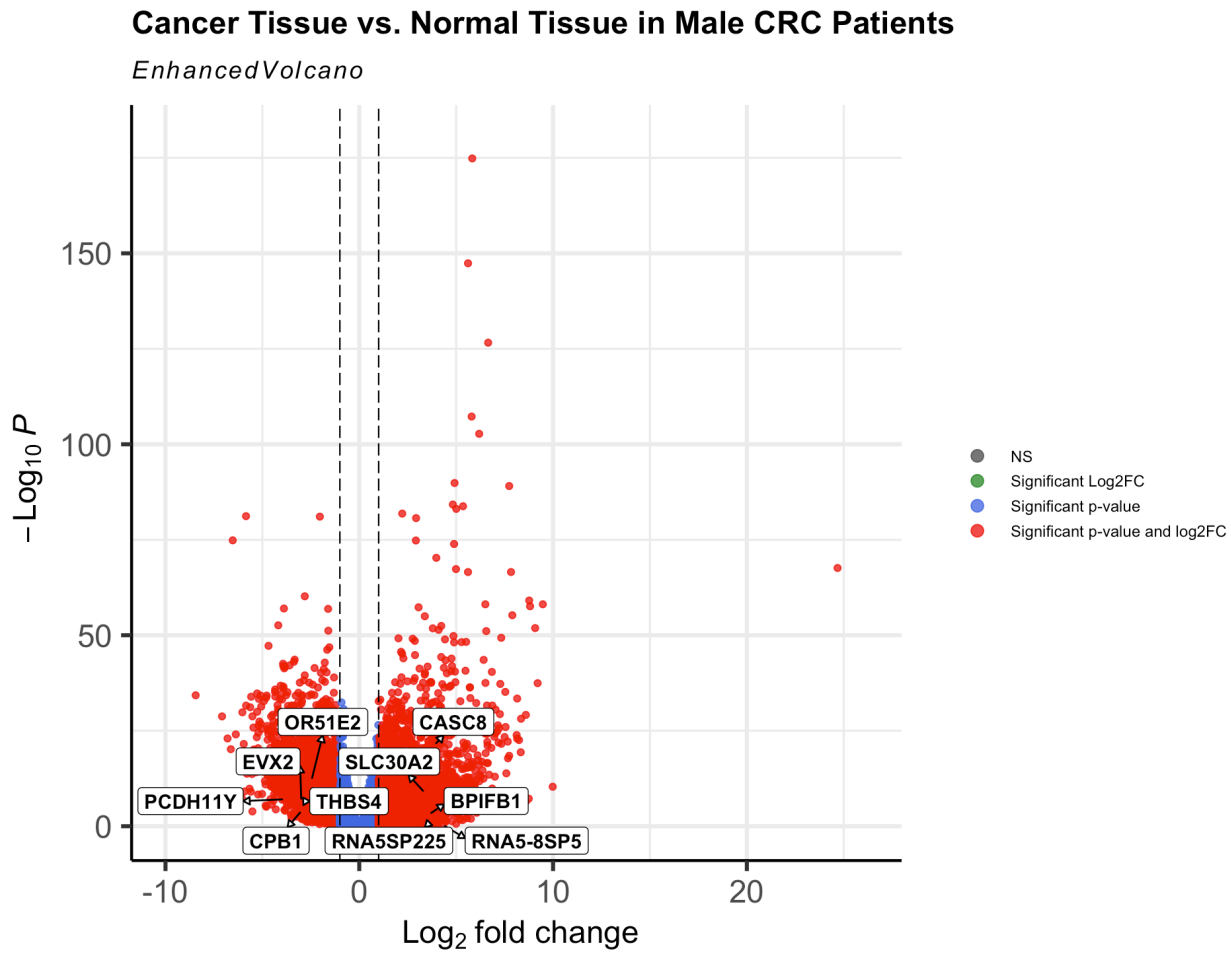

Figure 5. Volcano plot showing the differential expression of genes between cancer tissue and normal tissue in male CRC patients. The genes on the right are those that are overexpressed in cancer tissue while the genes on the left are those that are overexpressed in normal tissue. The labeled genes are those that have the highest log2FoldChange (absolute value) but are only significantly differentially expressed in male patients and not female patients. We only considered genes with a significant difference in expression as those with a Log2FoldChange greater than or equal to 1 or less than or equal to -1, which would represent at least a 2x difference in expression. The p-value threshold is set to 0.05.

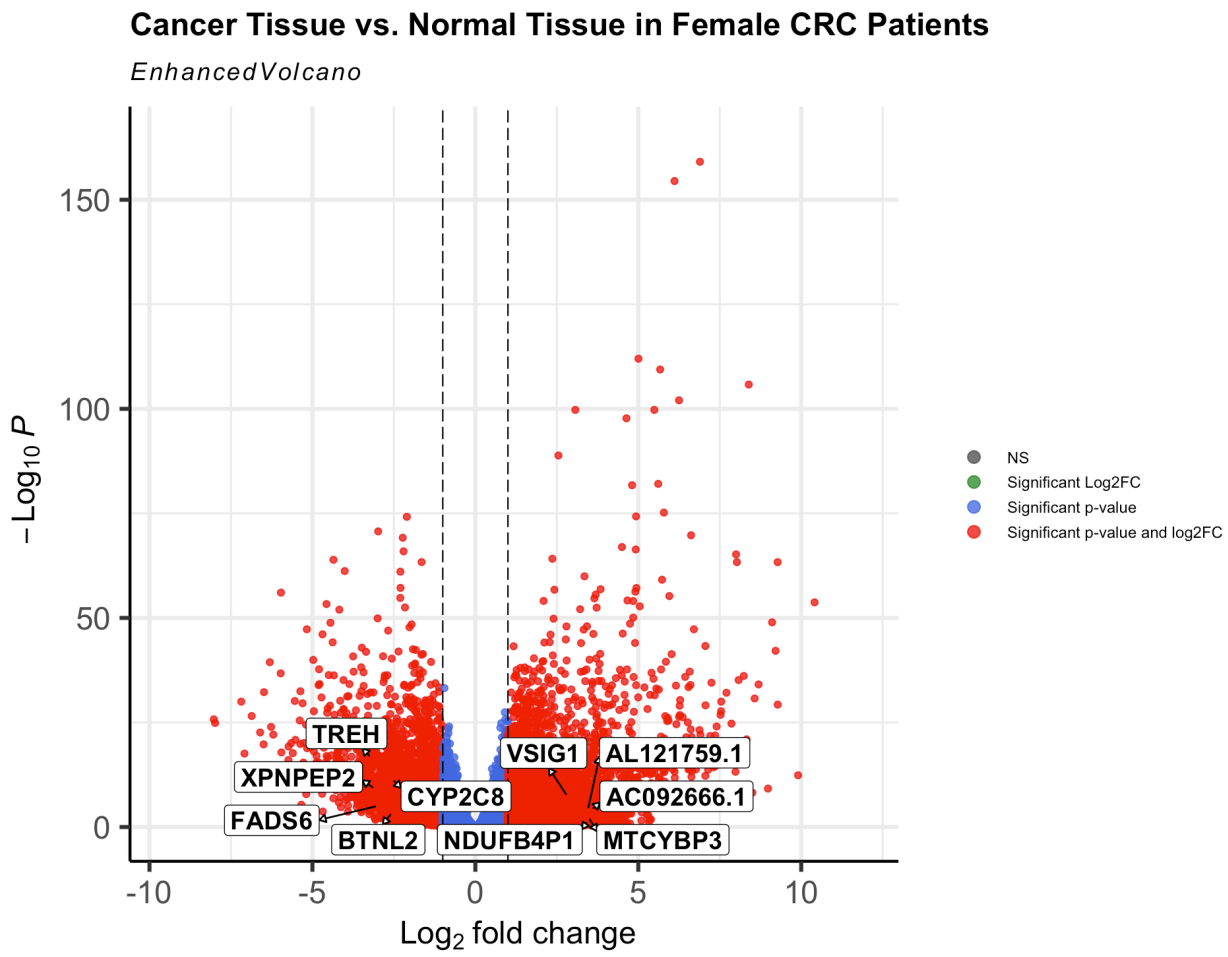

Figure 6. Volcano plot showing the differential expression of genes between cancer tissue and normal tissue in female CRC patients. The genes on the right are those that are overexpressed in cancer tissue while the genes on the left are those that are overexpressed in normal tissue. The labeled genes are those that have the highest log2FoldChange (absolute value) but are only significantly differentially expressed in female patients and not male patients. We only considered genes with a significant difference in expression as those with a log2FoldChange greater than or equal to 1 or less than or equal to -1, which would represent at least a 2x difference in expression. The p-value threshold is set to 0.05.

The expression of the top five most mutated genes was then analyzed through stratifying by tissue type. When comparing the differential expression of the top 5 most mutated genes between normal tissue and cancer tissue, there was no significant difference in the differential expression of these genes between normal and cancer tissue in both male and female patients. For instance, the *APC* gene had a log2FoldChange of -0.85 in male patients and -0.88 in female

patients, which means that the *APC* gene was expressed in normal tissue at almost twice the level as *APC* expression in cancer tissue for both male and female patients. Similarly, *TP53* had a log2FoldChange of 0.63 and 0.70 for male and female patients, respectively, indicating that cancer tissues overexpressed *TP53* at similar levels in both male and female patients.

However, when identifying the genes that were only overexpressed (genes with a Log2FoldChange greater than or equal to 1 or less than or equal to -1) in cancer tissue in male patients but not in female patients, a number of differences were found. The top five genes that were only overexpressed in cancer tissue in male patients were *SLC30A2*, *RNA5SP225*, *BPIFB1*, *CASC*, and *RNA5-8SP5*. Many of these genes like *BPIFB1* and *SLC30A2* are metabolic genes, playing roles in nutrient transport and inflammatory response. The top five genes that were underexpressed in cancer tissue compared to normal tissue in only male patients but not female patients were *PCDH11Y*, *EVX2*, *CPB1*, *THBS4*, and *OR51E2* (Figure 5). Contrasting the overexpressed genes in male cancer tissue, which play metabolic roles, genes like *PCDH11Y*, *THBS4*, and *OR51E2* are cell recognition and identification genes.

On the other hand, the top 5 genes that were only overexpressed in cancer tissue in female patients but not male patients were *NDUFB4P1*, *AL121759.1*, *MTCYBP3*, *AC092666.1*, and *VSIG1*, while the top 5 genes that were only underexpressed in cancer tissue in female patients were *TREH*, *XPNPEP2*, *FADS6*, *BTNL2*, and *CYP2C8* (Figure 6). Among the genes overexpressed in cancer tissue among female patients, *VSIG1* plays a role in cell-to-cell recognition in immune response. Many underexpressed genes in female cancer tissue like *XPNPEP2* and *FADS6*, however, are metabolic genes.

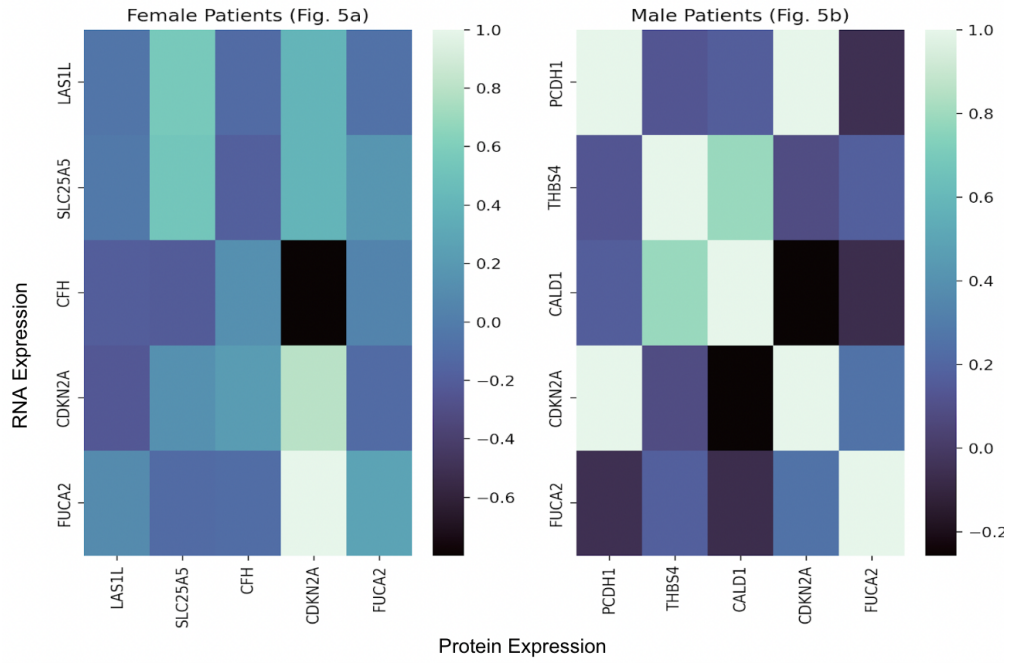

Figure 7. Analysis of correlation coefficients of RNA and protein expression and male and female patients shows that the relationship between a gene's RNA and protein expression are sexually differentiated. (a) In female patients, only CDKN2A has highly correlated RNA and protein expression. (b) In male patients, all genes of interest have very highly correlated RNA and respective protein expression.

From the differential expression analysis, five genes from the top ten most expressed genes for each sex were taken for further analysis using CPTAC proteomics data. Not all genes chosen were the five most expressed genes for each respective sex; because the CPTAC lacked proteomics data for some genes of interest for male patients, the RNA and protein expression of *CDKN2A* and *FUCA2* was analyzed for both sexes. *PCDH1* was used for male patients in this analysis, as *PCDH11Y* is a common variant of *PCDH1* found exclusively on the Y chromosome.

Through this analysis, we wanted to validate that changes in RNA expression would be associated with an increase in protein levels. The genes examined in male patients had very highly correlated RNA and respective protein expression. In addition to highly correlated RNA and respective protein expression, *PCDH1* expression was also highly correlated with *CDK2NA*

expression, suggesting that the two genes' expression may be dependent on each other among male patients (Figure 7b). On the other hand, only *CDKN2A* was found to have highly correlated RNA and respective protein expression among female patients with a correlation coefficient of 0.8. All other genes' RNA and respective protein expression examined among female patients had low to moderate correlation (Figure 5a). However, the correlation between *CDK2NA* protein expression and *FUCA2* RNA expression among female patients was very high. This relationship was absent among male patients, where the correlation coefficient between *CDKN2A* protein and *FUCA2* RNA expression was less than 0.3 (Figures 5a and 5b). This discrepancy suggests that the relationship between a gene's RNA and protein expression differs significantly between sexes.

Discussion

The survival plot in Figure 1 shows that female CRC patients had better overall survival rates than male CRC patients, which supports previous reports (Yang et al., 2017). To further investigate potential causes of this difference in survival rate based on sex, we explored the molecular differences between male and female CRC patients.

First, the top 5 most commonly mutated genes in both female and male CRC patients were found to be *APC*, *TP53*, *TTN*, *KRAS* and *PIK3CA* (Figure 2). However, visual analysis of the gene counts for each of these genes showed no significant differences between sexes (Figure 3). Further analysis showed that none of the top five most mutated genes showed significant differential expression (Table 1). Thus, while these genes are known to play significant roles in the progression and metastasis of colorectal cancer, they are unlikely to be responsible for the difference in survival between male and female CRC patients.

To identify potential genes that could impact the observed difference in survival rate between male and female CRC patients, we initially conducted a differential expression analysis of all genes between male and female patients. However, the results of this analysis identified genes that were predominantly sex-linked, which provided little information about genes that had an impact on CRC disease progression and survival. Thus, we expanded our analysis to compare the differential expression of genes between normal tissue and cancer tissue in male and female patients. This analysis allowed us to identify a number of genes that were significantly differentially expressed in cancer tissue compared to normal tissue in male patients but not female patients, and vice versa.

One differentially expressed gene that we identified that could have clinical significance was *VSIG1*, shown to be only highly overexpressed in female cancer tissue, but not male cancer tissue (Table 4). It was reported that *VSIG1* can play a role in reducing metastatic behavior from multiple cancer cell types (Bernal et al. 2020). As such, *VSIG1* being more highly expressed in the cancer tissue of female patients may contribute to the better survival rates of female patients.

Another significant gene in our analysis was *BPIFB1*, which was shown to be highly overexpressed in male cancer tissue, but not female cancer tissue (Table 3). In male patients, *BPIFB1* had a log2FoldChange of 3.69 in cancer tissue compared to normal tissue, while in female patients, *BPIFB1* only had a log2FoldChange of 0.71 in cancer tissue compared to normal tissue. Additionally, DESeq2 analysis also showed that male CRC patients generally overexpressed *BPIFB1* at double the rate of that in female patients. Thus, *BPIFB1* serves as another gene that can impact the difference in survival rate in male and female patients, as *BPIFB1* has been shown to play an important role in the development of tumors in gastric cancer and other cancer types (Li et al., 2020).

The RNA and protein expression analysis revealed that all genes' RNA and respective protein expression examined among male patients were very highly correlated, whereas the same trend was only found among one gene in female patients (Figures 5a and 5b). For female patients, *CDKN2A* protein expression was also found to have been highly correlated with *FUCA2* RNA expression (Figure 5a). *FUCA2* was found to be correlated to an immunosuppressive microenvironment and upregulated in many tumors in a previous study (Zhong et al. 2021); however, research on its links to *CDKN2A* is very limited. We suggest that future research directions for CRC among female patients will benefit from analyzing the relationship between *CDKN2A* and *FUCA2*.

THBS4, underexpressed among cancer tissue in male patients exclusively, is a cell migration gene and tumor suppressor gene previously found to be underexpressed in prostate cancer (Hou et al. 2020). Our findings about the sex-specific nature of *THBS4* support this study and highlight *THBS4* as an important therapeutic target for male patients. Furthermore for male patients, the high correlation between *PCDH1* expression and *CDKN2A* expression also implies that these two genes may be dependently expressed (Figure 5b). This supports a previous observations in pancreatic cancer patients where *CDKN2A* and *PCDH1* were commonly co-expressed (Iguchi et al. 2016). As both genes code proteins found to play roles in cell-to-cell recognition in immune response, our findings suggest that their functional relationship may also hold for CRC patients. Therefore, we propose that *PCDH1* and *CDKN2A* are possible therapeutic targets among male CRC patients. As female patients also display high correlation among *CDKN2A* RNA and protein expression, future research on *CDKN2A* in CRC among both sexes may be beneficial.

References

Bernal, C., Silvano, M., Tapponnier, Y., Anand, S., Angulo, C., & Ruiz i Altaba, A. (2020). Functional Pro-metastatic Heterogeneity Revealed by Spiked-scrnaseq is shaped by cancer cell interactions and restricted by VSIG1. *Cell Reports*, 33(6), 1–19.

Chang, C.-H., Lee, C.-H., Ho, C.-C., Wang, J.-Y., & Yu, C.-J. (2015). Gender-based impact of epidermal growth factor receptor mutation in patients with nonsmall cell lung cancer and previous tuberculosis. *Medicine*, 94(4), 1–8.

Hou, Y., Li, H., & Huo, W. (2020). THBS4 silencing regulates the cancer stem cell-like properties in prostate cancer via blocking the PI3K/akt pathway. *Prostate*, 80(10), 753–763.

Huang, D., Sun, W., Zhou, Y., Li, P., Chen, F., Chen, H., Xia, D., Xu, E., Lai, M., Wu, Y., & Zhang, H. (2018). Mutations of key driver genes in colorectal cancer progression and metastasis. *Cancer and Metastasis Reviews*, 37(1), 173–187.

Iguchi, E., Safgren, S. L., Marks, D. L., Olsen, R. L., & Fernandez-Zapico, M. E. (2016). Pancreatic Cancer, A Mis-interpreter of the Epigenetic Language. *Yale Journal of Biology and Medicine*, 89(4), 575–590.

Li, J., Xu, P., Wang, L., Feng, M., Chen, D., Yu, X., & Lu, Y. (2020). Molecular biology of BPIFB1 and its advances in disease. *Annals of Translational Medicine*, *8*(10), 651–651. https://doi.org/10.21037/atm-20-3462

Lopes-Ramos, C. M., Quackenbush, J., & DeMeo, D. L. (2020). Genome-wide sex and gender differences in cancer. *Frontiers in Oncology*, 10(1), 1–17.

National Institute of Health. (n.d.). *Clinical proteomic tumor analysis consortium (CPTAC)*. National Cancer Institute Genomic Data Commons.

Retrieved April 22, 2022, from https://gdc.cancer.gov/about-gdc/contributed-genomic-data-cancer-research/clinical-proteomic-tumor-analysis-consortium-cptac

National Cancer Institute. (n.d.). *Common Cancer Sites - Cancer Stat Facts*. SEER. Retrieved April 22, 2022, from https://seer.cancer.gov/statfacts/html/common.html#:~:text=Lung%20and%20bronchus20cancer%20is%20responsible%20for%20the%20most%20deaths,deadliest%20cancer2C%20causing%2048%2C220%20deaths

National Institute of Health. (n.d.). *The Cancer Genome Atlas Program*. National Cancer Institute. Retrieved April 22, 2022, from https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga

Oh, J.-H., Jang, S. J., Kim, J., Sohn, I., Lee, J.-Y., Cho, E. J., Chun, S.-M., & Sung, C. O. (2020). Spontaneous mutations in the single TTN gene represent high tumor mutation burden. *Npj Genomic Medicine*, 5(1), 1–11.

Press, O. A., Zhang, W., Gordon, M. A., Yang, D., Lurje, G., Iqbal, S., El-Khoueiry, A., & Lenz, H.-J. (2008). Gender-related survival differences associated with EGFR polymorphisms in metastatic colon cancer. *Cancer Research*, 68(8), 3037–3042.

Wang, L., Xiao, J., Gu, W., & Chen, H. (2016). Sex difference of EGFR expression and molecular pathway in the liver: Impact on drug design and cancer treatments? *Journal of Cancer*, 7(6), 671–680.

Yang, Y., Wang, G., He, J., Ren, S., Wu, F., Zhang, J., & Wang, F. (2017). Gender differences in colorectal cancer survival: A meta-analysis. *International Journal of Cancer*, 141(10), 1942–1949.

Zhong, A., Chen, T., Xing, Y., Pan, X., & Shi, M. (2021). Fuca2 is a prognostic biomarker and correlated with an immunosuppressive microenvironment in Pan-Cancer. *Frontiers in Immunology*, 12(1), 1–12.