2023

Advancing our understanding on the role of Bromodomain PHD-Finger Transcription Factor (BPTF) in Cancer through the Analysis of Publicly Available Databases.

Preksha Jerajani
Virginia Commonwealth University

Advancing our understanding on the role of Bromodomain PHD-Finger Transcription Factor (BPTF) in Cancer through the Analysis of Publicly Available Databases.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics at Virginia Commonwealth University

By
Preksha Jerajani
B.S in Bioinformatics, 2021
Virginia Commonwealth University

Director: Dr. Joseph Landry, Ph.D.
Assistant Professor
Department of Human and Molecular Genetics

Virginia Commonwealth University
Richmond, Virginia
May 2023

Acknowledgments

First and foremost, I would like to thank my advisor, Dr. Joseph Landry, and committee members, Dr. Amy Olex, Dr. Dana Lapato and Dr. Lamont Cannon. Without their guidance, support, and dedication to my success, none of this work could have been possible. I feel very fortunate they took a chance on advising me through this process. I thank the Department of Life Sciences and VCU Bioinformatics for their constant guidance and collaborative community. I want to extend my thank you to Dr. Allison Johnson for guiding me and having my back since day one of my undergraduate studies at Virginia Commonwealth University. I'd like to thank my parents Shital and Yatri Jerajani for always being there and supporting me. Their constant encouragement and love, drove me to do my best and challenge myself academically. I would like to extend my thanks to fellow classmates and lab members that have always been there in case I needed any help. I would like to specifically thank Christiane Morecock for always being there to help me out and providing guidance. I would also like to thank Akshaya Ranganathan and Anushka Jain for their help and their hard work on related projects. I'd like to thank my friends, Emily Herman, Swathi Sowmitran, Narthana Kambalapally, Barjaa Brown, Shannon Hendricks, Iulia Voina, Sehaj Kaur and Jamie-Jean Gilmer for keeping me balanced and motivated. Finally, I would like to thank my best friend and partner Ryan Butler for his unending support and love.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

DAVID – Database for Annotation, Visualization and Integrated Discovery

BLCA – Bladder Urothelial Carcinoma

BPTF – Bromodomain PHD Finger Transcription Factor

BRCA – Breast Invasive Carcinoma

NURF – Nucleosome Remodeling Factor

PRAD – Prostate Adenocarcinoma

THCA – Thyroid Carcinoma

UCEC – Uterine Corpus Endometrial Carcinoma

NMIBC – Non-Muscle Invasive Bladder Cancer

MIBC – Muscle Invasive Bladder Cancer

ER – Estrogen Receptors

PR – Progesterone Receptors

TNBC – Triple Negative Breast Cancer

HER2 – Human Epidermal Growth Factor Receptor 2

CRC – Chromatin Remodeling Complex

ISWI- Imitation Switch/Sucrose Non-Fermentable

SWI – Switch/Sucrose Non-Fermentable

CHRAC – Chromatin Accessibility Complex

ACF – ATP – Dependent Chromatin Assembly Factor

ZNF – Zinc Finger Proteins

TRRAP – Transformation/Transcription Domain Associated Protein

NSD – Nuclear Receptor Binding SET Domain

DMAP – DNA Methyltransferase Associated Protein

ATF1/2 – Activating Transcription Factor 1 and 2

SMARCC1/2 – SWI/SNF Related, Matrix Associated, Actin Dependent Regulator of Chromatin Subfamily C Member 1 and 2

NF2 – Neurofibromatosis Type 2

MYCBP2 – MYC Binding Protein 2

MAPK1 – Mitogen-Activated Protein Kinase 1

BRCA2 – Breast Cancer Gene 2

DAPK3 – Death Associated Protein Kinase 3

RHOC – Ras Homolog Family Member C

MYL – Myosin Light Chain 1

TOP1 – DNA Topoisomerase 1

RBBP6 – RB Binding Protein 6, Ubiquitin Ligase

TUT4 – Terminal Uridylyl Transferase 4

AKAP1 – A-Kinase Anchoring Protein 1

2. Abstract

Advancing our understanding on the role of Bromodomain PHD-Finger Transcription Factor (BPTF) in Cancer through the Analysis of Publicly Available Databases.

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Bioinformatics at Virginia Commonwealth University

By: Preksha Jerajani
B.S in Bioinformatics, 2021
Virginia Commonwealth University

Mentor: Dr. Joseph Landry, Ph.D.
Assistant Professor
Department of Human and Molecular Genetics

The initiation and progression of cancer are significantly influenced by genetic factors, and despite ongoing research, a definite cure for this disease remains elusive. Over time, researchers have explored various genes as potential targets for treating various types of cancer, and numerous innovative therapies continue to be developed. General chemotherapies face their own challenges like recurrence, while cancer specific drugs face the challenge of metastasis. Therefore, therapeutic strategies have slowly started to shift towards comprehending the pathways and corresponding genes that may be implicated in multiple types of cancer, with the aim of identifying more effective targets for drug development. One such potential target being the Bromodomain PHD-Finger Transcription Factor (BPTF), an essential subunit of Nucleosome Remodeling Factor (NURF) that has been reported through previous research to be a highly druggable potential target of NURF. This study seeks to improve our knowledge about the role of the BPTF gene in the onset of cancer and explore the correlated biological pathways and genes that might aid in further research and identification of cancer biomarkers using publicly available databases. Specifically, we use The Cancer Genome Atlas (TCGA) SpliceSeq to analyze alternative splicing events in both normal and tumor samples, and TCGA Wanderer to examine CpG methylation patterns in normal and tumor samples. Additionally, we conducted gene ontology analysis using the Database for Annotation, Visualization and Integrated Discovery (DAVID) to identify shared pathways and genes that may be relevant for further research and potential therapeutic targeting.

3. Introduction

3.1 Cancer Review

The definition of cancer is the uncontrollable growth and spread of the body's cells to other parts of the body. The role of genetics in the onset of cancer is significant, as genes control the way cells function, and genetic changes that lead to cancer may result from errors in cell division, DNA damage caused by harmful environmental substances, or inheritance (National Cancer Institute [NCI], 2011). Over the years, these factors have been studied, and therapies have been developed to target the overall process and the genes responsible for a more targeted approach. Advancements in technology, such as DNA sequencing, proteomics, and transcriptomics, have resulted in the discovery of more information about multiple pathways being correlated and working together to bring about the onset of cancer.

Metastasis is a significant challenge when considering cancer therapies and studying cancer, as it can spread to other parts of the body, making treatment for metastasized cancers complex. While treatment in these cases may help prolong people's lives, in other cases, the goal is to control the overall growth of the cancer and relieve symptoms that are causing the growth. Recurrence has also become a research topic over the last few decades. Recurrence of cancer is commonly seen in aggressive cancer types and can be resistant to previous treatment (National Cancer Institute [NCI], 2020). This phenomenon poses a different set of challenges in terms of prevention, diagnosis, and treatment. As one of the most complex diseases that has perplexed the scientific community for decades, different cancers and their therapies have been studied, yet there is still so much more to be discovered.

3.1.1. Bladder Urothelial Carcinoma

Bladder cancer is a common neoplasm of the urinary system and is considered one of the top ten malignant tumors (Kaseb, H., & Aeddula, N. R., 2022). Bladder urothelial carcinoma, which is the invasion of the basement membrane and lamina propria by neoplastic cells of urothelial origin, has an incidence rate of 350,000-380,000 cases per year worldwide (Kaseb, H., & Aeddula, N. R., 2022). According to Ferlay and Shin (2008), bladder cancer causes about 150,000 deaths annually. Bladder cancer is classified into non-muscle-invasive bladder cancer (NMIBC) and muscle-invasive bladder cancer (MIBC). NMIBC frequently recurs at a rate of 50-70% and progresses to MIBC at a rate of 15%. The progress of MIBC leads to more advanced

stages of cancer with a five-year survival rate of less than 50%. Bladder urothelial carcinoma is the most common pathological type of bladder cancer, accounting for 90% of tumors, and has a recurrence rate of 50-70% after surgery (Kaseb, H., & Aeddula, N. R., 2022). Due to the progressive nature of this cancer, there have been advances in molecular and drug treatments. The *FGFR3* receptor proteins and *TP53* are some genes that have been studied for their significance, and with new molecular findings and genetic engineering, molecular profiling is expected to be useful in providing better prognosis and more effective treatment (Bai, Y., et al., 2022).

3.1.2. Breast Invasive Cancer

Breast cancer is a complex disease that involves multiple cell types and stages. It is one of the most difficult cancers to prevent and the second leading cause of cancer deaths among women. Globally, it is responsible for about 570,000 deaths in 2015, and over 1.5 million women are diagnosed with breast cancer each year (Stewart, B. W., & Wild, C. P., 2014). This cancer has the ability to metastasize to distant organs such as the liver, lungs, and brain, making it challenging to treat. However, early detection is crucial for improving the prognosis and survival rate of patients (Sun, Y. S., et al., 2017).

Breast cancer diagnosis and therapy decisions are based on the immunohistochemistry (IHC) classifications of certain proteins/receptors including estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) (Onitilo, A. A., et al., 2009). The IHC profile determines whether a patient is positive or negative for these receptors and helps to inform treatment decisions. One of the most aggressive subtypes of breast cancer is triple negative breast cancer (TNBC), which is negative for ER, PR, and HER2 (Aysola, K., et al., 2013). TNBC is a heterogeneous disease that lacks effective therapies and has a lower overall survival rate.

Breast cancer initiation and progression theories differ among the various subtypes (Figure 1). Many risk factors increase the likelihood of developing breast cancer, including sex, age, family history, and genetic mutations, among others.

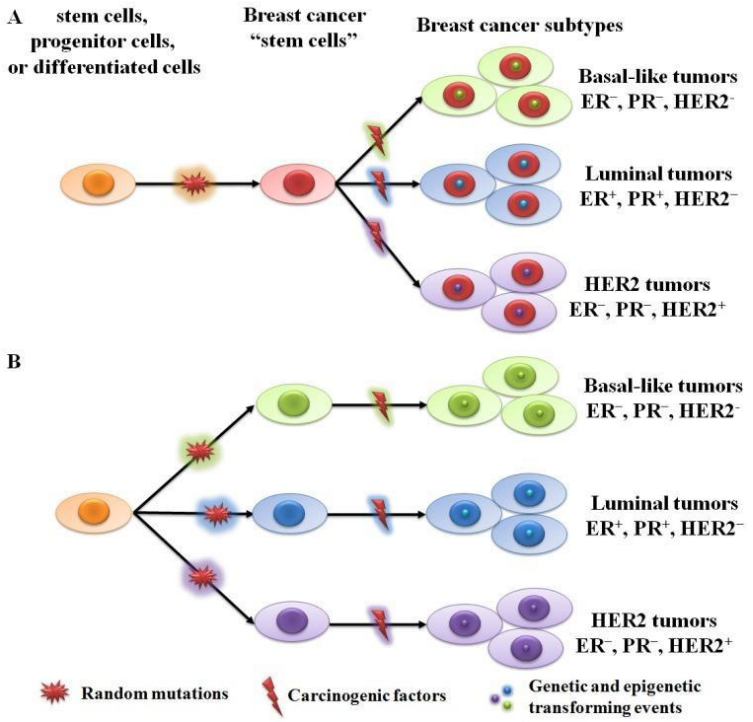

Figure 1: Breast Cancer Subtype Division Hypothesis. A) *A single breast cancer stem cell then gets characterized to different subtypes with different sets of proteins that are present or not based on carcinogenic factors.* ***B)*** *The stem cells get differentiated and further divide into separate categories to show breast cancer subtypes*. *Note.* Represented from "Risk Factors and Preventions of Breast Cancer", by Sun, Y. S., et.al, 2017, International Journal of Biological Sciences, 13(11), p.1389.

3.1.3. Prostate Adenocarcinoma

Prostate adenocarcinoma is the second most commonly diagnosed cancer in men worldwide, with over 1,276,106 new cases and 258,989 deaths in 2018. While this cancer may be asymptomatic in its early stages, frequent symptoms include difficulty with urination and nocturia. More severe symptoms include urinary retention and back pain (Bray et al., 2018). Studies have shown that dietary factors and physical activity can impact the development and progression of this disease, highlighting the need for a deeper understanding of risk factors and lifestyle factors (Figure 2) (Rawla, P., 2019). Researchers are focusing on identifying genes involved in the inheritance and somatic mutations that can be acquired from environmental changes. Genes that have been associated with the initiation and progression of prostate adenocarcinoma include PTEN, CDKN1B, and c-MYC (Hartmann, A., & Friess, H., 2017).

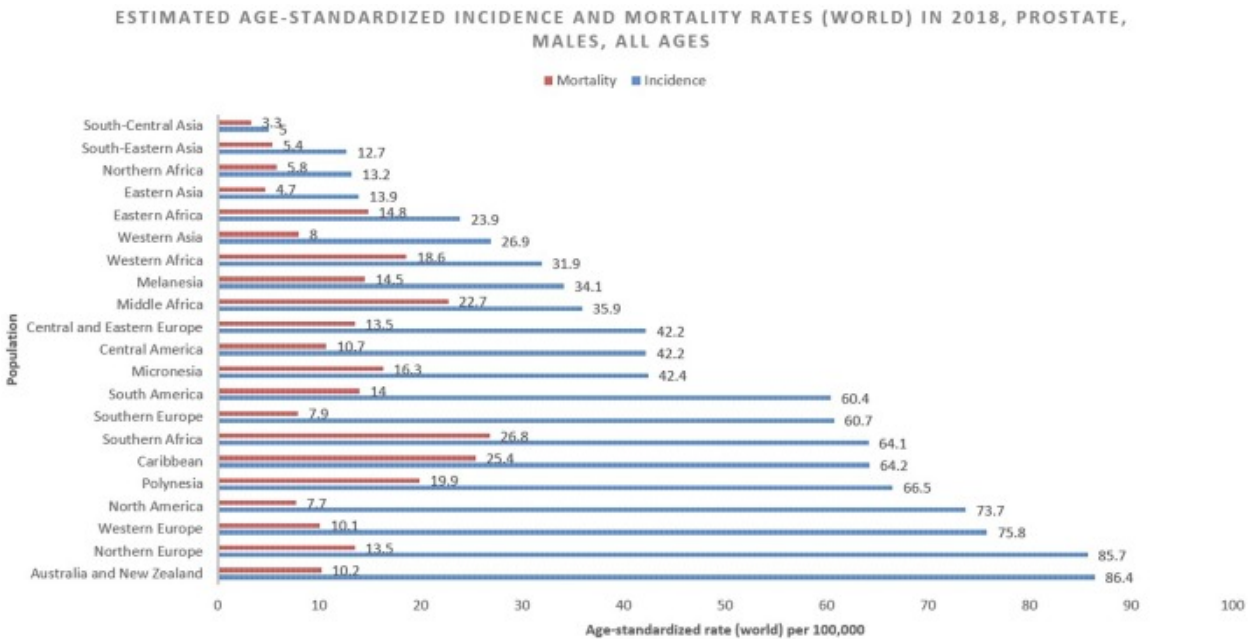

Figure 2: Prostate Adenocarcinoma Mortality and Incidence ASR Rate. *The bar chart illustrates the estimated age-standardized incidence and mortality rates in 2018 in males of all ages. The data was obtained from Globocan 2018. Note.* Represented from "Epidemiology of Prostate Cancer" by Rawla, P., 2019, World Journal of Oncology, 10(2), p.65.

3.1.4. Thyroid Carcinoma

As advancements in diagnostic imaging and surveillance continue, the incidence of thyroid cancer is on the rise. In the United States, it ranks as the fifth most common cancer among women (Cabanillas, M. E., 2016). However, a major challenge faced by physicians treating this cancer is to ensure that patients with lower risk disease or benign thyroid nodules are not overtreated. Genetic analysis of thyroid cancer through DNA sequencing has revealed that mutations in the mitogen-activated protein kinase (MAPK) cellular signaling pathway are present in most cases (Singh, A., et al., 2021). Additionally, other genetic alterations such as RET proto-oncogene, ALK, and BRAF have been shown to play a role in the development of thyroid carcinoma (Younis, E., 2017).

3.1.5. Uterine Corpus Endometrial Carcinoma

According to Rutgers (2015), endometrial cancer is the most commonly diagnosed female genital cancer and ranks fourth in incidence among women globally, following lung, colorectal, and breast cancers. Uterine corpus endometrial carcinoma (UCEC) is a frequent subtype of endometrial cancer, and its incidence increases with age, with most diagnoses occurring in

women between 45 and 65 years old. To aid in early detection and better treatment of UCEC, researchers have explored potential biomarkers such as activated leukocyte cell adhesion molecules, sperm-associated antigen 9, and heat shock protein family A, among others (Shen, L., et al., 2018).

3.2. Chromatin Remodeling Factors

When considering therapies and treatments for cancer, the most studied and utilized are metabolic enzymes/pathways and well-known oncogenes. According to *Therapeutic Targets in Cancer Cell Metabolism and Autophagy,* an article by Cheong, H., et.al (2012), a lot of these metabolic and genetic pathways have become targets for drug development. An aspect not studied in detail is the chromatin remodeling complexes (CRC) that have been noted to play a role in cancer onset and recurrence.

CRCs aid in modification and modulation of the chromatin. Cell proliferation is greatly affected by changes made to the chromatin structure, and these changes are often seen in cancer cells (Nair, S. S., & Kumar, R., 2012). The ATP-dependent chromatin remodeling factors serve to regulate DNA accessibility by repositioning, ejecting and/or modifying nucleosomes. CRCs are divided into four families that are differentiated based on their ATP domain. The families include the switch/sucrose nonfermentable, inositol-requiring mutant 80 families, chromodomain-helicase DNA binding protein and the imitation switch. All these complex families have many roles, for example, INO80 complexes aid in DNA damage repair, CHD complexes interact with may chromatin modifying complexes, and SWI aid in remodeling chromatins by nucleosome sliding, creating DNA loops in order to repress or enhance activation of genes.

These are some of the most studied families of CRC, however, the ISWI (imitation switch) which is the smaller family, has a specific complex known as the Nucleosome Remodeling Factor (NURF), that has been shown to play part in cancer biology in model organisms (Morecock, C., 2022). Along with the NURF complex, ISWI ATPase includes the chromatin assembly complex (CHRAC) and ATP-utilizing chromatin assembly and remodeling factor (ACF) complexes. Figure 4 illustrates the different classifications of the ISWI family and details the molecular components, subcellular localization, targeting inhibitors and the functional domains. First discovered in *Drosophila melanogaster* NURF was described as an ATP

dependent biochemical activity that helped mediate nuclease accessibility to reconstituted chromatin. When isolated and purified from human cells the complex was highly homologous to the one discovered in *Drosophila melanogaster* (Alkhatib, S. G., & Landry, J. W., 2011).

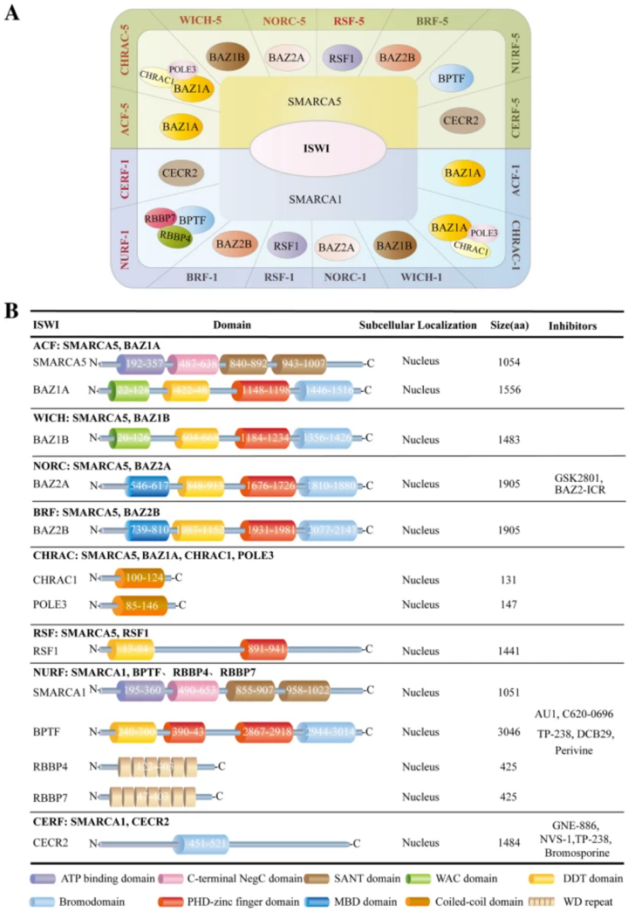

Figure 3: Classification of ISWI Units and Subunits. A) *The sixteen different types of complexes harboring either the SMARCA1 or SMARCA5 as ATPase subunits and 1-3 non-catalytic subunits.* ***B)*** *Representation of the ISWI, the domains (functional), the subcellular localization, the size in amino acids and the inhibitors for each of the ISWI members.Note.* Represented from "The Emerging Role of ISWI Chromatin Remodeling in Cancer" by Li, Y., et.al., 2021, Journal of experimental & clinical cancer research, 40(1), p.3.

3.3 Nucleosome Remodeling Factor (NURF)

The NURF complex has three subunits: histone binding protein (RBBP4/7), ATPase ISWI protein SNF2L which is encoded by *SMARCA1* and the bromodomain PHD-finger transcription factor (BPTF). The NURF complex is responsible for catalyzing ATP-dependent nucleosome sliding. BPTF is responsible for identifying the loci of methylation and acetylation on histones,

while RBBP4 and RBBP7 are components of multiple protein structures, and the role of SMARCA1 varies depending on the type of tumor. SMARCA1 and its homologs are mainly responsible for cell survival and cell cycle progression.

Figure 4A details the overall structure of NURF and its interaction with nucleosomes. NURF's structure and its individual components play an important role in interacting with other transcription factors and genes that have been related to cell division, cell differentiation and gene expression. For example, MITF also known as microphthalmia associated transcription factor plays an essential role in the differentiation, survival and proliferation of normal melanocytes and in controlling the melanoma cell physiology. The MITF interacts with the NURF complex and this interaction discovery led to investigations of their role in melanoma cells. Investigations by Koludrovic, D., et.al in 2015, found that *BPTF*, a component of NURF, is essential to produce differentiated adult melanocytes. They discovered that MITF and BPTF co-regulate *PREX1* expression in melanocytes in vitro. The *PREX1* gene encodes a protein that acts as a guanine nucleotide exchange factor for the RHO family of small GTP binding proteins. This gene is not only seen in relation to melanoma but also plays a role in breast cancer. It is a key mediator in ErbB signaling in breast cancer and has been implicated in mammary tumorigenesis and metastatic dissemination.

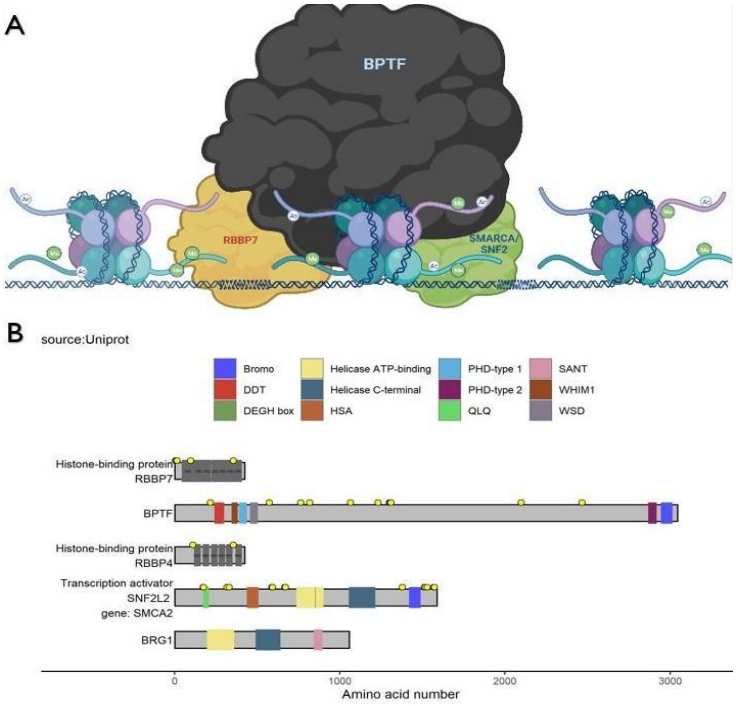

Figure 4: NURF complex and its Protein Domains. A) *NURF subunits contribute to the sliding of the nucleosomes along the DNA.* ***B)*** *The possible units of the NURF complex's functional domains and phosphorylation sites are done to scale. (A)* created with BioRender.com*, B)* Represented from "Analyses of Internal CRISPR Screen and RNA-Seq data, and Publicly Available Genomic Datasets to Study the Roles of Epigenetics in Cancer Biology." by Morecock, C.*, 2022, VCU* Scholars Compass).

3.4 Bromodomain PHD-Finger Transcription Factor

The significance of the pathogenic variants of BPTF was reported in 2017 in other neurodevelopmental disorders with dysmorphic faces and distal limb anomalies. Over the years it has been identified as a pro-tumorigenic factor. BPTF, being the largest subunit of NURF, has a crucial function in ensuring the proper functioning of the NURF complex. Thus, the amplification of BPTF in cancerous cells and possible alterations in the BPTF gene that encodes this protein have demonstrated the potential to increasingly promote cell proliferation.As seen in Figure 5 BPTF is located on chromosome 17. BPTF consists of a DDT domain, two PHD-Fingers and a bromodomain. DDT better characterized as the DNA-binding homeobox-containing proteins and the different transcription and chromatin remodeling factor which is a domain generally of 60 amino acids that contains the regions of conserved phenylalanine and leucine residues that aids in DNA binding and is essential (Doerks, T., et.al., 2001). PHD-fingers are structurally conserved modules that are usually found in proteins that modify chromatin and mediate molecular interactions in transcription and it generally recognizes methylated lysine residues in histone tails. The bromodomain determines the binding affinities and aids in acetylation binding. The features of BPTF allow for it to bind to multiple histones like H4K16ac and H3K4me3.

In recent decades, researchers have observed that the amplification of BPTF promotes malignancy. This was demonstrated through knockdown analysis, which showed that reducing BPTF levels slowed down cell proliferation. In BLCA the Circ-BPTF which is derived from the BPTF exons has been shown to be overexpressed compared to normal cell lines, in BRCA the copy number of BPTF was gained in 34.1% and amplified in 8.2% of the breast cancer tissue cohorts, and in THCA the gene along with others like *NCOR2* and *ANK3* was found to have a high frequency of mutations.

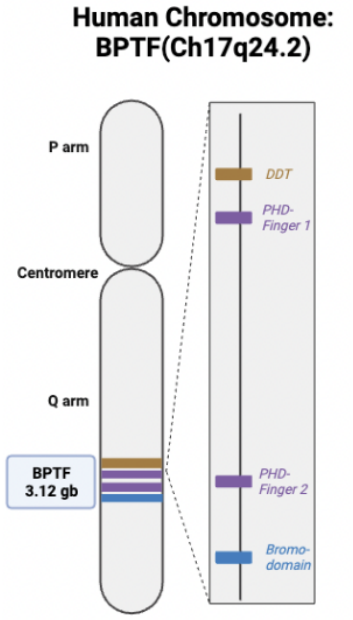

Figure 5: BPTF Ch17q24.2. *The location of the BPTF gene is on Chromosome 17 in the Q arm. There are four main functional domains, DDT, PHD-Fingers 1, PHD-Finger 2 and the Bromodomain. BPTF is about 3.12 gb in length.* (created with BioRender.com)

Furthermore, the overexpression of BPTF leads to downstream changes, such as it affecting MAPK signaling which regulates cell cycle and survival, and c-MYC which is a transcription factor that has been shown to be overexpressed in many human cancers. c-MYC is a well-studied gene/pathway and currently has drug therapies that target the pathway. Since BPTF interaction with c-MYC occurs in most if not all tumor types, it can be a potential target for treating c-MYC driven tumors (Zahid, H., et.al., 2021). Although there is significance noted in these interactions, there are not many detailed studies on the overall effect of overexpression of BPTF and how the interactions between BPTF and other co-related genes can be targeted. Along with the overexpression of the BPTF gene, somatic mutations of the gene were found to be correlated with tumor volume. The gain of function mutation of BPTF could confer the abnormal expression and activate sets of target genes that could contribute to tumorigenesis and maintenance thereafter of the cancer (Duan, C., et al., 2018).The goal is to gain a deeper understanding of BPTF and the NURF complex as sources of potential therapy targets in different cancer types.

3.5 Alternative Splicing

mRNAs are the molecular templates for the synthesis of proteins. In humans, pre-mRNA splicing is essential for the expression of more than 95% of the genes. Removal of introns from the precursor messenger RNA is an essential step for the expression of most eukaryotic genes. Alternative splicing gives rise to a diverse set of mRNAs that get translated and turned to proteins, which eventually aids in normal cellular function. Different splicing events are common in cancers and are associated with mutations and altered expression of the components of the machinery. Cancer cells usually generate advantageous splicing variants that provide a growth advantage. The cells often show increased splicing complexity and diversity compared to normal cells, indicating that they have developed mechanisms to regulate and fine-tune the splicing process. These changes in efficiency can be targeted by developing inhibitors or targeted therapies. The analyses of over 8000 tumors across 32 cancer types revealed thousands of variants that are not present in non-malignant tissue that were likely to generate cancer-specific markers and neoantigens (Bonnal, S. C., et al., 2020). Since BPTF is known to be overexpressed in most cancer types, we can analyze if there are any specific splicing patterns that occur more often in a specific type of cancer compared to the normal splicing patterns in normal cells. We can also illustrate if there is significance in the splicing patterns of BPTF in these cancer types.

3.6 CpG Methylation

DNA methylation is a type of epigenetic mechanism that involves the transfer of a methyl group onto the 5th carbon position of the cytosine to form a 5-methylcytosine. DNA methyltransferase mediates this process by aiding in creating a covalent bond between the methyl group and the 5-carbon on the cytosine base. Furthermore, this process only occurs on the cytosine linked directly to a guanine by the phosphodiesterase link which forms the CpG dinucleotide pair. Methylation usually occurs near the promoter regions of different genes, and generally methylation patterns in normal cells differ from those in cancer cells. Since cancer cells go through extensive genetic changes, using this method to gain insight on the different processes by which the cancer is developed can lead to better understanding and more effective targeting for therapies. DNA is usually hypomethylated which refers to the loss of the methyl group in the 5-methylcytosine nucleotide than hypermethylated during carcinogenesis leading to a net decrease in the genomic 5-methylcytosine content. There are also considerable differences in methylation based not only on the species but is also tissue specific therefore one must consider

the changes in between normal and tumor methylation from the same tissue to conduct proper analysis. There are many drugs developed that target these epigenetic changes, specifically some that demethylate DNA (Ammerpohl, O., et.al., 2016). These drugs normally target hypermethylated regions in tumor suppressor genes that have been dysregulated in many cancers. Figure 6 illustrates how hypomethylation and hypermethylation affect the transcription of the RNA. The hypermethylation in and around the CpG island leads to the overall shutdown of the transcription. DNA methylation could be interesting to look at for individualized antitumor targeting, which would be a more personalized method of therapy.

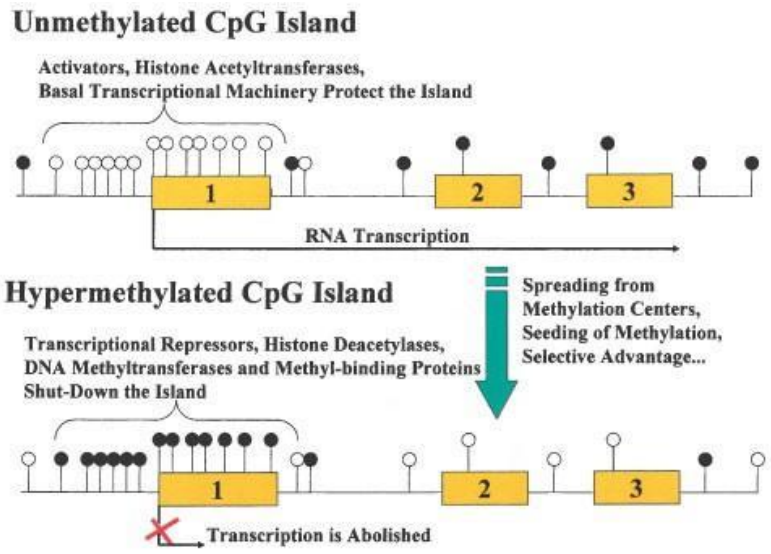

Figure 6: Hypomethylation and Hypermethylation of CpG Islands. *The CpG island of a tumor suppressor gene is represented in normal and a tumor cell. The black dots represent methylated CpG islands while the white dots represent the unmethylated CpG islands. Note.* Represented from "CpG island hypermethylation and tumor suppressor genes: a booming present, a brighter future.", by Esteller, M., 2002, Oncogene, 21(35), p.5429.

3.7 Gene Ontology

Gene Ontology (GO) is a means of describing biological knowledge in three aspects: molecular functions, cellular components, and biological processes. According to the *GeneOntology Unifying Biology* website, an example of GO annotation is the cytochrome c gene product, which can be described by its molecular function as an oxidoreductase, its biological process as oxidative phosphorylation, and its cellular component as the mitochondrial matrix. These three aspects are interdependent and work in tandem to provide a more comprehensive understanding

of related pathways and genes. In cancer research, GO analysis has been employed to identify co-related pathways and genes, and to uncover potential biomarkers for targeted therapies. For instance, in a study on secondary bone cancer conducted by Dr. Shikha Vashisht and Ganesh Bagler in 2012, a cancer gene network was constructed to represent the protein interactome of cancer genes. This resulted in a more comprehensive understanding of the role of molecular regulators in cancer. Additionally, GO is highly valuable in the pharmaceutical and drug development industry, as it enables the combination of gene and pathway knowledge to create a working network that can be used to identify essential and non-essential pathways.

3.8 Rationale for Study

The overexpression of BPTF has been observed in various cancer types, based on previous research. However, the underlying reasons behind this overexpression are still unclear. Therefore, this study aims to investigate the mechanisms that drive changes in the expression of BPTF and analyze the genes that are co-expressed with BPTF in five specific types of cancer. Through carrying out an examination of the expression of BPTF and its related genes, this study aims to provide a comprehensive understanding of the effects of BPTF changes on the development, metastasis, and recurrence of the five cancers of interest.

3.9 Research Aims

The three main aims of the study are:

Aim 1:

To identify the alternative splicing patterns of the BPTF gene between normal and tumor patient samples for BLCA, BRCA, PRAD, THCA and UCEC and determine if they affect patient survival.

Aim 2:

To understand CpG island methylation patterns on the BPTF gene between normal and tumor samples through interrogating publicly available databases for the BLCA, BRCA, PRAD, THCA and UCEC cancer types.

Aim 3:

Utilize the gene ontology (biological processes, cellular components, and the molecular functions) of the genes that have been shown to have been co-expressed with BPTF in cancer and to analyze these pathways and genes across the five cancer types to aid in finding pathways that might commonly be affected and eventually targeted for a better therapy.

4. Methods and Materials

4.1 Databases and Data Acquisition

4.1.1. The Cancer Genome Atlas

The Cancer Genome Atlas (TCGA) is a cancer genomics research program that molecularly characterized over 20,000 primary cancers and their normal samples spanning 33 cancer types. This project was launched by the NIH in December 2005 and over time the data collected by this program generated over 2.5 PB of genomic, epigenomic, proteomic and transcriptomic publicly available data that can be used by the research community.

4.1.2. Splicing Database: TCGA SpliceSeq

The TCGA SpliceSeq Version 2.1 (https://bioinformatics.mdanderson.org/public-software/tcgaspliceseq/) is a Java application that allows for the investigation of alternative mRNA splicing patterns in data from high throughput mRNA sequencing studies (Ryan, M. C., et al., 2012). The source code and tool installation can be found at https://bioinformatics.mdanderson.org/public-software/spliceseq/installation/ for more details and the creators can be contacted using their help and support contact information. The splice graph maps the sequence reads and quantifies the inclusion level of each exon and the splice junction. The graphs are traversed to predict the protein isoforms that are likely to result from the observed exon junctions and the UniPort annotations are mapped as well. The splicing events are quantified through calculating the percent spliced in (PSI) values. This is a ratio of normalized read counts indicated over the total normalized reads for the event. The formula for the PSI value is the number of inclusive reads over the total of inclusive and exclusive reads of the exon. Therefore, when a PSI value of 0.5 is noted in the database it indicates that the exon is included in half of the expressed isoforms. The PSI value of 0 would indicate that the exon is never included in any of the transcripts and a PSI value of 1 would indicate that the exon is always included in all the transcripts.

4.1.3. Methylation Database: TCGA Wanderer

Using the TCGA database Wanderer (http://maplab.imppc.org/wanderer/index.html) was built as an intuitive tool allowing for real time access and visualization of DNA methylation profiles

(Díez-Villanueva, et.al., 2015). For a given gene the tool provides detailed individual β-values of all the HumanMethylation450 probes within or in the vicinity of the gene. The Wilcoxon rank sum test is performed on normal versus tumor provided there are more than two observations in each of the groups. The green colored probes indicate the CpG islands for the gene query and additionally the interactive tool marks statistically significant probes with asterisks. The database datasets are available for download and the datasets include β-values for normal samples and tumor sample with patient/sample ID. The Wanderer code repository is available through https://sourceforge.net/projects/tcga-wanderer/ and was created using R/Shiny and a PostgreSQL backend using an eXtreme programming software development methodology.

4.1.4. CBioPortal

CBioPortal version 5.3.5 (https://github.com/cBioPortal/) for Cancer Genomics was developed at the Memorial Sloan Kettering Cancer Center and is a public site that is hosted by the Center for Molecular Oncology at MSK. This open-access resource is available for interactive exploration of multidimensional cancer genomics datasets (Cerami, E., et al., 2012 and Gao, J.,et al, 2013). The portal itself supports and stores a diverse set of data like DNA copy number data, mRNA seq, microRNA expression data and de-identified clinical data for all types of cancer. The GitHub link for the datahub is https://github.com/cBioPortal/datahub and contact information for the team is available at the github directory link above.

4.1.5. DAVID

The Database for Annotation, Visualization and Integrated Discovery version2023q1 (https://david.ncifcrf.gov/) is a bioinformatics resource system for functional enrichment analysis, functional annotation and ID conversion of gene lists. This tool was released in 2003 and has gone through multiple updates throughout the years to include new information and methods (Díez-Villanueva, A., et.al., 2015) . The download client code is available for the different software packages at this link https://david.ncifcrf.gov/content.jsp?file=WS.html. This tool can provide high throughput functional annotation tools rarely found in other similar works. Specifically, the gene functional classification tool that can quickly group large lists of genes into functional groups and the functional annotation clustering tool that can condense redundant and heterogeneous annotation terms into groups to ease interpretation.

4.2. Methods

The data downloaded from these databases was cleaned and analyzed in R using different R packages which included but were not limited to tidyverse (Wickham, H., et.al., 2019) , patchwork (Pedersen, T., 2022), dplyr (Wickham, H., et.al., 2023), ggplot2 (Wickham, H, 2016), knitr (Xie, Y., 2023), gt_summary (Iannone, R., et.al., 2023) and janitor (Firke, S., 2023).

4.2.1. Splicing Data: TCGA SpliceSeq

The data was collected using the MD Anderson Cancer Center SpliceSeq web-based tool as a text file that included the PSI values and their patient data from the TCGA program, and the rest of the settings were left as defaults to get a good range of data (data that included more than 75% PSI values). Figure 7C illustrates the exons of BPTF as a whole. The alternative splicing events that were part of the datasets included alternative donor events of exon 23.2 as seen in Figure 7A. Exon skip of exons 5 and 6 as well as exon skip of exon 5 was noted as seen in Figure 7B.

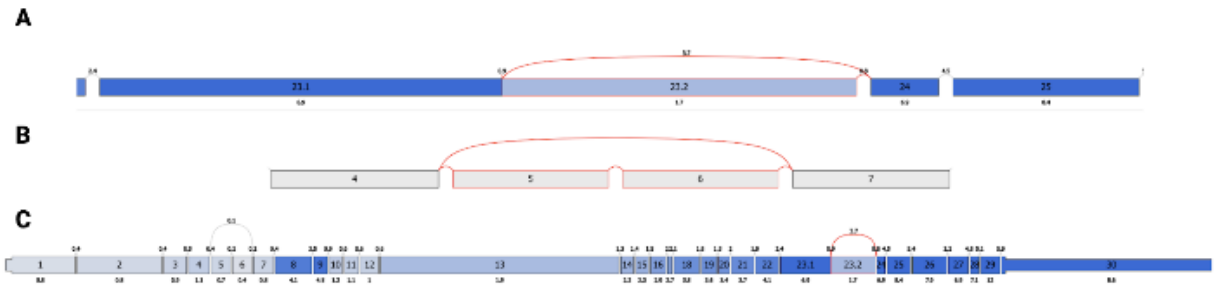

Figure 7: Exon view of BPTF and Alternative Splicing. *The exons of BPTF and each of the exons are shaded based on expression intensities.* ***A)*** *The exon alternative donor event of exon 23.2* ***B)*** *The exon skip event of exons 5:6 and exon 5.* ***C)*** *BPTF Exons 8, 9, 23.1, 24, 25, 26, 27, 28, 29, and 30 all showing intense expression compared to the other exons. (Created with BioRender.com).*

The dataset was then run through R and tbl_summary() function was used to obtain the data demographics for the datasets. Table S-1 illustrates the overall distribution of all the variables that were in the dataset and describes the missingness related to the raw data. Significance of the splicing patterns seen in the five cancer types was calculated using the Welch's Two-Sided T-Test, which is a test that compares the means of two independent groups and assumes that both groups of data are sampled from populations that follow a normal distribution but not the same variance. Previous research conducted significance testing using the Wilcoxon-rank sum test

which is a non-parametric alternative to the unpaired two sample t-test used for data that is not normally distributed. The normality of the normal and tumor PSI values was tested using the Shapiro-Wilk normality test. Results showed that the normal PSI samples of BLCA, PRAD, THCA, and UCEC were potentially normally distributed, while the normal PSI values of BRCA and the tumor PSI values of BLCA, BRCA, PRAD, THCA, and UCEC were not normally distributed. As a result, Wilcoxon-rank sum tests were performed and significance was recorded in the study to account for the observed differences in distribution as per the Shapiro-Wilk normality test.

Splicing Survival Analysis

The patient data for survival was gathered from CBioPortal's PanCancer datasets for each of the cancer types and the dataset was cleaned to keep overall survival, disease free survival and progression free survival status and months of the correlated SpliceSeq patients with the PSI values. Overall survival is the length of time from the date of diagnosis/start of treatment, it is usually a five year survival rate. Disease free survival is the length of time after primary treatment for a cancer ends that the patient survives without any signs or symptoms of the cancer. Finally, progression free survival is the length of time during and after the treatment of the disease that the patient lived with the disease, but it does not get worse.

The Kaplan Meier Survival curves were calculated using the survfit() function within the survminer and survival packages for each of the cancer types. The PSI values were placed into two categories: high and low inclusion. High inclusion had PSI values greater than 0.55, while the low inclusion PSI values included less than 0.45 values. High inclusion in terms of PSI meant that the exons during the event were spliced in 55% or more of the times in the transcripts seen in the patient samples. While low inclusion was indicative that the exons involved in the event were spliced in less than 45% of the time. Finally, the survival was conducted based on comparing the high and low inclusion of the exons in the tumor samples for the three survival types over a five year period noted in days. The p-values noted are based on the log-rank test where in the null hypothesis is that each of the strata has the same survival probability.

4.2.2. Methylation Data: TCGA Wanderer

Methylation expression for the tumor and their matched normal tissue samples were gathered from TCGA Wanderer. The HumanMethylation450 BeadChip probe measurements were given

as beta (β) values that ranged from 0 to 1, where a value of 0 indicated unmethylated probes while the value of 1 indicated fully methylated probes. The Shapiro-Wilk normality test was performed for the 16 probes within the normal and tumor data downloaded from TCGA Wanderer. The Shapiro-Wilk normality test informed on what statistical test would be used to accurately measure significance and to understand the distribution of small sample size in normal samples for each of the probes. The normality test indicated the tumor beta values for the sixteen probes were not normally distributed while the normal samples of the 16 probes indicated some probes had beta values that were potentially normally distributed. The beta values for the 16 probes that ranged throughout the BPTF gene were downloaded and were run through the Wilcoxon- Rank Sum Test that compared the means between the normal tissue β-values and the tumor tissue β-values. The ggplot2 package was used to illustrate the boxplots that show the overall distribution of the probe values between the normal and tumor samples. The significance shown in the boxplot uses the Wilcoxon test to conduct significant difference examination. Finally, the fold change of methylation was calculated between the mean of normal and tumor beta values.The fold change was calculated by dividing the average of normal beta values by the average of tumor beta values to indicate whether the probe registered more methylation in normal samples or in the tumor samples.

4.2.3. Gene Ontology DAVID

Initially, a gene list of co-expressed genes with BPTF was obtained from CBioPortal's Pan-Cancer database by querying BPTF as the gene of interest. This database provided data on BPTF as well as the genes correlated with it, determined by Spearman correlation calculated by the CBioPortal interface. This gene list was specific to each cancer type. As DAVID's functional clustering annotation tool has a gene limit intake of 2000, only the most significant genes ($p < 0.05$) were selected for analysis. The default selection of GOTERM_BP_DIRECT, GOTERM_CC_DIRECT, and GOTERM_MF_DIRECT were used to select the terms for biological processes, cellular components, and molecular function, respectively, which had a percentage of involved genes over 92% of the total, and the number of genes involved from the list provided was also listed. The DAVID tool provided a text file containing the gene ontology terms, their counts, percentage, P Values, False Discovery Rates (FDR) values, Fold Enrichment, Bonferroni, Benjamini values, and the genes involved in those processes. To compare highly related gene ontology terms between cancer types, FDR values were used to create a heatmap.

Due to the extremely small FDR values, the values were log scaled to provide a more precise view of the comparisons between cancer types and their highly related gene ontology terms.

5.Results

5.1. Alternative Splicing

The alternative splicing patterns between the PSI values of TCGA tumor patient samples and their adjacent normal patient samples were analyzed through conducting the Wilcoxon t-test shown in Table 1. In BRCA all three alternative splicing events were seen to be significantly different between the normal and tumor samples with alternate donor of exon 23.2 ($p = 1.31e^{-14}$), exon skip 5:6 ($p = 5.96e^{-38}$), and exon skip 5 ($p = 6.20e^{-35}$). In the cases of BLCA and UCEC only the alternate donor of exon 23.2 splicing event was found to have significant different between the PSI values of normal and tumor samples with BLCA ($p = 1.68e^{-06}$), and UCEC ($p = 3.53e^{-06}$). PRAD alternate donor of exon 23.2 ($p = 6.55e^{-09}$), exon skip 5:6 ($p = 7.24e^{-05}$), and THCA alternate donor ($p = 1.94e^{-09}$), exon skip ($p = 2.34e^{-06}$) were the alternative splicing events that showed significance based on a cutoff of $p < 0.05$.

Table 1: Wilcoxon T-Test of Significance. The significant BPTF splicing events for the different cancer types ($p < 0.05$) and NS indicates no significance.

Significant BPTF Splicing Types in Five Cancers

	Wilcoxon T-Test Pvalues		
Cancer	Alternate Donor 23.2	Exon Skip 5:6	Exon Skip 5
Breast Invasive Carcinoma	1.313e-14	5.968e-38	6.204e-35
Bladder Urothelial Carcinoma	1.685e-06	NS	NS
Prostate Adenocarcinoma	6.558e-09	7.242e-05	NS
Thyroid Carcinoma	1.943e-09	2.347e-06	NS
Uterine Corpus Endometrial Carcinoma	3.539e-06	NS	NS

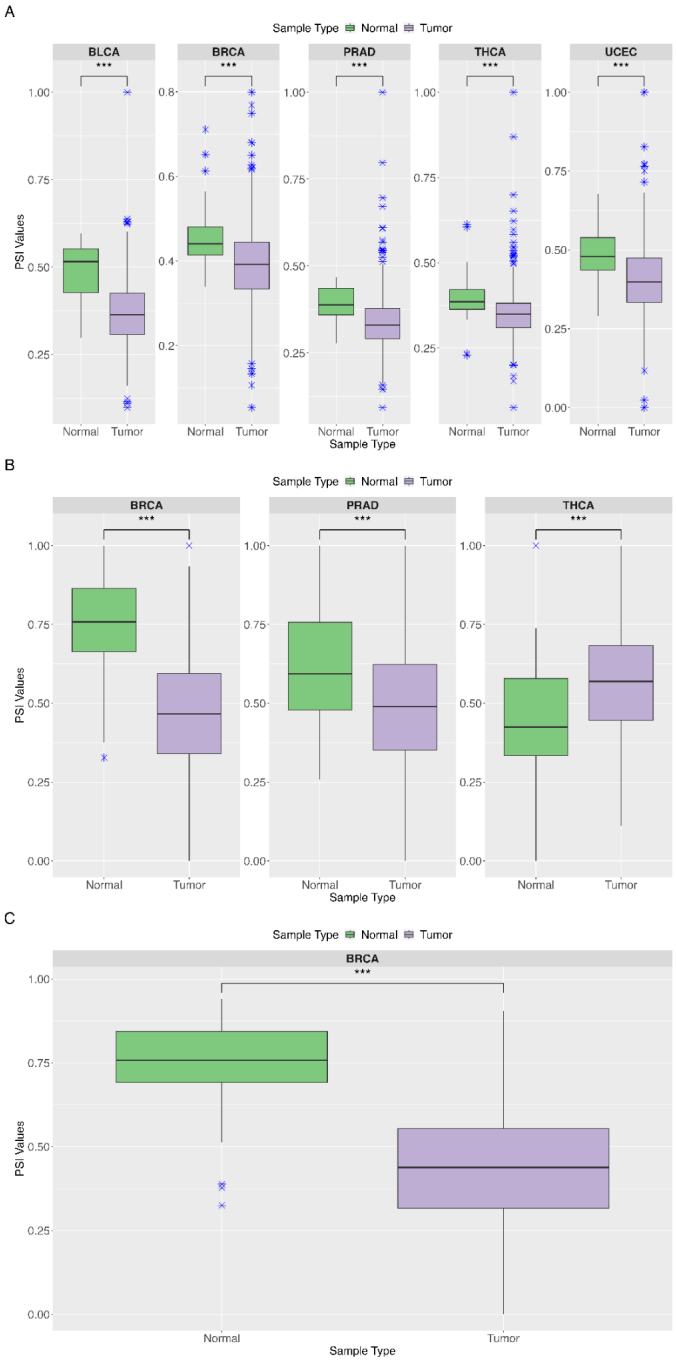

Figure 8: Significant Splicing Events PSI Comparison for Five Cancers. *Tumor and their adjacent normal samples PSI values compared for significant differences (blue asterisks indicate outlier values).* ***A)*** *Alternate Donor of Exon 23.2.* ***B)*** *Exon Skip of Exons 5 and 6.* ***C)*** *Exon Skip of Exon 5.*

Survival Analysis

Based on the significance analysis conducted between the PSI-values of normal and tumor samples further analysis was conducted to potentially understand if alternative splicing patterns affect the survival of the tumor positive patients over a five-year period. The PSI values placed into the high and low inclusion categories yielded tumor patient samples and their respective counts are noted in Table 2 below. The low inclusion category consistently had a good amount of data for the analysis while the high inclusion categories for the alternate donor event for exon 23.2 had a number of tumor samples < 30 for BLCA, PRAD and THCA cancer types.

Table 2: Tumor Sample Counts for Survival. The sample counts for each alternative splicing event for the two different PSI levels.

Number Tumor Samples in High and Low Inclusion

	PSI Inclusion Levels	
Cancer	High Inclusion	Low Inclusion
BLCA AD 23.2	19	294
BRCA AD 23.2	38	644
PRAD AD 23.2	10	372
THCA AD 23.2	9	374
UCEC AD 23.2	43	298
BRCA Exon Skip 5:6	331	342
PRAD Exon Skip 5:6	210	132
THCA Exon Skip 5:6	243	92
BRCA Exon Skip 5	361	333

The survival curves in the alternate donor event of exon 23.2 do not yield any significant difference in survival of the patients with low inclusion of the exon or high inclusion of the exon. The log rank statistical p-values seen in Figure 9 for BLCA were $p = 0.64$, $p = 0.2$ and $p = 0.11$ for overall, disease free and progression free survival respectively. Figure 10 for BRCA noted log rank $p = 0.65$, $p = 0.38$ and $p = 0.46$ for overall, disease free and progression free survival. The risk tables shown in each of the subfigures (A, B, C) indicated the number of patients at each interval (in days) that were present for the strata (high and low inclusion).

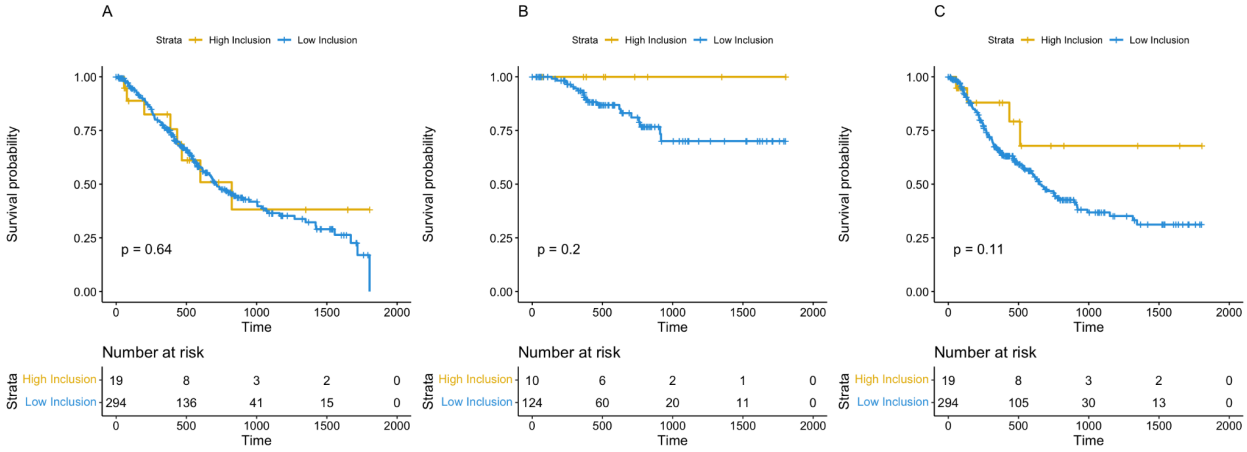

Figure 9: BPTF Alternate Donor Exon 23.2 Survival Curves in BLCA.*Comparing tumor patients survival by the relative inclusion of exon 23.2.* ***A)*** *Overall Survival* ***B)*** *Disease Free Survival* ***C)*** *Progression Free Survival*

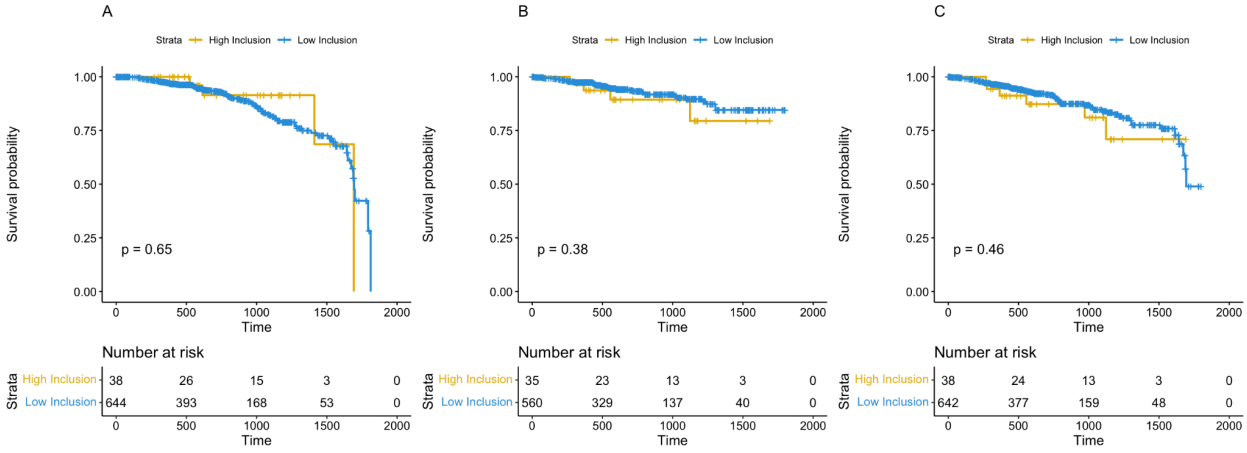

Figure 10: BPTF Alternate Donor Exon 23.2 Survival Curves in BRCA. *Comparing tumor patients survival by the relative inclusion of exon 23.2.* ***A)*** *Overall Survival* ***B)*** *Disease Free Survival* ***C)*** *Progression Free Survival*

Figures 11 and 12 for cancers PRAD and THCA the survival difference between the high and low inclusion of exon 23.2 between tumor patient samples indicated log rank p-values for overall, disease free and progression free survival of p =0.67, p = 0.32 and p = 0.15 respectively for PRAD and p = 0.61, p = 0.4, and p =0.92 for THCA. Figure 13 survival curves for UCEC's low and high inclusion survival differences for the tumor patient samples indicated log rank p-values of p = 0.66,p = 0.23 and p = 0.53 for the three survival types.

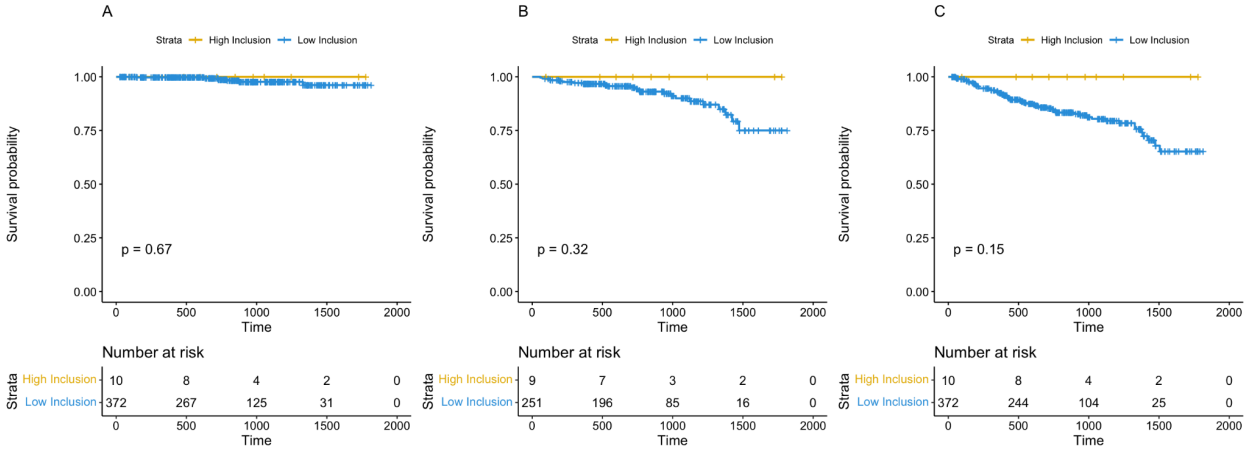

Figure 11: BPTF Alternate Donor Exon 23.2 Survival Curves in PRAD. *Comparing tumor patients survival by the relative inclusion of exon 23.2.* ***A)*** *Overall Survival* ***B)*** *Disease Free Survival* ***C)*** *Progression Free Survival*

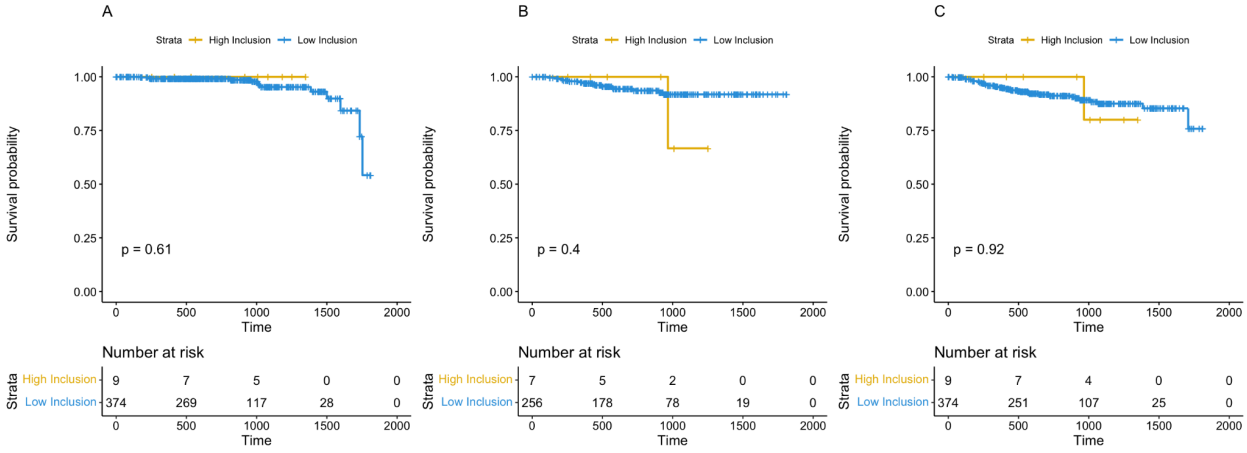

Figure 12: BPTF Alternate Donor Exon 23.2 Survival Curves in THCA. *Comparing tumor patients survival by the relative inclusion of exon 23.2.* ***A)*** *Overall Survival* ***B)*** *Disease Free Survival* ***C)*** *Progression Free Survival*

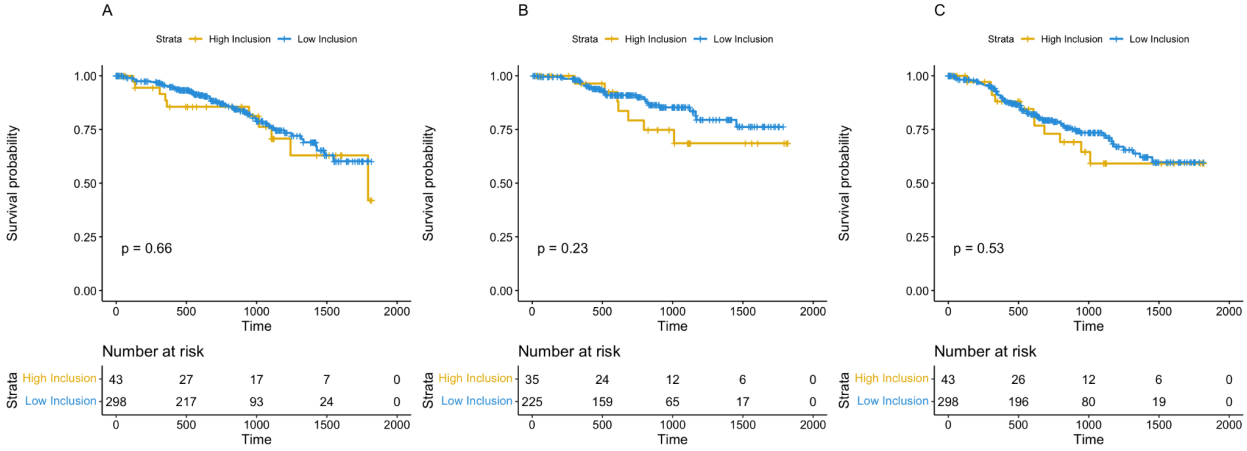

Figure 13: BPTF Alternate Donor Exon 23.2 Survival Curves in UCEC. *Comparing tumor patients survival by the relative inclusion of exon 23.2.* ***A)*** *Overall Survival* ***B)*** *Disease Free Survival* ***C)*** *Progression Free Survival*

Figure 14, 15 and 16 for BRCA, PRAD and THCA noted log rank p = 0.2, p = 0.60 and p = 0.097 for BRCA (Figure 14), p = 0.54, p = 0.50, and p = 0.41 for PRAD (Figure 15) and p = 0.24, p = 0.39 and p = 0.46 for THCA (Figure 16). These p-values are for the survival difference between the high and low inclusion of exons 5 and 6 in overall, disease free and progression free survival respectively for the tumor patient samples. The risk tables shown in each of the subfigures (A, B, C) indicated the number of patients at each interval (in days) that were present for the strata (high and low inclusion).

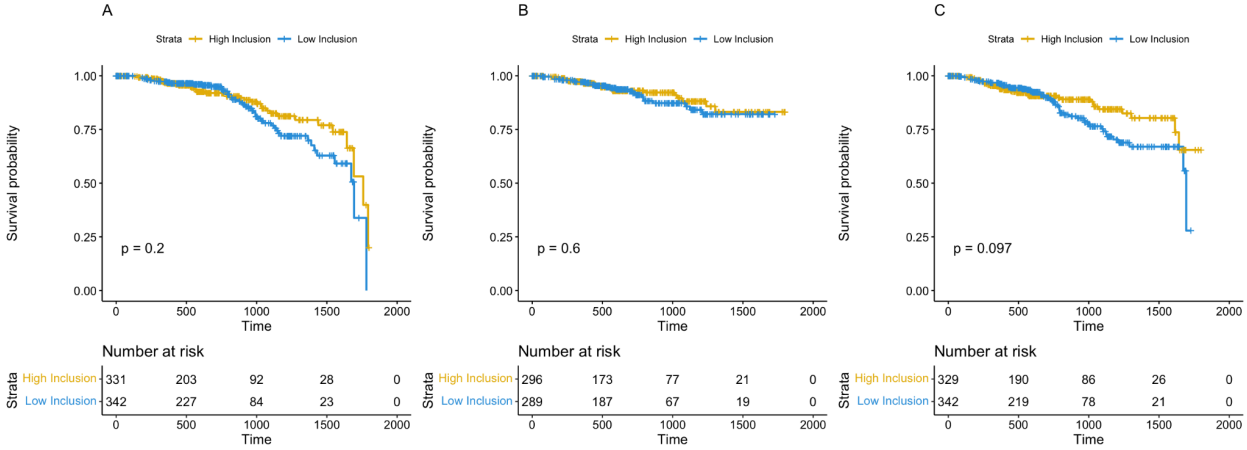

Figure 14: BPTF Exon Skip Exons 5 and 6 Survival Curves in BRCA. *Comparing the tumor patient survival by the relative inclusion of exons 5 and 6.* ***A)*** *Overall Survival* ***B)*** *Disease Free Survival* ***C)*** *Progression Free Survival*

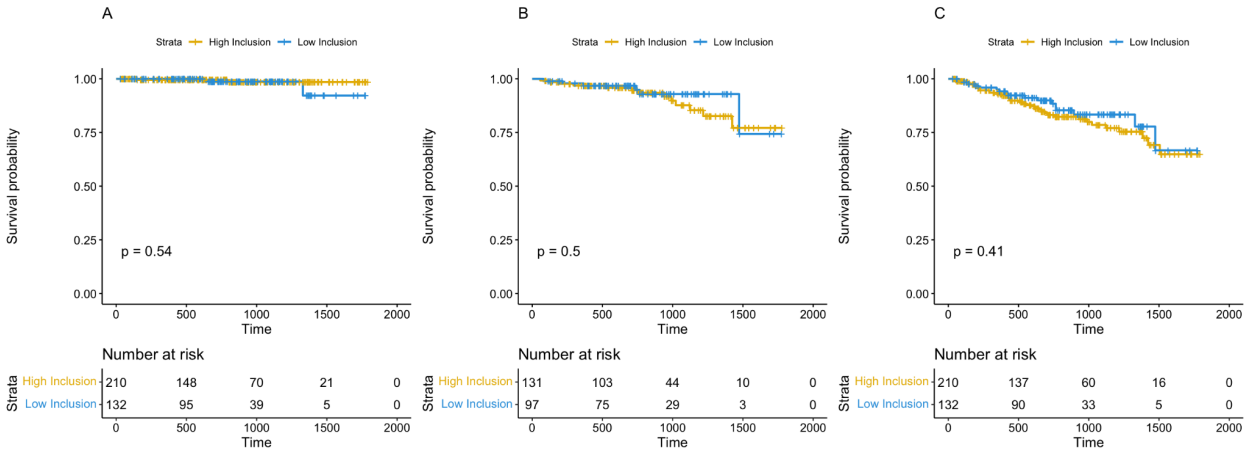

Figure 15: BPTF Exon Skip Exons 5 and 6 Survival Curves in PRAD. *Comparing the tumor patient survival by the relative inclusion of exons 5 and 6.* ***A)*** *Overall Survival* ***B)*** *Disease Free Survival* ***C)*** *Progression Free Survival*

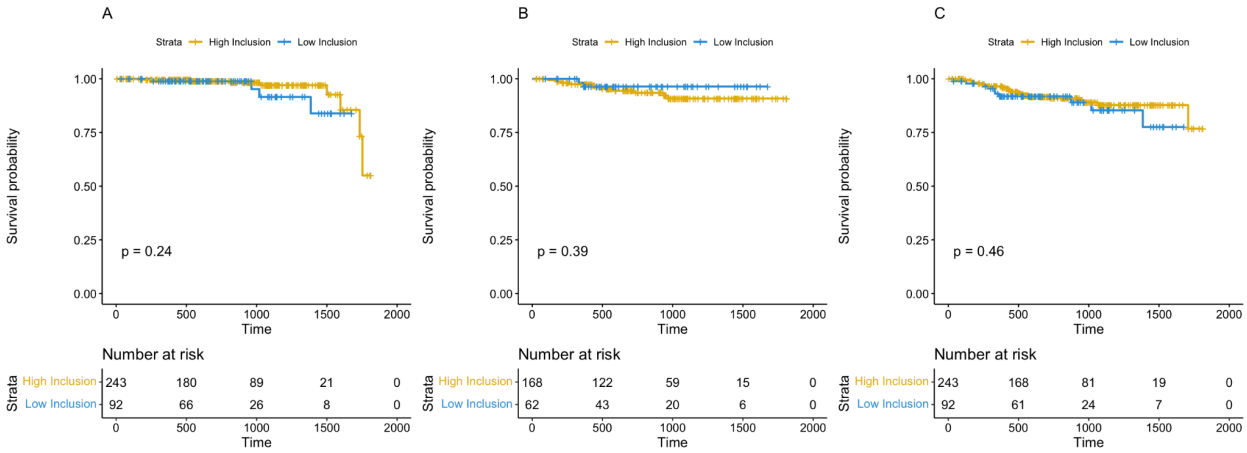

Figure 16: BPTF Exon Skip Exons 5 and 6 Survival Curves in THCA. *Comparing the tumor patient survival by the relative inclusion of exons 5 and 6.* ***A)*** *Overall Survival* ***B)*** *Disease Free Survival* ***C)*** *Progression Free Survival*

The log rank p-values for the survival difference between the high and low inclusion of exon 5 in overall, disease free and progression free survival respectively for the BRCA tumor patient samples were p = 0.59, p = 0.96 and p = 0.34 . The risk tables shown in each of the subfigures (A, B, C) indicated the number of patients at each interval (in days) that were present for the strata (high and low inclusion).

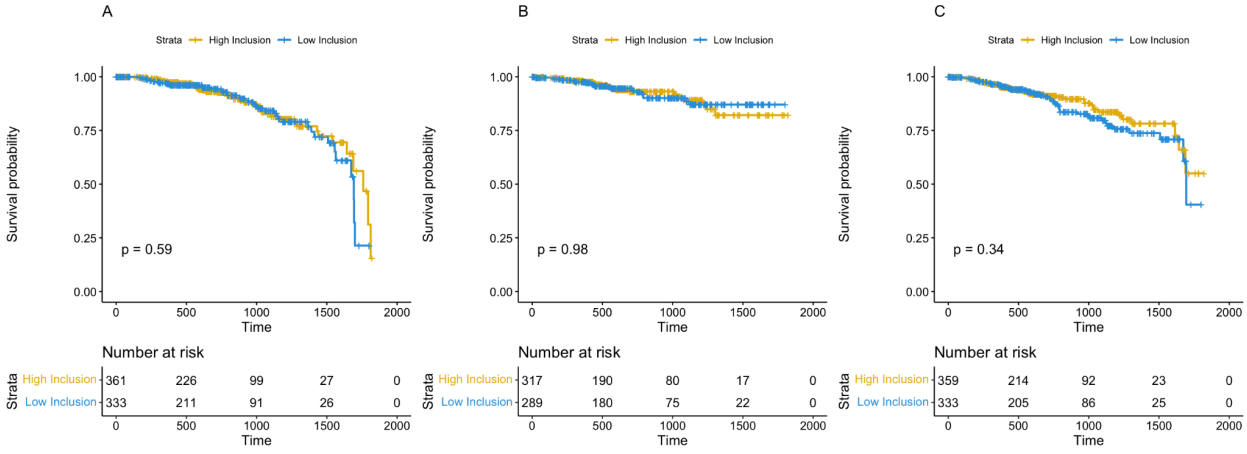

Figure 17: BPTF Exon Skip Exon 5 Survival Curves in BRCA. *Comparing the tumor patient survival by the relative inclusion of exon 5.* ***A)*** *Overall Survival* ***B)*** *Disease Free Survival* ***C)*** *Progression Free Survival*

5.2. CpG/DNA Methylation

There were 16 HumanMethylation450 probes noted on the BPTF gene, most of which were at the beginning of the BPTF gene as seen in Figure 18. This figure shows the placement of these probes where the methylation beta values were recorded. Significant difference between the tumor and their adjacent normal samples was conducted first by using the Welch's test on all the 16 probes. Probes cg22443205 and cg16028064 which are noted to be the CpG islands of the gene are of main interest, however probes cg12398397, cg22972262, cg16435601 and cg03215657 are all of interest because they are located at the beginning of the gene and would be able to inform on the ability to access the gene for translation to occur.

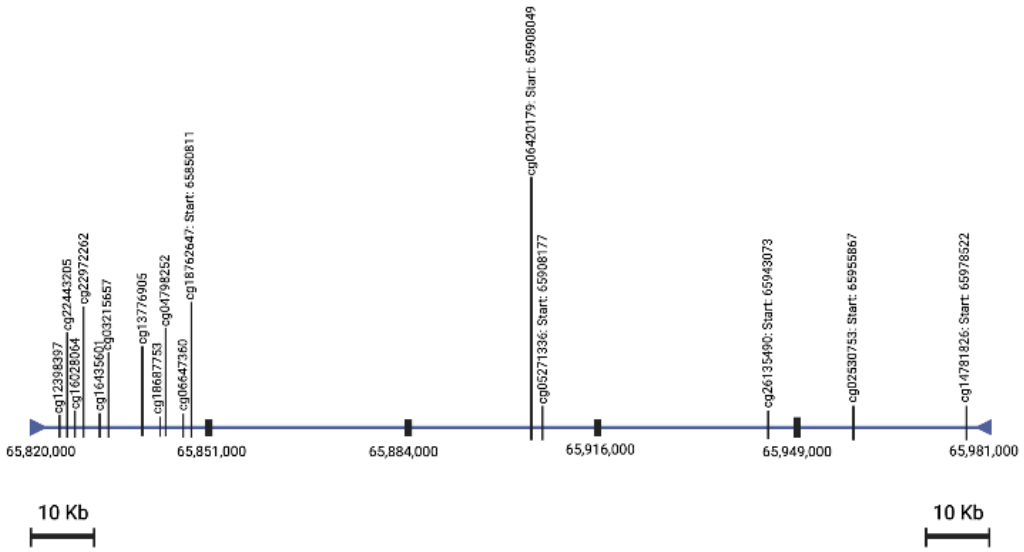

Figure 18: HumanMethylation450 probes on the BPTF gene region. *A scaled BPTF gene with the methylation probe locations on the gene. Probes cg22443205 and cg16028064 are noted to be the CpG islands for the gene (Created with BioRender.com).*

In BLCA there were many probes out of the sixteen of interest that indicated significant differences between the normal β-values and tumor β-values (Figure 19 and Table-3). As seen in Table S-3 BLCA probes cg12398397 ($p = 8.82e^{-06}$), cg22443205 ($p = 3.87e^{-04}$), cg16435601 ($p = 2.93e^{-04}$), and cg03215657($p = 3.85e^{-04}$) that are located in/near CpG methylation islands were found to be significant and their fold change was calculated to 1.333, 1.088 and 1.396, 1.062 respectively. The fold changes indicate that overall, the normal BPTF gene had a slightly higher mean β-value than tumor cells which in turn informs that normal BPTF gene is

slightly more methylated than the tumor BPTF gene in those locations. The other probes downstream of the gene (in order of occurrence) that were of noted significance were cg06647360($p = 5.7e^{-04}$), cg18762647($p = 2.06e^{-05}$), cg06420179($p = 9.2e^{-03}$), cg05271336($p = 3.27e^{-03}$), cg02530753($p = 4.53e^{-02}$) and cg14781826($p = 1.37e^{-04}$). These downstream probes had a fold change of about 1.008 to 1.04 however, the cg14781826 probe had a fold change of 0.857 as seen in Table S-3.

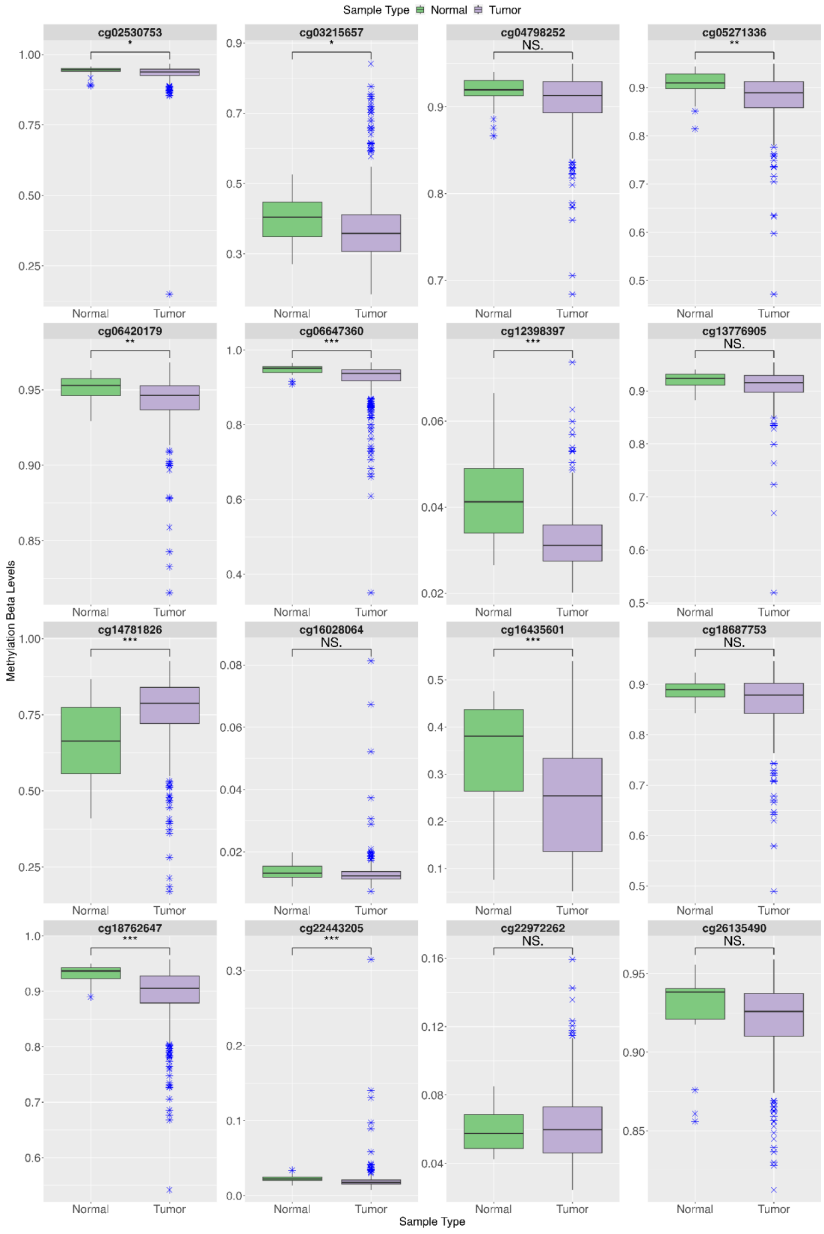

Figure 19: BLCA β-values Distribution Comparison between Normal and Tumor. *The sixteen probes compared for methylation beta levels (range 0 to 1). The stars represent the level of*

significance based on the Wilcoxon test and NS indicates not significant. The blue stars are noted to be beta value outliers.

The methylation analysis in BRCA illustrated through Figure 20 and Table S-3 indicated that probes cg12398397 and cg16028064 had significant differences in methylation between normal and tumor BPTF gene. The cg16028064 probe is located on the CpG island and had a p-value of $2.94e^{-13}$ and a fold change of 0.208 which indicates that this region was more methylated in the tumor genes overall compared to the normal gene. The cg12398397 probe that is located before the CpG island has a p-value of $2.74e^{-12}$ and a fold change of 1.607 which indicates that the mean beta values had higher values in the normal gene compared to tumor gene beta values. This would that signify the normal gene was slightly more methylated overall in this position compared to the tumor gene transcript. The other downstream probes that were significant were cg06647360($p = 7.36e^{-03}$), cg06420179($p = 9.4e^{-03}$), cg05271336($p = 4.47e^{-02}$), cg02530753($p = 1.72e^{-02}$) and cg14781826($p = 2.47e^{-14}$) and their fold changes ranged from 0.77 to 1.008 as seen in Table S-3.

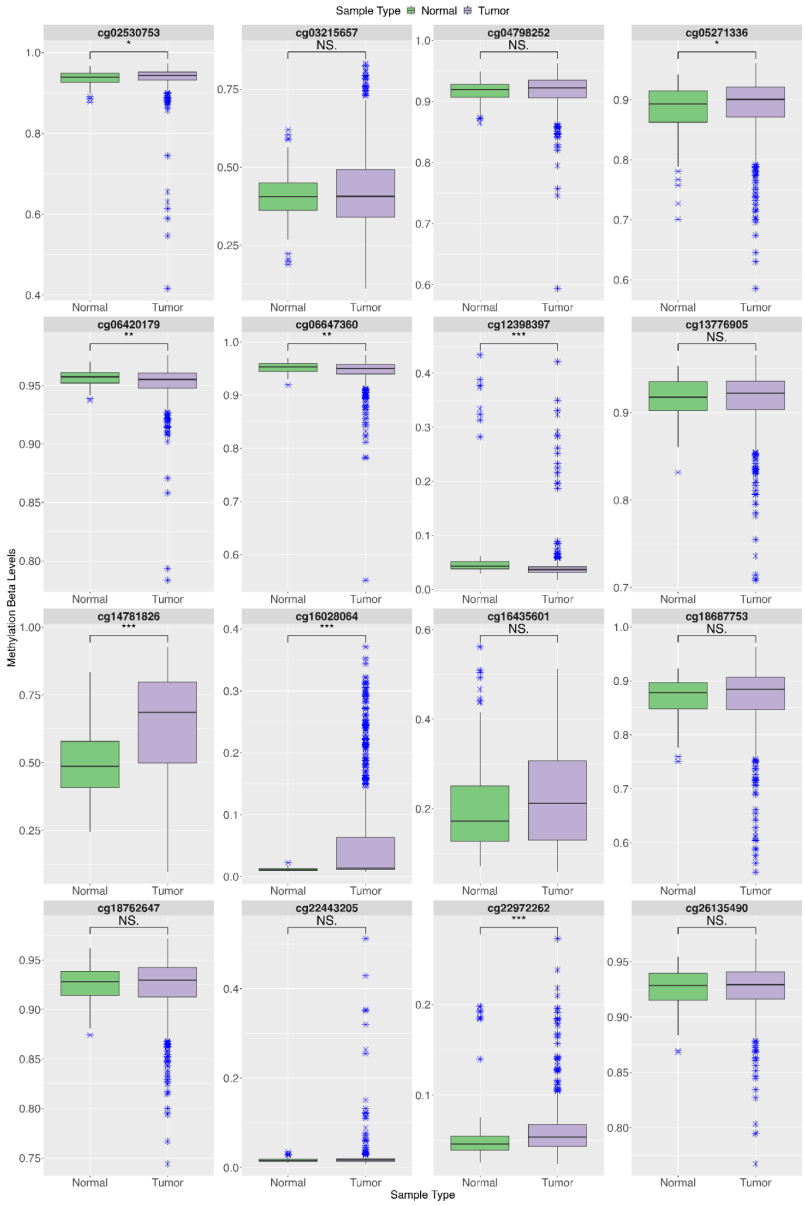

Figure 20: ***BRCA β-values Distribution Comparison between Normal and Tumor.*** *The sixteen probes compared for methylation beta levels (range 0 to 1). The stars represent the level of significance based on the Wilcoxon test and NS indicates not significant. The blue stars are noted to be beta value outliers.*

Within the PRAD methylation analysis there was slight significance noted for probes cg12398397 ($p = 2.87e^{-03}$) with a fold change of 1.04 which indicated almost the same mean methylation between the normal and the tumor samples. The downstream probes that had noted significance in PRAD for difference between normal and tumor samples were cg13776905 ($p = 9.51e^{-04}$), cg18687753($p = 1.29e^{-02}$), cg04798252($p = 6.8e^{-06}$),

cg06647360($p = 2.9e^{-03}$), cg18762647($p = 2.3e^{-10}$), cg06420179($p = 3.6e^{-03}$), cg05271336($p = 1.16e^{-05}$), cg26135490($p = 8.63e^{-06}$), cg02530753($p = 2.66e^{-07}$) and cg14781826 ($p = 7.56e^{-09}$). Their fold changes for these probes ranged from 0.745 and 1.039 as seen in Table S-4.

In THCA methylation analysis there were three probes in/near the CpG islands the first being cg12398397 ($p = 6.36e^{-04}$) with a fold change of 1.05, and the CpG island probes cg22443205 ($p = 1.41e^{-03}$) with a fold change of 0.909 and cg16028064 ($p = 1.83e^{-02}$) with a fold change of 1.04. The downstream probes that had noted significance in THCA for difference between the normal and tumor samples beta values were cg13776905 ($p = 9.66e^{-03}$), cg05271336($p = 4.73e^{-02}$), cg26135490 ($p = 2.95e^{-02}$), cg02530753($p = 2.17e^{-02}$) and cg14781826 ($p = 7.71e^{-05}$). The fold change for the downstream probe samples as seen in Table S-4 were around 0.99.

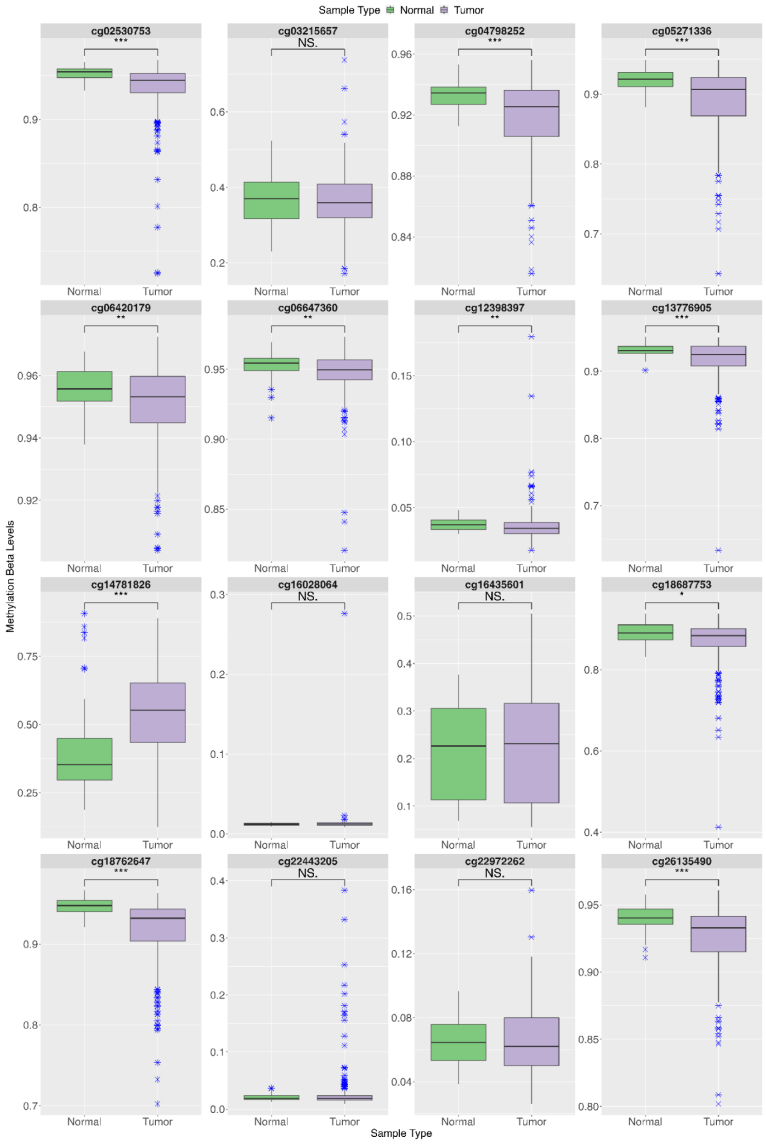

Figure 21: PRAD β-values Distribution Comparison between Normal and Tumor. *The sixteen probes compared for methylation beta levels (range 0 to 1). The stars represent the level of significance based on the Wilcoxon test and NS indicates not significant. The blue stars are noted to be beta value outliers.*

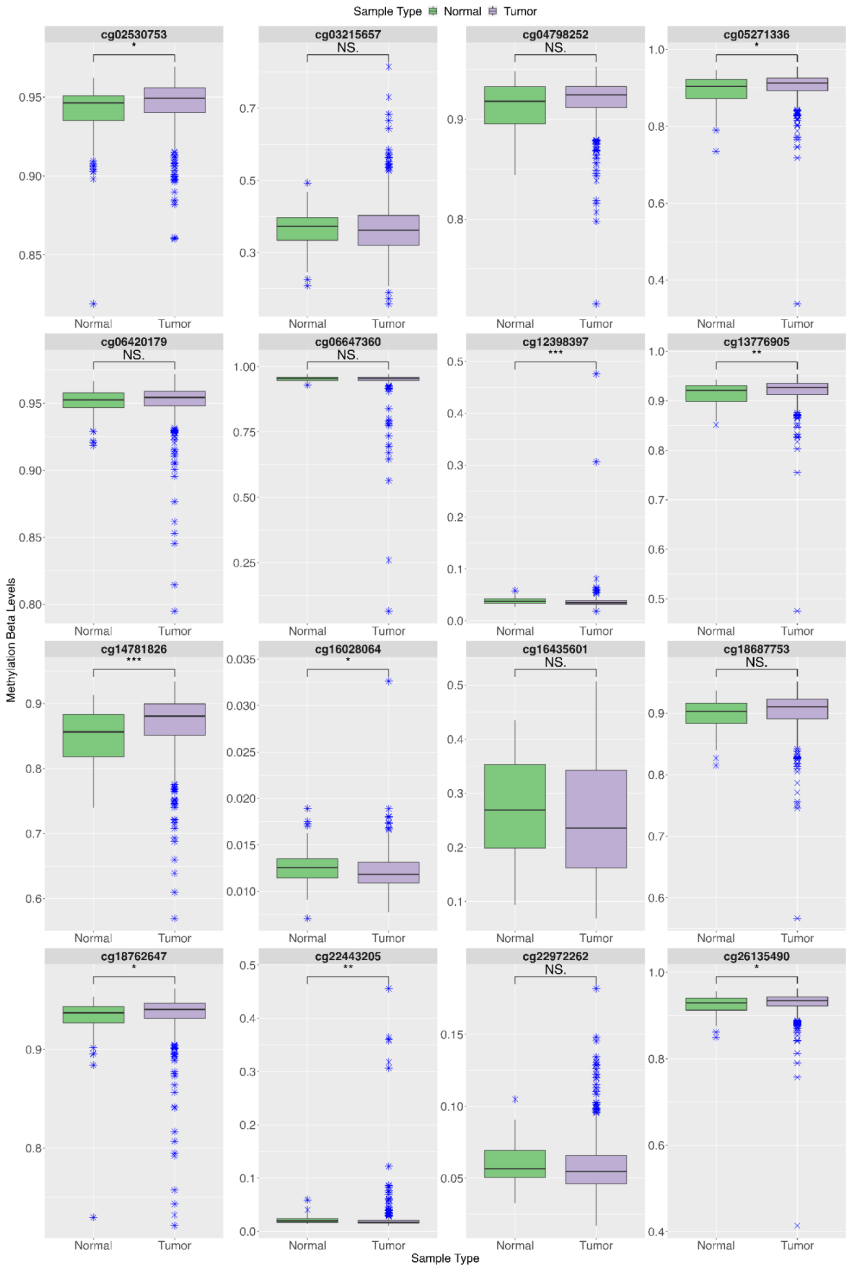

Figure 22: THCA β-values Distribution Comparison between Normal and Tumor. *The sixteen probes compared for methylation beta levels (range 0 to 1). The stars represent the level of significance based on the Wilcoxon test and NS indicates not significant. The blue stars are noted to be beta value outliers.*

The methylation probes cg12398397, g22443205 and cg16435601 were found to be significantly different with p-values of $7.89e^{-09}$, $1.40e^{-12}$, and $4.03e^{-05}$ respectively in UCEC methylation analysis. Their respective fold changes were calculated to be 1.26, 1.334, and 1.226.

These values inform that the overall mean methylation beta values were slightly greater in normal than in tumor BPTF gene. The downstream probes that had noted significance in UCEC for difference between normal and tumor samples were cg06647360($p = 2.19e^{-11}$), cg18762647($p = 3.06e^{-07}$), cg02530753($p = 4.08e^{-03}$) and cg14781826 ($p = 1.75e^{-14}$). The fold changes noted for these probes ranged from 1.003 to 1.03 except for the cg14781826 which had a fold change value of 0.5614 (Table S-5).

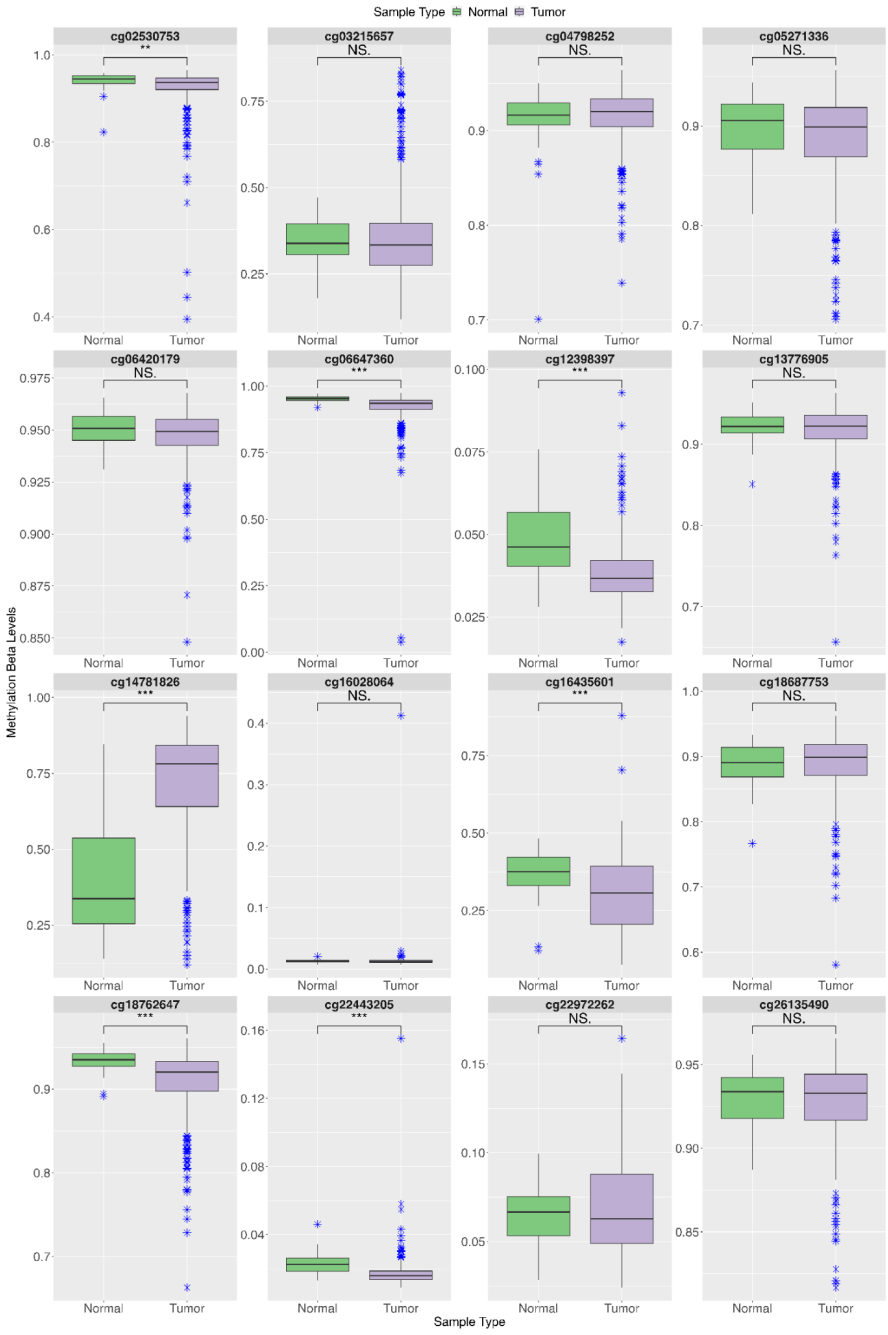

Figure 23: UCEC β-values Distribution Comparison between Normal and Tumor. *The sixteen probes compared for methylation beta levels (range 0 to 1). The stars represent the level of significance based on the Wilcoxon test and NS indicates not significant. The blue stars are noted to be beta value outliers.*

5.3. Gene Ontology

The positively and negatively co-expressed genes with BPTF in the tumor cells gave insight into the pathways and co-expression profiles that aid in better understanding the BPTF gene. The first

aspect that was checked was the biological processes as seen in Figure 24A where their FDR values were converted into log space and then visualized through a heatmap. The same concept was used for the other two aspects of the cellular component and the molecular function seen in Figure 24B and Figure 24C respectively. The FDR values for each of the three aspects are noted in Table S-6, Table S-7 and Table S-8. Figure 24A indicates that all these biological processes are common between the five different cancer types and are highly significant (FDR < 0.05). However, the processes of translation, regulation of transcription DNA-templated and regulation of transcription from RNA polymerase 2 promoter have the most significant FDR values related to the biological processes in all five cancers of interest. The translation biological process for BLCA (FDR = $1.20e^{-05}$), BRCA (FDR = $2.77e^{-13}$), PRAD (FDR = $2.84e^{-02}$), THCA (FDR = $9.01e^{-23}$) and UCEC (FDR = $1.60e^{-06}$) all had common genes that had been correlated to BPTF expression. Some of the genes included *EGFR*, *MRPL's* (multiple variations) and *RPL's*. The Spearman's correlation noted for the EGFR gene for BLCA(0.251), BRCA(0.116), PRAD(0.672), THCA(0.666) and UCEC(0.528).The regulation of transcription DNA-templated had FDR values of $4.24e^{-15}$, $2.63e^{-07}$, $5.06e^{-11}$, $8.63e^{-10}$, $1.54e^{-07}$ and regulation of transcription from RNA polymerase 2 promoter had FDR values of $1.34e^{-17}$, $1.20e^{-07}$, $3.55e^{-09}$, $8.34e^{-12}$, and $1.98e^{-09}$ for BLCA, BRCA, PRAD, THCA and UCEC respectively. A few genes related to these processes along with common genes seen in the translation biological process were *ZNF, TRRAP, NSD's, DMAP1 ATF1, ATF2* and *SMARCC1/2*. The Spearman's correlation noted for the *ATF1/ATF2* gene for BLCA(0.324/0.619), BRCA(0.388/0.653), PRAD(0.666/0.800), THCA(0.515/0.810) and UCEC(0.465/0.723). The Spearman's correlation noted for the *SMARCC1/2* gene for BLCA(0.468/0.389), BRCA(0.355/0.255), PRAD(0.646/0.255), THCA(0.720/0.347) and UCEC(0.558/0.130). Spearman's correlation noted for the *TRRAP* gene to *BPTF* for BLCA(0.670), BRCA(0.514), PRAD(0.783), THCA(0.807) and UCEC(0.610). *DMAP1* spearman's correlation to BPTF for BLCA is -0.336, BRCA is -0.489, PRAD is -0.586, THCA is -0.484 and for UCEC it is -0.467.

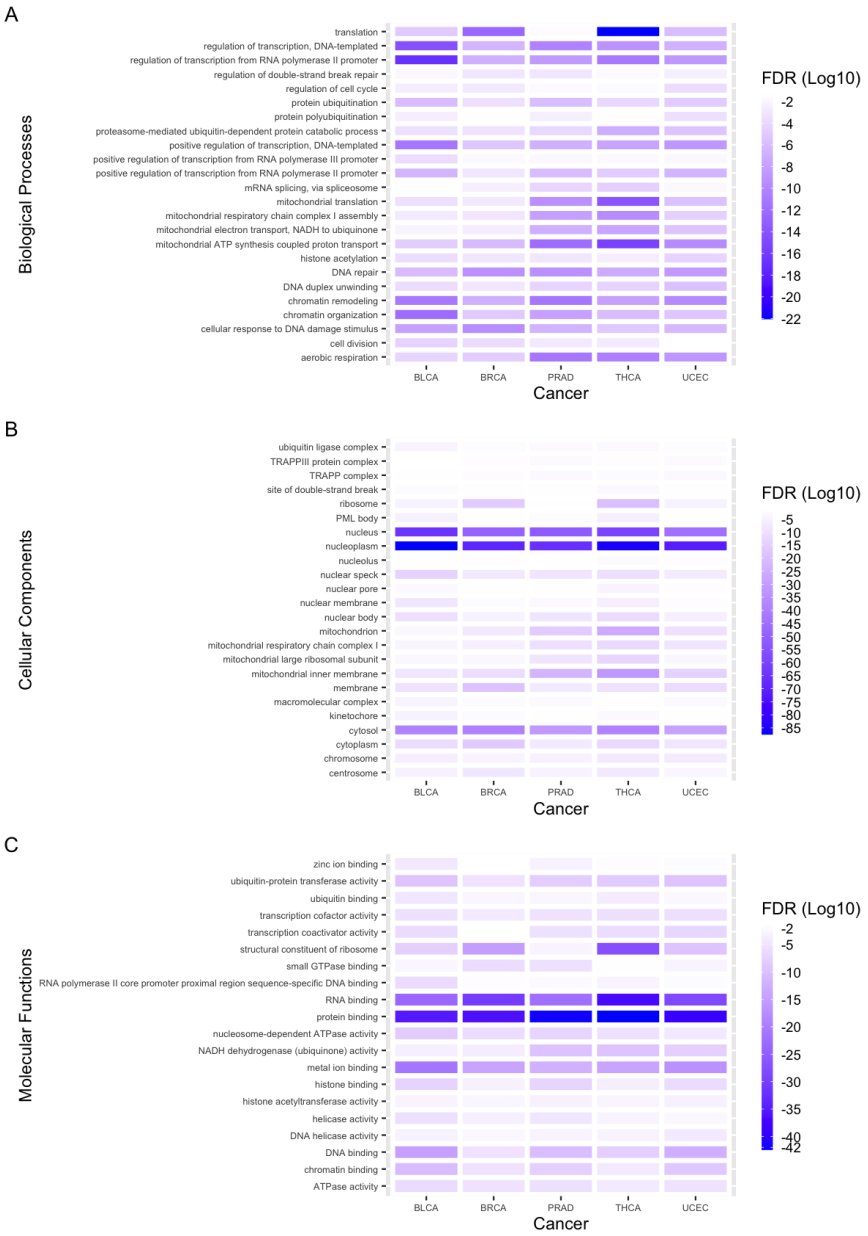

Figure 24: Gene Ontology Terms and Log FDR values for BPTF Correlated Genes. A) *Heatmap of Biological Process terms of five cancers related to BPTF and its co-expressed genes.* ***B)*** *Heatmap of Cellular Component terms of five cancers related to BPTF and its co-expressed genes.* ***C)*** *Heatmap of Molecular Function terms of five cancers related to BPTF and its co-expressed genes.*

The cellular components section in Figure 24B shows a clear high significance of the nucleus and the nucleoplasm with FDR log values. The nucleus FDR values for BLCA, BRCA, PRAD, THCA and UCEC were $1.02e^{-65}$, $2.08e^{-50}$, $1.32e^{-52}$, $1.77e^{-60}$, and $2.26e^{-45}$

respectively. As for the nucleoplasm the values for those cancers were $3.35e^{-88}$, $1.00e^{-69}$, $8.57e^{-67}$, $4.10e^{-86}$, and $4.77e^{-72}$. A few related genes to these cellular components included *MYCBP2, MAPK1, BRCA2, DAPK3* and *RHOC*. Spearman correlation noted for the *MYCBP2*, *MAPK1, BRCA2, DAPK3* and *RHOC* genes for BLCA (0.521, 0.406, 0.570, -0.426, -0.445 respectively), BRCA(0.487, 0.411, 0.486, -0.425, -0.370 respectively), PRAD(0.730, 0.601, 0.545, -0.577, -0.610 respectively), THCA(0.819, 0.496, 0.496, -0.536, -0.701 respectively) and UCEC(0.745, 0.518, 0.648, -0.346, -0.396 respectively).

The molecular function (Figure 24C) is the final aspect of the gene ontology analysis and the most significant functions in all five cancer types were RNA binding and protein binding. The FDR values for RNA binding were BLCA (FDR = $3.20e^{-24}$), BRCA (FDR = $4.29e^{-31}$), PRAD (FDR = $7.41e^{-23}$), THCA (FDR = $2.00e^{-37}$) and UCEC (FDR = $5.70e^{-29}$) and for the protein binding the values were $5.47e^{-36}$, $7.06e^{-37}$, $1.11e^{-42}$, $3.88e^{-43}$ and $1.31e^{-39}$ respectively. A few of the common genes were *MYL, TOP1, RBBP6, TUT4,* and *AKAP1*. Spearman correlation noted for the *TOP1, RBBP6, TUT4,* and *AKAP1* genes for BLCA (0.550, 0.526, 0.596, 0.487 respectively), BRCA(0.523, 0.446, 0.295, 0.348 respectively), PRAD(0.550, 0.676, 0.721, 0.327 respectively), THCA(0.485, 0.450, 0.797, 0.258 respectively) and UCEC(0.469, 0.447, 0.672, 0.410 respectively).

6.Discussion

Alternative Splicing

As alternative splicing has been related to the biological processes of multiple malignancies, understanding the significant splicing events occurring of the BPTF gene and conducting survival analysis could serve as a potential prognostic indicator. The three main splicing events that were noted of high significance were the alternate donor (5' prime donor) of exon 23.2, exon skip of both exons 5 and 6 and then an exon skip of just exon 5. The alternate donor site was noted to be significant in all five cancer types while the exon skip 5:6 was only significant in BRCA, PRAD and THCA cancers and further exon skip 5 was only significant in BRCA. Exon 23.2 has no noted functional significance while exons 5 and 6 have been noted to have interactions with KEAP1. KEAP1 also known as the Kelch-like ECH-associated protein 1 is a component of the E3- ubiquitin ligase complex that controls the stability and accumulation of

NRF2 (nuclear factor (erythroid-derived 2)-like 2). Cancer cells acquire malignancy by perverting NRF2. In normal conditions NRF2 levels are very low but it can accumulate if it escapes the KEAP1 trapping mechanism (Taguchi, K., & Yamamoto, M., 2017). The pathway through which this interaction is occurring and the binding motif between the KEAP1 gene and the BPTF gene has yet to be determined. However, in Alzheimer's disease the FAC1 (Fetal Alz-50 reactive clone protein) which is an alternatively spliced gene without the bromodomain of BPTF that spans from exon 28 through exon 29 has been noted to in its full length interact with human KEAP1 proteins. There is noted competition between the FAC1 and NRF2 for binding human KEAP1 that indicates that there is an interaction between the three genes and their proteins. As for exon 23.2 the alternate donor site does shorten the overall transcript that gets translated into protein. This shortening of the gene might lead to downstream effects in binding with other proteins and structural competency might be compromised. Furthermore the zinc-finger and Ca2+ interacting regions of BPTF which are located through exons 26 and 27 might be affected structurally as the proteins are made.

The differences in splicing and the proteins that are eventually created play a crucial role in understanding how other interacting proteins might be affected and based on these understandings one can potentially edit the splicing mechanisms to produced only beneficial transcripts that do not lead to the overproduction of the harmful splicing patterned BPTF. An analysis of protein structure based on the splicing events, which have been noted to be significantly different, and their interactive locations with other cofactors and co-expressed genes might aid in furthering the understanding of how these splicing events impact the transcription and translation of the BPTF gene.

The TCGA SpliceSeq dataset itself had many flaws as to how much of the data was presented and the amount of information that was left out that could have aided in understanding variances seen in splicing events based on gender, age, weight and other epigenetic factors that have been known to play a role in initiation and progression of cancer. Table S-1 contains the demographics of the dataset downloaded from the TCGA SpliceSeq. The ratio of the tumor to normal patient samples and the distribution differences seen led to difficulty in further analysis. For example, the overall missingness and involvement of null values within the PSI column were dealt with by omitting those values leading to the loss of some data. A more accurate result on the significance

of the splicing event can be achieved through the involvement of additional normal patient samples proportional to the sample size of cancer patients.

Splicing Survival Analysis

In order to glimpse at the potential effects on patient survival of these splicing events occurring more often or less often in the transcripts tested from tumor patients the PSI values were split into two categories. In the first case of the alternate donor of 23.2 high inclusion meant that the inclusion of 23.2 was seen more often in read transcripts and therefore the PSI values were between 0.55 and 1. While, the low inclusion of the exon 23.2 was indicative of PSI values less than 0.45 which meant that there were more read transcript that included the event of alternate donor site for exon 23.2 (ie. missing the 23.2 and connecting the exon 23.1 3' prime end to exon 24 5' prime). The patient survival difference over 5 years compared the tumor patients with high inclusion of exon(s) to the tumor patients with low inclusion exon(s) in the three alternative splicing events seen in this study.

The alternate donor event of exon 23.2, exon skip event of exons 5 and 6 and finally the exon skip event of just exon 5 indicated no significant difference at a 0.05 p-value cutoff in the tumor patient survival between the two groups of high and low inclusion. However, at a p-value of 0.1 cutoff the BRCA progression free exon skip of exons 5 and 6 yields slight significant difference in survival between the tumor patients with high inclusion of the exons and patients with low inclusion of the exons at p-value of 0.097.

The overall analysis of the splicing events in terms of the survival of the patients has pointed to some interesting results however, due to the missingness and the amount of confounding factors that come along with understanding if the splicing event itself had an effect of the survival of the tumor patients, this analysis might need to be conducted after understanding the changes that could occur in the BPTF gene while the treatment is occuring, and any changes that might occur due to the treatment. Additionally, having more complete data with a lot of null or NA values would lead to more accurate results and could allow the analysis of other factors such as age, height, weight and tumor status which are variables that were included in the dataset but were missing in some cancer types or were noted as zero's where using that variable would yield improper results.

There are studies such as the *Future directions of high throughput splicing assay in precision medicine* by Rhine, C. L., et.al., 2019, being conducted to address splicing databases and analysis

of different variants by addressing limitations that can be overcome like tissue-specific splicing, effect of surrounding sequence context, intronic variants, synthesizing large exons and also amplifying the complex libraries for these different variants.

To further analyze the survival data, the protein configurations of the alternatively spliced BPTF transcripts should be examined to better understand their interactions with other genes, particularly those that are involved in cancer. Previous research has shown that exons 5 and 6 of BPTF interact with *KEAP1*, which has been implicated as a gene that plays a role in cancer development, along with *NRF2*. Additionally, investigating the structure of BPTF and its interactions with other NURF subunits, such as *SMARCA/SNF2* and *RBBP7,* as well as co-expressed genes, may yield valuable insights.

DNA/CpG Methylation

The methylation patterns seen around and in the CpG islands of the BPTF gene indicate that the gene was overall more methylated in normal patient samples than that of the tumor patient samples. As seen in Figure 6 the hypermethylation leading to the shutdown of transcription of the gene aids in the regulation of how often a gene is being transcribed and then eventually translated. In the case of tumor suppressor genes like *TP53, BRCA1* and BRCA2 are hypermethylated which lead to their inactivation. Their ability to suppress tumors is suppressed by this event. BPTF a transcription factor acts a bit differently where having it regulated through methylation is crucial in order to not overproduce the gene and eventually the protein. In BLCA, BRCA and UCEC the CpG island probes and the probes around the islands indicated that there was higher methylation seen in normal cells than those from tumor cells. Furthermore, in most cases the BPTF gene was highly methylated throughout the gene in the normal cells compared to the tumor cells (other than in THCA) indicative of their maintenance of methylation pattern. As the BPTF gene is overexpressed in many if not all cancer types, the hypomethylation patterns indicate that there is a possibility that this hypomethylation is leading to the gene being transcribed and eventually translated more often.

In order to further this analysis, the MRS can be calculated for the gene using the beta values and weights to more accurately predict methylation patterns on each sight. MRS are the methylation risk scores that are calculated by using weighted sums of the individual's methylation markers beta values of a pre-selected number of CpG sites (Hüls, A., & Czamara, D., 2020). This method not only allows for understanding epigenetic biomarkers but also helps with association analyses

in which a single CpG site did not give significance. It can also aid in predicting the individual risk of disease or treatment success. Further analysis of the acetylation patterns of the BPTF gene would also be beneficial in identifying a more holistic pattern that affects the transcription of the gene.

Gene Ontology Analysis

The biological processes, cellular components and the molecular functions noted to be of high significance indicate pathways and their respective genes that are highly related and co-expressed with BPTF. Since BPTF is a transcription factor and is a conserved gene across species, the involvement of its co-expressed genes in biological processes like translation and transcription is expected. The other significant biological processes that are involved that have been major players of cancer are DNA repair mechanisms, cellular response to DNA damage, cell division and many others. The cellular components of high significance like nucleus, nucleosome and cytosol as well as the molecular functions of high significance like RNA and protein binding are all interconnected in many ways. These processes, components and functions have also been noted to play major roles in controlling and contributing to the initiation and progression of cancer. The genetic information that is consistently maintained to ensure stability. Mutations are common but sometimes can be harmful and can damage mechanisms that can lead to disease. DNA damage and error in the DNA repair mechanisms could lead to activation of oncogenes and inactivation of tumor suppressor genes. Since BPTF is a transcription factor and has been known to interact with other genes that are tumor suppressors and oncogenes these pathways can aid in better understanding on the effect BPTF overexpression might have on its co-expressed genes.

The overexpression of BPTF and the positively correlated co-expression of some of the genes involved in these processes gives a better outlook on what BPTF's role might be in these five cancer types. The co-expression of related genes like *EGFR, MYCBP2, MAPK1, BRCA2, ATF2, ATF1* and many more have shown medium to high positive spearman correlation values (cBioPortal) when comparing the mRNA expression levels (RSEM normalized Illumina HiSeq) of BPTF and the respective genes. *MAPK* signaling cascades consist of the interaction of one or more growth factors with their specific receptors. *MAPK1* along with other analogs of the protein were noted to be positively correlated to BPTF. The *MAPK1* and *BPTF* pathways also cross paths with the *EGFR* which is the epidermal growth factor receptor whose physiological function

is to regulate the epithelial tissue development and homeostasis (Sigismund, S., et.al., 2018). A study conducted by Dongyu Bai et al. through RT-qPCR, western blotting bioinformatic analysis and immunohistochemistry indicated that BPTF and *Raf1* were overexpressed in T-cell lymphoma tissues compared to normal and that BPTF promoted the activation of the MAPK pathway. This would lead to the activation of *EGFR* which is commonly upregulated in many cancers like breast, pancreatic and metastatic colorectal cancer (Wee, P., & Wang, Z., 2017). *SMARCC1* and *SMARCC2* are part of the SWI/SNF family and have been noted to have diagnostic value in HCC patients. The *SMARCC1* gene has shown to affect immune infiltration and potentially play a role in tumor promoting role in HCC (Cai, X., Zhou, et.al., 2021). Another pathway that could be studied further is related to the *ATF1, ATF2 , BRCA1* and *BRCA2*.
Poly-ADP ribose polymerase also known as PARP functions in various DNA damage repair pathways and in DNA replication. PARP inhibitors impairs the repair of DNA breaks which leads to the inhibition of repairs. Homologous recombination (HR) contributes to the repair of DNA double stranded breaks, PARP trapping and collapsed forks. The alterations in HR factors like *BRCA1/2* can cause hereditary cancers like breast and ovarian cancer.
The overexpression of BPTF in many cancers is not alone sufficient to cause the initiation and progression of cancer, however if the genes in pathways have had several significant mutational changes and are constantly being co-expressed at a high rate in these different cancer types it illustrates the importance of controlling the expression of BPTF that downstream could lower the transcription of these mutated/ non-functional genes.

7. Conclusion

In conclusion, the analysis of BPTF and the publicly available datasets is crucial for advancing cancer research and developing improved databases that will facilitate further studies in the field. By understanding the role of BPTF in cancer initiation and progression, researchers can focus on developing targeted therapies that will improve patient outcomes. Furthermore, the analysis of publicly available datasets provides a wealth of information that can inform research and identify areas where more focused studies are needed.
As the databases continue to expand, collaborations among researchers and organizations are essential to ensure that the data is consistent and complete. This will enable researchers to conduct more in-depth analyses and develop better predictive models for cancer. Additionally,

the use of publicly available datasets will enhance the reproducibility and transparency of cancer research by enabling other researchers to verify and build upon existing studies.

The results of this study provided significant insights into the potential mechanisms of BPTF overexpression in certain cancers through the analysis of methylation using the TCGA Wanderer bioinformatics tool. While the study only examined five different cancer types, the findings suggest that differences in the alternative splicing of BPTF transcripts such as the alternate donor of exon 23.2 and exon skip of exons 5 and 6, may contribute to the observed overexpression, highlighting the need for further analysis of these transcripts through structural and interaction studies using the TCGA SpliceSeq database or a much more comprehensive splicing database. Although there were no significant differences in the survival analysis between the high and low inclusion of the exons of interest from the splicing analysis in tumor patients. There were interesting trends between the two categories of relative inclusion which can be studies in other cancer types and analyzed with more restrictive cutoff levels.

Furthermore, the gene ontology and co-expressed gene analysis provided valuable information on the biological processes, cellular components, and molecular functions associated with BPTF, as well as its co-expressed genes such as *MAPK1, EGFR, BRCA2* and so on, providing a deeper understanding of the potential mechanisms of BPTF in the initiation and progression of cancer. Even though this study does not provide any direct evidence of interactions between BPTF and these co-related genes, the ontology analysis does highlight the potential genes that potentially do have interactive power and should be researched further to establish interactive status. By expanding these findings to other cancer types and analyzing a wider range of transcripts and co-expressed genes, we can gain a more comprehensive understanding of the role of BPTF in cancer.

Ultimately, the analysis of BPTF and the publicly available datasets has the potential to transform cancer research by leading to the development of more efficient and long-term cures for all forms of cancer. By leveraging the power of bioinformatics, researchers can better understand the complexities of cancer biology and develop personalized treatment plans that will improve patient outcomes. Therefore, it is imperative to continue to invest in this area of research to accelerate the development of more effective cancer therapies.

8.Future Directions

While this study has shed light on certain aspects of the BPTF gene, there is still much to be explored. Specifically, a more comprehensive analysis involving all 33 cancer types using the methodology outlined in this study is necessary to gain a more thorough understanding of how BPTF splicing, methylation, and gene ontology evolve across different cancer types. Such a meta-analysis could serve as a valuable resource for future research in both bioinformatics and wet-bench experimental work, by identifying important pathways and genes involved in the initiation and progression of cancer. Ultimately, a more holistic view of the workings of the BPTF gene can help pave the way towards better diagnostic and treatment strategies for cancer patients.

Despite the limitations of this study utilizing publicly available data, there is potential to overcome these challenges by promoting collaborations and expanding the databases. The Cancer Genome Atlas project provides extensive characteristics such as genomic and transcriptomic information for analysis. Addressing missing data, standardizing data recording and storage of patient samples and their attributes would lead to more robust analyses using publicly available tools, which in turn could guide researchers in identifying potential targets for therapies.

Furthermore, as part of my research in the lab, I am currently working with more sophisticated bioinformatics tools and larger databases, such as CBioPortal and CIBERSORTx, to assess BPTF mRNA expression and copy number alterations in different types of cancer. CBioPortal is a comprehensive platform for exploring multidimensional cancer genomics data, including TCGA and other large-scale cancer genomics datasets (Cerami, E., et al., 2012). Meanwhile, CIBERSORTx is a versatile computational method that quantifies cell fractions from bulk tissue gene expression profiles Newman, A. M., et al., 2019).

To analyze the immune composition of tumor biopsies based on BPTF gene analysis in our lab, we use both mRNA sequencing and microarray data. The CIBERSORTx interface requires a signature matrix that includes signature gene expressions for deconvolving cell types of interest (Figure 25). The tool leverages support vector regression and robust mathematical optimization techniques to enhance deconvolution performance. Specifically, CIBERSORTx applies a machine learning algorithm to infer the proportions of cell types present in a bulk tissue sample from the expression levels of signature genes. The tool uses an optimization approach based on

quadratic programming to estimate the cell fractions that best fit the gene expression profile of the sample (Newman, A. M., et al., 2019).

In summary, our research team is using bioinformatics tools and resources, such as CBioPortal and CIBERSORTx, to study BPTF expression and copy number variations in different cancer types. By leveraging the capabilities of CIBERSORTx, we are able to analyze the immune composition of tumor biopsies with greater accuracy and efficiency, providing insights that could help improve cancer diagnosis and treatment.

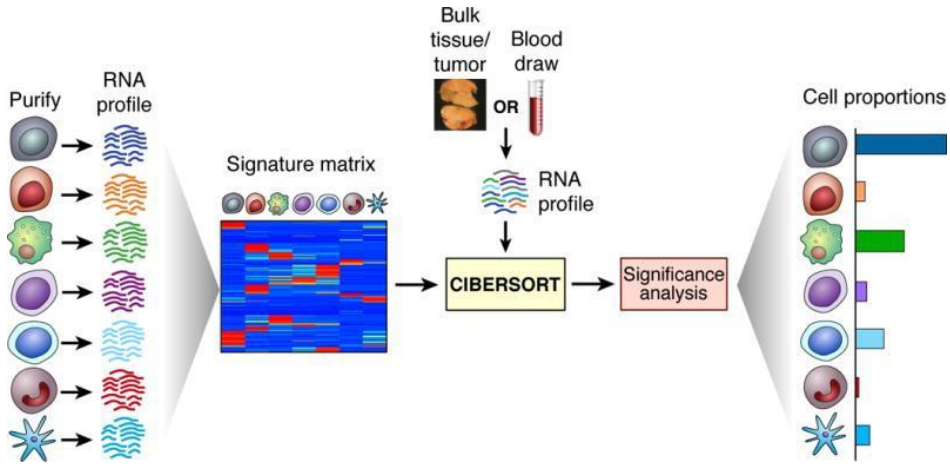

Figure 25: CIBERSORT Analysis Pathway. *The cell proportionality calculated based on the signature matrix that has the deconvoluted profiles from many normalized samples from previous studies is placed as a control. The RNA profiles of interest gathered from the source of interest are also entire and a fraction is delivered in matrix format with the different cell types as columns and the individual samples as the rows. Note.* Represented from "Profiling tumor infiltrating immune cells with CIBERSORT" by Chen, B., et.al., 2018, Methods in molecular biology, 1711, p.14.

9.Supplemental Data

Table S-1: Alternative Splicing Dataset Demographics. The data distribution of all the variables included in the TCGA SpliceSeq database dataset for the three splicing events for the five cancer types.

TCGA Splicing Data: Patient Characteristics
Alternate Donor Exon 23.2

Characteristic	BLCA, N = 425†	BRCA, N = 1,207†	PRAD, N = 549†	THCA, N = 588†	UCEC, N = 580†
Gender					
	0 / 406 (0%)	36 / 1,129 (3.2%)	2 / 499 (0.4%)	7 / 512 (1.4%)	0 / 545 (0%)
FEMALE	105 / 406 (26%)	1,081 / 1,129 (96%)	0 / 499 (0%)	369 / 512 (72%)	545 / 545 (100%)
MALE	301 / 406 (74%)	12 / 1,129 (1.1%)	497 / 499 (100%)	136 / 512 (27%)	0 / 545 (0%)
Missing (NA)	19	78	50	76	35
Age	69 (11) (34,90)	58 (13) (0,90)	61 (7) (41,78)	46 (16) (15,89)	64 (12) (0,90)
Missing (NA)	19	114	52	83	35
Height(Centimeters)	170 (57) (0,196)	0 (0) (0,0)	0 (0) (0,0)	0 (0) (0,0)	161 (37) (0,183)
Missing (NA)	19	114	52	83	35
Weight(Kilograms)	75 (33) (0,292)	0 (0) (0,0)	0 (0) (0,0)	0 (0) (0,0)	83 (30) (0,209)
Missing (NA)	19	114	52	83	35
Tumor Status					
	0 / 406 (0%)	36 / 1,129 (3.2%)	2 / 499 (0.4%)	7 / 512 (1.4%)	0 / 545 (0%)
null	38 / 406 (9.4%)	124 / 1,129 (11%)	91 / 499 (18%)	44 / 512 (8.6%)	38 / 545 (7.0%)
TUMOR FREE	233 / 406 (57%)	875 / 1,129 (78%)	313 / 499 (63%)	411 / 512 (80%)	429 / 545 (79%)
WITH TUMOR	135 / 406 (33%)	94 / 1,129 (8.3%)	93 / 499 (19%)	50 / 512 (9.8%)	78 / 545 (14%)
Missing (NA)	19	78	50	76	35
Vital Status					
	0 / 406 (0%)	36 / 1,129 (3.2%)	2 / 499 (0.4%)	7 / 512 (1.4%)	0 / 545 (0%)
Alive	299 / 406 (74%)	989 / 1,129 (88%)	489 / 499 (98%)	491 / 512 (96%)	500 / 545 (92%)
Dead	107 / 406 (26%)	104 / 1,129 (9.2%)	8 / 499 (1.6%)	14 / 512 (2.7%)	45 / 545 (8.3%)
Missing (NA)	19	78	50	76	35
Percent Spliced In	0.37 (0.10) (0.10,1.00)	0.40 (0.09) (0.05,0.80)	0.34 (0.09) (0.09,1.00)	0.35 (0.08) (0.08,1.00)	0.40 (0.12) (0.00,1.00)
Missing (NA)	2	2	5	4	7
Patient Sample Type					
Normal	19 / 425 (4.5%)	113 / 1,207 (9.4%)	52 / 549 (9.5%)	71 / 588 (12%)	35 / 580 (6.0%)
Tumor	406 / 425 (96%)	1,094 / 1,207 (91%)	497 / 549 (91%)	517 / 588 (88%)	545 / 580 (94%)

† n / N (%); Median (SD) (Minimum,Maximum)

TCGA Splicing Data: Patient Characteristics
Exon Skip 5

Characteristic	BRCA, N = 1,207†
Gender	
	36 / 1,129 (3.2%)
FEMALE	1,081 / 1,129 (96%)
MALE	12 / 1,129 (1.1%)
Missing (NA)	78
Age	58 (13) (0,90)
Missing (NA)	114
Height(Centimeters)	0 / 1,093 (0%)
Missing (NA)	114
Weight(Kilograms)	0 / 1,093 (0%)
Missing (NA)	114
Tumor Status	
	36 / 1,129 (3.2%)
null	124 / 1,129 (11%)
TUMOR FREE	875 / 1,129 (78%)
WITH TUMOR	94 / 1,129 (8.3%)
Missing (NA)	78
Vital Status	
	36 / 1,129 (3.2%)
Alive	989 / 1,129 (88%)
Dead	104 / 1,129 (9.2%)
Missing (NA)	78
Percent Spliced In	0.46 (0.20) (0.00,0.94)
Missing (NA)	296
Patient Sample Type	
Normal	113 / 1,207 (9.4%)
Tumor	1,094 / 1,207 (91%)

† n / N (%); Median (SD) (Minimum,Maximum)

TCGA Splicing Data: Patient Characteristics
Exon Skip 5:6

Characteristic	BRCA, N = 1,207†	PRAD, N = 549†	THCA, N = 588†
Gender			
	36 / 1,129 (3.2%)	2 / 499 (0.4%)	7 / 512 (1.4%)
FEMALE	1,081 / 1,129 (96%)	0 / 499 (0%)	369 / 512 (72%)
MALE	12 / 1,129 (1.1%)	497 / 499 (100%)	136 / 512 (27%)
Missing (NA)	78	50	76
Age	58 (13) (0,90)	61 (7) (41,78)	46 (16) (15,89)
Missing (NA)	114	52	83
Height(Centimeters)	0 / 1,093 (0%)	0 / 497 (0%)	0 / 505 (0%)
Missing (NA)	114	52	83
Weight(Kilograms)	0 / 1,093 (0%)	0 / 497 (0%)	0 / 505 (0%)
Missing (NA)	114	52	83
Tumor Status			
	36 / 1,129 (3.2%)	2 / 499 (0.4%)	7 / 512 (1.4%)
null	124 / 1,129 (11%)	91 / 499 (18%)	44 / 512 (8.6%)
TUMOR FREE	875 / 1,129 (78%)	313 / 499 (63%)	411 / 512 (80%)
WITH TUMOR	94 / 1,129 (8.3%)	93 / 499 (19%)	50 / 512 (9.8%)
Missing (NA)	78	50	76
Vital Status			
	36 / 1,129 (3.2%)	2 / 499 (0.4%)	7 / 512 (1.4%)
Alive	989 / 1,129 (88%)	489 / 499 (98%)	491 / 512 (96%)
Dead	104 / 1,129 (9.2%)	8 / 499 (1.6%)	14 / 512 (2.7%)
Missing (NA)	78	50	76
Percent Spliced In	0.49 (0.21) (0.00,1.00)	0.50 (0.22) (0.00,1.00)	0.55 (0.20) (0.00,1.00)
Missing (NA)	130	113	90
Patient Sample Type			
Normal	113 / 1,207 (9.4%)	52 / 549 (9.5%)	71 / 588 (12%)
Tumor	1,094 / 1,207 (91%)	497 / 549 (91%)	517 / 588 (88%)

† n / N (%); Median (SD) (Minimum,Maximum)

Table S-2: BPTF Exons on Chromosome 17. BPTF exons, chromosome number, exon start and end positions based on the TCGA SpliceSeq database.

BPTF Gene Exons

Exon	Chromosome	Exon Start	Exon End
1.0	17	65821780	65822453
2.0	17	65850056	65850878
3.0	17	65862580	65862803
4.0	17	65870933	65871136
5.0	17	65871672	65871860
6.0	17	65882244	65882432
7.0	17	65887960	65888150
8.0	17	65889486	65889841
9.0	17	65890150	65890281
10.0	17	65899905	65900034
11.0	17	65900818	65900956
12.0	17	65905698	65905877
13.0	17	65906993	65909303
14.0	17	65914830	65914954
15.0	17	65916131	65916259
16.0	17	65918956	65919106
17.0	17	65920663	65920705
18.0	17	65924471	65924717
19.0	17	65925452	65925603
20.0	17	65928027	65928135
21.0	17	65936555	65936772
22.0	17	65940266	65940488
23.1	17	65941525	65942012
23.2	17	65942013	65942441
24.0	17	65943842	65943924
25.0	17	65944197	65944422
26.0	17	65955657	65955991
27.0	17	65960328	65960520
28.0	17	65962688	65962772
29.0	17	65971888	65972074
30.0	17	65978368	65980494

Table S-3: Significant Methylation Probes and Fold Change. The significant probes based on Wilcoxon Test comparing the normal and tumor beta methylation values of bladder urothelial carcinoma and breast invasive carcinoma. The fold change was calculated based on the average of normal beta values over the average of tumor beta values.

Significant Methylation Probes and Fold Change between Normal and Tumor Samples in Bladder Urothelial Carcinoma

Methylation Probes	Wilcoxon Pvalue	Fold Change (Normal/Tumor)
cg12398397	8.816287e-06	1.333597e+00
cg22443205	3.871271e-04	1.088616e+00
cg16435601	2.934226e-04	1.396757e+00
cg03215657	3.854899e-02	1.062243e+00
cg06647360	5.706201e-04	1.029907e+00
cg18762647	2.064174e-05	1.041583e+00
cg06420179	9.217278e-03	1.008522e+00
cg05271336	3.273558e-03	1.031704e+00
cg02530753	4.535624e-02	1.009673e+00
cg14781826	1.372386e-04	8.574999e-01

Significant Methylation Probes and Fold Change between Normal and Tumor Samples in Breast Invasive Carcinoma

Methylation Probes	Wilcoxon Pvalue	Fold Change (Normal/Tumor)
cg12398397	2.737993e-12	1.607664e+00
cg16028064	2.949000e-13	2.087730e-01
cg22972262	5.798517e-05	9.523159e-01
cg06647360	7.359282e-03	1.008033e+00
cg06420179	9.399234e-03	1.004298e+00
cg05271336	4.470224e-02	9.912485e-01
cg02530753	1.725618e-02	9.989184e-01
cg14781826	2.474528e-14	7.780636e-01

Table S-4: Significant Methylation Probes and Fold Change. The significant probes based on Wilcoxon Test comparing the normal and tumor beta methylation values of prostate adenocarcinoma and thyroid carcinoma. The fold change was calculated based on the average of normal beta values over the average of tumor beta values.

Significant Methylation Probes and Fold Change between Normal and Tumor Samples in Prostate Adenocarcinoma

Methylation Probes	Wilcoxon Pvalue	Fold Change (Normal/Tumor)
cg12398397	2.873313e-03	1.039143e+00
cg13776905	9.517346e-04	1.015000e+00
cg18687753	1.295735e-02	1.023326e+00
cg04798252	6.827904e-06	1.017241e+00
cg06647360	2.976990e-03	1.006002e+00
cg18762647	2.311476e-10	1.029996e+00
cg06420179	3.626461e-03	1.005579e+00
cg05271336	1.168047e-05	1.032317e+00
cg26135490	8.639130e-06	1.014526e+00
cg02530753	2.662843e-07	1.015890e+00
cg14781826	7.568057e-09	7.456007e-01

Significant Methylation Probes and Fold Change between Normal and Tumor Samples in Thyroid Carcinoma

Methylation Probes	Wilcoxon Pvalue	Fold Change (Normal/Tumor)
cg12398397	6.366349e-04	1.052635e+00
cg22443205	1.416353e-03	9.091633e-01
cg16028064	1.834788e-02	1.047260e+00
cg13776905	9.667742e-03	9.921533e-01
cg18762647	1.906598e-02	9.950842e-01
cg05271336	4.738749e-02	9.849759e-01
cg26135490	2.953298e-02	9.948875e-01
cg02530753	2.176396e-02	9.924304e-01
cg14781826	7.714886e-05	9.751168e-01

Table S-5: Significant Methylation Probes and Fold Change. The significant probes based on Wilcoxon Test comparing the normal and tumor beta methylation values of uterine corpus endometrial carcinoma. The fold change was calculated based on the average of normal beta values over the average of tumor beta values.

Significant Methylation Probes and Fold Change between Normal and Tumor Samples in Uterine Corpus Endometrial Carcinoma

Methylation Probes	Wilcoxon Pvalue	Fold Change (Normal/Tumor)
cg12398397	7.891450e-09	1.264626e+00
cg22443205	1.423781e-12	1.334654e+00
cg16435601	4.039807e-05	1.226272e+00
cg06647360	2.190113e-11	1.034001e+00
cg18762647	3.067688e-07	1.026310e+00
cg02530753	4.081343e-03	1.017019e+00
cg14781826	1.757989e-16	5.614152e-01

Table S-6: Gene Ontology: Biological Processes related to BPTF. The biological processes related to BPTF in all five cancers with the GO term code, the biological process terms, and the FDR values to note for significance of each process in each of the cancer types.

Biological Processes Gene Ontology Related to BPTF

		FDR Values				
GO Term	Term	BRCA	BLCA	PRAD	THCA	UCEC
GO:0006412	translation	2.77e-13	1.2e-05	0.02844	9.01e-23	1.6e-06
GO:0006974	cellular response to DNA damage stimulus	3.79e-10	1.15e-08	2.15e-07	9.4e-06	5.07e-07
GO:0006281	DNA repair	7.08e-10	1.04e-06	5.15e-10	6.52e-08	3.11e-09
GO:0006357	regulation of transcription from RNA polymerase II promoter	1.2e-07	1.34e-17	3.55e-09	8.34e-12	1.98e-09
GO:0006338	chromatin remodeling	1.2e-07	8.8e-12	5.27e-12	9.15e-09	2.05e-10
GO:0006355	regulation of transcription, DNA-templated	2.63e-07	4.24e-15	5.06e-11	8.63e-10	1.54e-07
GO:0042776	mitochondrial ATP synthesis coupled proton transport	8.94e-07	1.57e-05	7.94e-13	3.08e-16	2.33e-10
GO:0045893	positive regulation of transcription, DNA-templated	6.02e-06	5.57e-12	1.15e-07	1.94e-08	1.91e-09
GO:0006325	chromatin organization	9.61e-06	8.04e-13	9.19e-09	1.22e-06	5.26e-06
GO:0009060	aerobic respiration	1.87e-05	5.69e-05	5.27e-12	2.23e-11	1.58e-09
GO:0051301	cell division	0.00015	4.6e-05	0.001662	0.00194	0.04737
GO:0016567	protein ubiquitination	0.000384	1.26e-06	1.37e-06	9.33e-05	1.5e-05
GO:0043161	proteasome-mediated ubiquitin-dependent protein catabolic process	0.000612	0.000335	8.8e-05	7.48e-08	3.88e-06
GO:0032981	mitochondrial respiratory chain complex I assembly	0.00082	0.001915	1.05e-08	3e-10	3.23e-05
GO:2000779	regulation of double-strand break repair	0.000847	0.01542	0.000783	0.01949	0.004078
GO:0032508	DNA duplex unwinding	0.000853	0.000204	6.5e-05	5.11e-05	2.89e-06
GO:0016573	histone acetylation	0.000978	0.000227	0.000963	0.003371	5.04e-05
GO:0045944	positive regulation of transcription from RNA polymerase II promoter	0.001022	3.65e-07	1.28e-06	1.32e-05	3.24e-07
GO:0051726	regulation of cell cycle	0.001131	0.002878	0.0233	0.02494	0.000226
GO:0032543	mitochondrial translation	0.001241	0.000335	5.15e-10	1.03e-14	2.68e-06
GO:0006120	mitochondrial electron transport, NADH to ubiquinone	0.002711	0.008046	1.22e-07	1.72e-08	4.19e-06
GO:0000398	mRNA splicing, via spliceosome	0.004492	0.04073	8e-05	2.8e-05	0.01605
GO:0045945	positive regulation of transcription from RNA polymerase III promoter	0.0168	0.000253	0.01559	0.01644	0.01492
GO:0000209	protein polyubiquitination	0.03859	0.003233	0.004472	0.03604	0.000292

Table S-7: Gene Ontology: Cellular Components related to BPTF. The cellular components related to BPTF in all five cancers with the GO term code, the cellular component terms, and the FDR values to note for significance of each process in each of the cancer types.

Cellular Components Gene Ontology Related to BPTF

		FDR Values				
GO Term	Term	BRCA	BLCA	PRAD	THCA	UCEC
GO:0005654	nucleoplasm	1e-69	3.35e-88	8.57e-67	4.1e-86	4.77e-72
GO:0005634	nucleus	2.08e-50	1.02e-65	1.32e-52	1.77e-60	2.26e-45
GO:0005829	cytosol	6.15e-40	5.56e-40	2e-32	2.28e-39	1.75e-29
GO:0016020	membrane	2.35e-19	3.5e-10	9.35e-08	2.79e-10	2.2e-12
GO:0005737	cytoplasm	1.45e-17	4.3e-12	4.9e-08	3.75e-13	4.46e-09
GO:0005840	ribosome	1.22e-16	7.44e-06	0.02817	2.65e-20	8.44e-06
GO:0005743	mitochondrial inner membrane	9.33e-12	5.66e-09	8.89e-24	8.91e-33	4.77e-15
GO:0005813	centrosome	5.25e-09	3.34e-06	7.27e-06	1.38e-08	0.000167
GO:0005739	mitochondrion	2.67e-08	0.000216	7.14e-17	1.11e-26	1.08e-10
GO:0016607	nuclear speck	4.04e-08	5.11e-15	7.77e-10	2.29e-11	1.46e-07
GO:0005747	mitochondrial respiratory chain complex I	2.6e-06	4.9e-05	2.45e-11	1.26e-12	1.21e-09
GO:0016604	nuclear body	8.52e-06	4.53e-11	1.6e-09	1.89e-12	2.87e-06
GO:0005694	chromosome	1.64e-05	8.42e-07	3.64e-06	4.12e-08	1.43e-07
GO:0005762	mitochondrial large ribosomal subunit	0.000439	0.000293	7.77e-10	1.04e-13	0.000106
GO:0031965	nuclear membrane	0.003506	4.55e-09	0.000509	6.32e-07	0.009181
GO:0000776	kinetochore	0.003506	9.92e-06	0.02628	0.000977	0.02435
GO:0032991	macromolecular complex	0.003684	2.67e-05	0.001739	0.01477	0.00051
GO:0000151	ubiquitin ligase complex	0.003684	1.73e-05	0.001739	0.000978	0.0044
GO:0030008	TRAPP complex	0.005333	0.02581	0.000927	0.003445	0.000852
GO:0005730	nucleolus	0.00584	0.01712	0.04412	0.001629	0.006161
GO:1990072	TRAPPIII protein complex	0.008727	0.04563	0.001395	0.005983	0.001296
GO:0035861	site of double-strand break	0.01318	0.00412	0.02982	0.000458	0.0294
GO:0005643	nuclear pore	0.0163	0.000842	0.01959	2.65e-05	0.007328
GO:0016605	PML body	0.02955	3.34e-06	0.00291	2.55e-06	0.0294

Table S-8: Gene Ontology: Molecular Functions related to BPTF. The molecular functions related to BPTF in all five cancers with the GO term code, the molecular function terms, and the FDR values to note for significance of each process in each of the cancer types.

Molecular Functions Gene Ontology Related to BPTF

		FDR Values				
GO Term	Term	BRCA	BLCA	PRAD	THCA	UCEC
GO:0005515	protein binding	7.06e-37	5.47e-36	1.11e-42	3.88e-43	1.31e-39
GO:0003723	RNA binding	4.29e-31	3.2e-24	7.41e-23	2e-37	5.7e-29
GO:0003735	structural constituent of ribosome	1.37e-15	1.61e-08	0.001237	3.58e-28	2.74e-10
GO:0046872	metal ion binding	6.85e-15	6.02e-22	5.34e-13	6.98e-15	1.23e-17
GO:0070615	nucleosome-dependent ATPase activity	8.9e-07	4.17e-09	5.69e-08	4.14e-06	2.39e-05
GO:0031267	small GTPase binding	1.12e-06	0.00245	3.64e-06	0.02694	0.000831
GO:0003677	DNA binding	6.09e-06	1.92e-15	5.29e-11	1.55e-08	4.13e-13
GO:0004842	ubiquitin-protein transferase activity	6.29e-06	3.33e-10	7.44e-09	2.82e-09	2.74e-10
GO:0003682	chromatin binding	6.29e-06	3.41e-11	1.63e-08	5.34e-05	1.07e-09
GO:0016887	ATPase activity	6.29e-06	5.26e-07	2.78e-06	6.48e-05	4.96e-06
GO:0003712	transcription cofactor activity	6.36e-05	3.35e-06	5.99e-06	2.86e-06	2.07e-06
GO:0008137	NADH dehydrogenase (ubiquinone) activity	0.000151	0.000373	2.43e-10	2.63e-10	1.42e-08
GO:0004386	helicase activity	0.000307	2.43e-06	1.59e-05	0.000958	0.003404
GO:0042393	histone binding	0.000535	3.91e-08	7.66e-08	0.000234	6.45e-07
GO:0004402	histone acetyltransferase activity	0.002125	0.001137	0.000397	0.001724	0.000408
GO:0043130	ubiquitin binding	0.00278	1.73e-05	0.002392	0.000117	0.002987
GO:0003678	DNA helicase activity	0.003766	0.000651	0.001158	0.001045	7.76e-05
GO:0000978	RNA polymerase II core promoter proximal region sequence-specific DNA binding	0.01604	6.97e-07	0.009167	0.000915	0.01265
GO:0003713	transcription coactivator activity	0.03858	6.97e-07	3.64e-06	9.04e-07	8.87e-08
GO:0008270	zinc ion binding	0.03858	3.27e-05	0.000725	0.01447	0.009812

10.References

Alkhatib, S. G., & Landry, J. W. (2011). The nucleosome remodeling factor. FEBS letters, 585(20), 3197–3207. https://doi.org/10.1016/j.febslet.2011.09.003

Ammerpohl, O., Haake, A., Kolarova, J., & Siebert, R. (2016). Quantitative DNA Methylation Profiling in Cancer. In R. Grützmann & C. Pilarsky (Eds.), Cancer Gene Profiling (pp. 69-85). Methods in Molecular Biology, Vol. 1381. Humana Press. https://doi.org/10.1007/978-1-4939-3204-7_5

Aysola, K., Desai, A., Welch, C., Xu, J., Qin, Y., Reddy, V., Matthews, R., Owens, C., Okoli, J., Beech, D. J., Piyathilake, C. J., Reddy, S. P., & Rao, V. N. (2013). Triple Negative Breast Cancer - An Overview. Hereditary genetics : current research, 2013(Suppl 2), 001. https://doi.org/10.4172/2161-1041.S2-001

Bai, Y., Wang, H., Wu, X., Weng, M., Han, Q., Xu, L., Zhang, H., Chang, C., Jin, C., Chen, M., Luo, K., & Teng, X. (2022). Study on Molecular Information Intelligent Diagnosis and Treatment of Bladder Cancer on Pathological Tissue Image. Frontiers in medicine, 9, 838182. https://doi.org/10.3389/fmed.2022.838182

Bonnal, S. C., López-Oreja, I., & Valcárcel, J. (2020). Roles and mechanisms of alternative splicing in cancer - implications for care. Nature reviews. Clinical oncology, 17(8), 457–474. https://doi.org/10.1038/s41571-020-0350-x

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 68(6), 394–424. https://doi.org/10.3322/caac.21492

Cabanillas, M. E., McFadden, D. G., & Durante, C. (2016). Thyroid cancer. Lancet, 388(10061), 2783–2795. https://doi.org/10.1016/S0140-6736(16)30172-6

Cai, X., Zhou, J., Deng, J., & Chen, Z. (2021). Prognostic biomarker SMARCC1 and its association with immune infiltrates in hepatocellular carcinoma. Cancer cell international, 21(1), 701. https://doi.org/10.1186/s12935-021-02413-w

Cerami, E., Gao, J., Dogrusoz, U., Gross, B. E., Sumer, S. O., Aksoy, B. A., Jacobsen, A., Byrne, C. J., Heuer, M. L., Larsson, E., Antipin, Y., Reva, B., Goldberg, A. P., Sander, C., & Schultz, N. (2012). The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer discovery, 2(5), 401–404. https://doi.org/10.1158/2159-8290.CD-12-0095

Cheong, H., Lu, C., Lindsten, T., & Thompson, C. B. (2012). Therapeutic targets in cancer cell metabolism and autophagy. Nature biotechnology, 30(7), 671–678. https://doi.org/10.1038/nbt.2285
Crawford, E.D. (2003). Epidemiology of prostate cancer. Urology, 62(6 Suppl 1), 3-12. https://doi.org/10.1016/j.urology.2003.10.013. PMID: 14706503.

Dennis, G., Jr, Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., & Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology, 4(5), P3.

Díez-Villanueva, A., Mallona, I., & Peinado, M.A. (2015). Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. Epigenetics & Chromatin, 8, 22. https://doi.org/10.1186/s13072-015-0014-8

Doerks, T., Copley, R., & Bork, P. (2001). DDT--a novel domain in different transcription and chromosome remodeling factors. Trends in Biochemical Sciences, 26(3), 145–146. https://doi.org/10.1016/s0968-0004(00)01769-2

Duan, C., Wang, H., Chen, Y., et al. (2018). Whole exome sequencing reveals novel somatic alterations in neuroblastoma patients with chemotherapy. Cancer Cell International, 18, 21. https://doi.org/10.1186/s12935-018-0521-3

Endo, S., Yoshino, Y., Shirota, M., Watanabe, G., & Chiba, N. (2021). BRCA1/ATF1-Mediated Transactivation is Involved in Resistance to PARP Inhibitors and Cisplatin. Cancer Research Communications, 1(2), 90–105. https://doi.org/10.1158/2767-9764.CRC-21-0064

Esteller, M. (2002). CpG island hypermethylation and tumor suppressor genes: A booming present, a brighter future. *Oncogene*, *21*(35), 5427-5440. https://doi.org/10.1038/sj.onc.1205600

Ferlay, J., Bray, F., Colombet, M., Mery, L., Pineros, M., Znaor, A., Soerjomataram, I., et al. (2019). Global cancer observatory: cancer today. Lyon, France: International Agency for Research on Cancer. Available from: https://gco.iarc.fr/today. Accessed 02 February 2019.

Ferlay, J., Shin, H. R., Bray, F., Forman, D., Mathers, C., & Parkin, D. M. (2010). Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. International Journal of Cancer, 127(12), 2893–2917. https://doi.org/10.1002/ijc.25516

Firke S (2023). *janitor: Simple Tools for Examining and Cleaning Dirty Data*. R package version 2.2.0, https://CRAN.R-project.org/package=janitor

Gao, J., Aksoy, B. A., Dogrusoz, U., Dresdner, G., Gross, B., Sumer, S. O., Sun, Y., Jacobsen, A., Sinha, R., Larsson, E., Cerami, E., Sander, C., & Schultz, N. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Science Signaling, 6(269), pl1. https://doi.org/10.1126/scisignal.2004088

Hartmann, A., & Friess, H. (2017). Adenocarcinomas ☆. In Reference Module in Life Sciences. https://doi.org/10.1016/B978-0-12-809633-8.06013-1

Hüls, A., & Czamara, D. (2020). Methodological challenges in constructing DNA methylation risk scores. Epigenetics, 15(1-2), 1–11. https://doi.org/10.1080/15592294.2019.1644879
Iannone R, Cheng J, Schloerke B, Hughes E, Lauer A, Seo J (2023). *gt: Easily Create Presentation-Ready Display Tables*. R package version 0.9.0, https://CRAN.R-project.org/package=gt

Kaseb, H., & Aeddula, N. R. (2022). Bladder Cancer. In StatPearls. StatPearls Publishing.
Li, Y., Gong, H., Wang, P., Zhu, Y., Peng, H., Cui, Y., Li, H., Liu, J., & Wang, Z. (2021). The Emerging Role of ISWI Chromatin Remodeling Complexes in Cancer. Journal of Experimental & Clinical Cancer Research, 40(1), 346. https://doi.org/10.1186/s13046-021-02151-x.

Koludrovic, D., Laurette, P., Strub, T., Keime, C., Le Coz, M., Coassolo, S., Mengus, G., Larue, L., & Davidson, I. (2015). Chromatin-Remodelling Complex NURF Is Essential for Differentiation of Adult Melanocyte Stem Cells. *PLoS genetics*, *11*(10), e1005555. https://doi.org/10.1371/journal.pgen.1005555

Mayes, Alkhatib, S. G., Peterson, K., Alhazmi, A., Song, C., Chan, V., Blevins, T., Roberts, M., Dumur, C. I., Wang, X.-Y., & Landry, J. W. (2016). BPTF Depletion Enhances T-cell-Mediated Antitumor Immunity. Cancer Research (Chicago, Ill.), 76(21), 6183–6192. https://doi.org/10.1158/0008-5472.CAN-15-3125

McDonald, E. S., Clark, A. S., Tchou, J., Zhang, P., & Freedman, G. M. (2016). Clinical Diagnosis and Management of Breast Cancer. Journal of Nuclear Medicine, 57(Supplement 1), 9S-16S. https://doi.org/10.2967/jnumed.115.157834

Morcock. C. (2022). Analyses of Internal CRISPR Screen and RNA-Seq data, and Publicly Available Genomic Datasets to Study the Roles of Epigenetics in Cancer Biology. VCU Scholars Compass.

Nair, S. S., & Kumar, R. (2012). Chromatin remodeling in cancer: a gateway to regulate gene transcription. Molecular oncology, 6(6), 611–619. https://doi.org/10.1016/j.molonc.2012.09.005

Newman, A. M., Steen, C. B., Liu, C. L., Gentles, A. J., Chaudhuri, A. A., Scherer, F., Khodadoust, M. S., Esfahani, M. S., Luca, B. A., Steiner, D., Diehn, M., & Alizadeh, A. A. (2019). Determining cell type abundance and expression from bulk tissues with digital cytometry. Nature Biotechnology, 37(7), 773–782. https://doi.org/10.1038/s41587-019-0114-2

Onitilo, A. A., Engel, J. M., Greenlee, R. T., & Mukesh, B. N. (2009). Breast cancer subtypes based on ER/PR and Her2 expression: comparison of clinicopathologic features and survival. Clinical Medicine & Research, 7(1-2), 4–13. https://doi.org/10.3121/cmr.2009.825

Pedersen T (2022). *patchwork: The Composer of Plots*. R package version 1.1.2, https://CRAN.R-project.org/package=patchwork

Prout GR Jr, Barton BA, Griffin PP, et al. (1992). Treated history of noninvasive grade 1 transitional cell carcinoma. The National Bladder Cancer Group. Journal of Urology, 148, 1413-1419.

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rawla, P. (2019). Epidemiology of prostate cancer. World Journal of Oncology, 10(2), 63–89. https://doi.org/10.14740/wjon1191

Rhine, C. L., Neil, C., Glidden, D. T., Cygan, K. J., Fredericks, A. M., Wang, J., Walton, N. A., & Fairbrother, W. G. (2019). Future directions for high-throughput splicing assays in precision medicine. Human Mutation, 40(9), 1225–1234. https://doi.org/10.1002/humu.23866

Rutgers, J. K. (2015). Update on pathology, staging and molecular pathology of endometrial (uterine corpus) adenocarcinoma. Future Oncology, 11(23), 3207–3218. https://doi.org/10.2217/fon.15.262

Ryan, M. C., Cleland, J., Kim, R., Wong, W. C., Weinstein, J. N., & SpliceSeq: A resource for analysis and visualization of RNA-Seq data on alternative splicing and its functional impacts. (2012). Bioinformatics, 28(18), 2385–2387. https://doi.org/10.1093/bioinformatics/bts452

Shen, L., Liu, M., Liu, W., Cui, J., & Li, C. (2018). Bioinformatics analysis of RNA sequencing data reveals multiple key genes in uterine corpus endometrial carcinoma. Oncology Letters, 15(1), 205-212. doi: 10.3892/ol.2017.7346

Sigismund, S., Avanzato, D., & Lanzetti, L. (2018). Emerging functions of the EGFR in cancer. Molecular Oncology, 12(1), 3-20. doi: 10.1002/1878-0261.12155

Singh, A., Ham, J., Po, J. W., Niles, N., Roberts, T., & Lee, C. S. (2021). The genomic landscape of thyroid cancer tumorigenesis and implications for immunotherapy. Cells, 10(5), 1082. doi: 10.3390/cells10051082

Stewart, B. W., & Wild, C. P. (2014). World cancer report 2014. Geneva, Switzerland: WHO Press.

Sun, Y. S., Zhao, Z., Yang, Z. N., Xu, F., Lu, H. J., Zhu, Z. Y., Shi, W., Jiang, J., Yao, P. P., & Zhu, H. P. (2017). Risk factors and preventions of breast cancer. International Journal of Biological Sciences, 13(11), 1387-1397. doi: 10.7150/ijbs.21635

Taguchi, K., & Yamamoto, M. (2017). The KEAP1-NRF2 system in cancer. Frontiers in Oncology, 7, 85. doi: 10.3389/fonc.2017.00085

Takahashi, M., Lio, C. J., Campeau, A., Steger, M., Ay, F., Mann, M., Gonzalez, D. J., Jain, M., & Sharma, S. (2021). The tumor suppressor kinase DAPK3 drives tumor-intrinsic immunity

through the STING-IFN-β pathway. Nature Immunology, 22(4), 485-496. doi: 10.1038/s41590-021-00896-3

Vashisht, S., & Bagler, G. (2012). An approach for the identification of targets specific to bone metastasis using cancer genes interactome and gene ontology analysis. PloS One, 7(11), e49401. doi: 10.1371/journal.pone.0049401

Wajed, S. A., Laird, P. W., & DeMeester, T. R. (2001). DNA methylation: an alternative pathway to cancer. Annals of Surgery, 234(1), 10-20. doi: 10.1097/00000658-200107000-00003
Wang, Y., Liu, J., Huang, B., Xu, Y., Li, J., Huang, L., ... Wang, X. (2015). Mechanism of alternative splicing and its regulation (Review). Biomedical Reports, 3, 152-158. https://doi.org/10.3892/br.2014.407

Wee, P., & Wang, Z. (2017). Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. Cancers, 9(5), 52. https://doi.org/10.3390/cancers9050052

What is cancer? National Cancer Institute. (n.d.). Retrieved April 22, 2023, from https://www.cancer.gov/about-cancer/understanding/what-is-cancer

WHO: Geneva, Switzerland. Breast cancer. http://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4, https://ggplot2.tidyverse.org.

Wickham H, François R, Henry L, Müller K, Vaughan D (2023). *dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr.

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... & Wilson, J. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686. https://doi.org/10.21105/joss.01686

Xie Y (2023). *knitr: A General-Purpose Package for Dynamic Report Generation in R*. R package version 1.42, https://yihui.org/knitr/.

Younis E. (2017). Oncogenesis of Thyroid Cancer. Asian Pacific journal of cancer prevention : APJCP, 18(5), 1191–1199. https://doi.org/10.22034/APJCP.2017.18.5.1191

Zahid, H., Olson, N. M., & Pomerantz, W. (2021). Opportunity knocks for uncovering the new function of an understudied nucleosome remodeling complex member, the bromodomain PHD finger transcription factor, BPTF. Current opinion in chemical biology, 63, 57–67. https://doi.org/10.1016/j.cbpa.2021.02.003

Vita

Preksha Jerajani was born on August 14, 1999, in Anand, Gujarat in India. She moved to the United States when she was twelve years old. She graduated Chantilly High School, Fairfax Virginia in 2017. She received her Bachelor of Science in Bioinformatics from Virginia Commonwealth University, Richmond Virginia in 2021. She became the Vice President of the Bioinformatics Graduate Student Organization and began to pursue her graduate studies in Bioinformatics and Cancer Biology. She interned as a local start-up company to gain more experience in the Cancer research field.