Jenna Jacobs, Joshua Gabbay, Kayla Xu

An Investigation into Glioblastoma Multiforme Survival

Abstract

Glioblastoma multiforme (GBM) is a highly fatal cancer with limited treatments and therapies compared to other tumor-related diseases. This project aims to analyze GBM using data from the Cancer Genome Atlas (TCGA) and Clinical Proteomic Tumor Analysis Consortium (CPTAC) across its genomic, transcriptomic, proteomic, and clinical properties in order to identify any notable differences between patients who survived the disease and patients who didn't. Our results found differences in race, age, mutation types, RNA expression, and protein expression between survivor and non-survivor patients, as well as a suggestion that regulatory elements like non-coding RNA play a large role in GBM survivability.

Introduction

Glioblastoma multiforme (GBM) is the uncontrolled division and replication of astrocyte cells in the central nervous system (Rasheed et al., 2021). It is in the top-ten of tumor-related deaths and the most aggressive and most common form of brain cancer, with an incident rate of 2.82 to 5.10 per 100,000 Americans annually (Lin et al., 2021). Despite increasing rates of GBM worldwide, the disease lacks the breadth of treatment options that other cancers have. On average, patients die within two years of an initial diagnosis (Shen et al., 2019). Previous studies suggest that while brain tumors often resemble their parent tumors on a genomic level, they differ on a transcriptomic level, suggesting that further investigation into the multi-omic profile of this type of brain cancer is warranted (Shen et al., 2019). Risk factors of GBM brain cancer include those that are environmental and also genetic. Some environmental risk factors include ionising radiation exposure. This can come from CT scans and electromagnetic radiation where exposure comes from mobile phones and also atomic bombs and hair dyes (Rasheed et al., 2021).

As described by previous research, some frequently found mutation gene sites in GBM include PTEN, EGFR, CDKN2A, CDH1 (Shen et al., 2019). PTEN is most notably a tumor suppressor gene which has previously been correlated with decreased survival rates in GBM patients (Yang et al., 2017). EGFR, or epidermal growth factor receptor, can be an extremely oncogenic gene in GBM when mutated (Gan et al., 2013). CDKN2A is another tumor suppressor gene that when mutated promotes tumor growth (Wang et al., 2021). Rare germline variation in CDH1 has been found to be more highly represented in GBM patients, presenting a potential genetic risk factor (Förster et al., 2021).

The Cancer Genome Atlas (TCGA) is a cancer genomics database that includes genomic, transcriptomic, epigenomic, and proteomic data from 33 different cancer types (“The Cancer Genome Atlas”). TCGA is incredibly important because it is a huge genome database for use throughout the research community, giving cancer researchers a huge amount of information to study. From TCGA, we can extract clinical data, radiation data, and gene mutation data. The Clinical Proteomic Tumor Analysis Consortium (CPTAC) is a database similar to TCGA, except that this database focuses on proteomic data. Using data from TCGA and CPTAC, researchers can analyze a problem from a “multi-omic” perspective, meaning they address a problem from a genomic, transcriptomic, epigenomic, and proteomic perspective. This often leads to realizations that could not have been possible if a researcher was simply looking at just the genome. This data can then be used in statistical analyses and survival plots. We manipulated and analyzed this data using both R and python.

This project consists of a multi-comic analysis of GBM patient data from TCGA, focused on identifying differences between patients who did and didn’t survive. We suspect that there will be significant differences in the genomic, transcriptomic, proteomic, and clinical profiles between surviving patients and non-survivor patients. Our results can provide future resources for the continued study of GBM brain cancer at the genomic, transcriptomic, proteomic, and clinical level.

Methods

Brain cancer clinical data was accessed from TCGA using code “TCGA-GBM” with R in RStudio. For Figure 1a, all patients with NA values in clinical information for age were excluded. For Figure 1b, all patients with NA values in the clinical information of sex were removed. For Figure 1c, all patients with NA values in the clinical information of race were excluded. For each, using the clinical data of “days to death” and “days to the last follow-up”, a survival time dataset was created. Then, a death event dataset was created using each patient's vital status. For each figure, a Kaplan Meier plot was created based on age, sex, and race, respectively. Both these clinical variables sort the patients into manageable yet large pools to base other plots on. Additionally, a DESeq2 analysis was run on GBM patients using clinical, genetic, and transcriptomic data and sorted by survivors and non-survivors to demonstrate the differing levels of transcriptome expression. Protein data was downloaded from CPTAC.

Onocoplots, lolliplot plots, and mutation co-occurance heat maps were created using the maftools R library. Boxplots of relative protein expression were created with python using the pandas, numpy, and matplotlib libraries.

Results

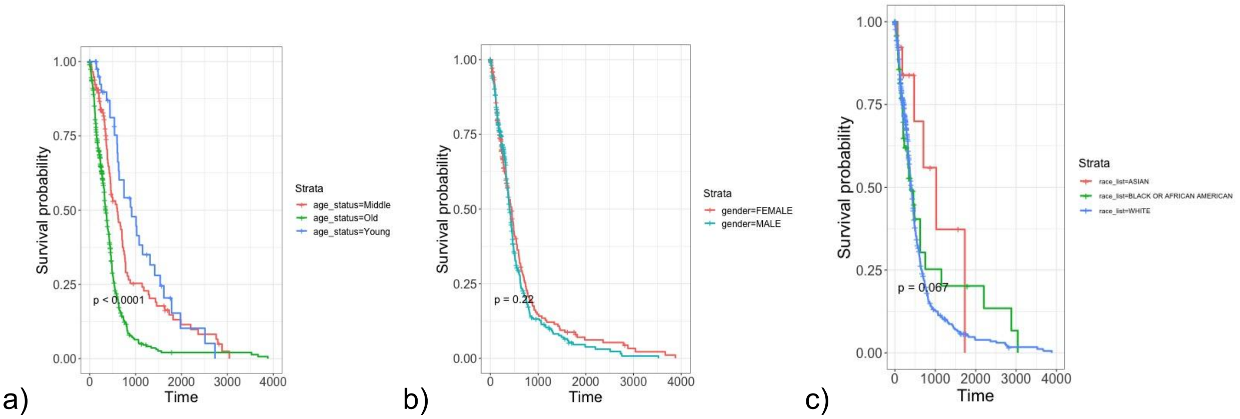

Figure 1. Kaplan-Meier survival plot by age, sex, and race of patients. In a), young patients were below 36, middle-age patients were between 36 and 50, and old patients were above 50. b) shows survival by patient sex. c) shows survival by reported race.

We began analysis by looking at how clinical features of GBM patients could be related to patient vital status. In Figure 1a), a kaplan-meier survival plot stratified by age demonstrates a statistically significant ($p < 0.0001$) difference between the survival rate of young (<36 years), middle-age(36-50 years), and old (>50 years) patients, corresponding to the patient's age at initial diagnosis. The KM plot in Figure 2a) suggests that survival rate does not differ with sex, however, with a p-value of 0.22. In addition, Figure 1c) shows that white patients have the lowest survival rate, followed by black or African-American patients, and finally Asian patients, with a p-value of 0.067. However, due to the small size of the African American and Asian population within the data and the p-value above a standard threshold of 0.05, we cannot confidently conclude that there is any difference in GBM survivability by race using the data avaliable in TCGA.

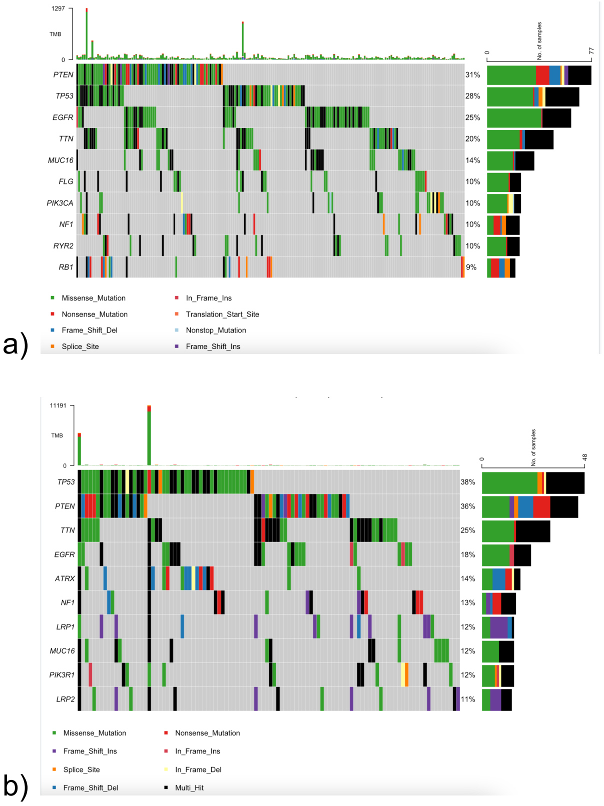

Fig 2. Oncoplot for GBM patients survivros (a) and non-survivors(b).

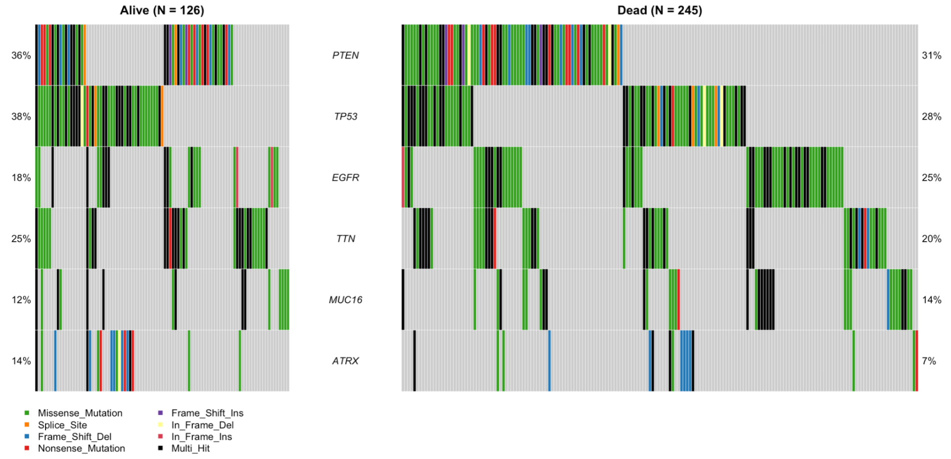

Fig 3. Co-oncoplot shows significant differences in commonly mutated genes between patients who survived and died from GBM.

Between survivor and non-survivor GBM patients, there are similarities in the genes with the most frequent mutations (PTEN, TP53, EGFR, TTN) (Figure 3), but each respective group has a number of unique genes with frequent mutations. The order of which genes were most commonly mutated was also slightly different between survivor (Figure 2a) and non-survivor (Figure 2b) GBM patients, though the top four genes are PTEN, TP53, EGFR, and TTN for both.

Even among the similar genes of these groups, there are differences in the location of the mutation, the frequency of the mutations, and the type of mutation. The most frequent type of mutation between both groups was a missense mutation (Figure 2a, 2b).

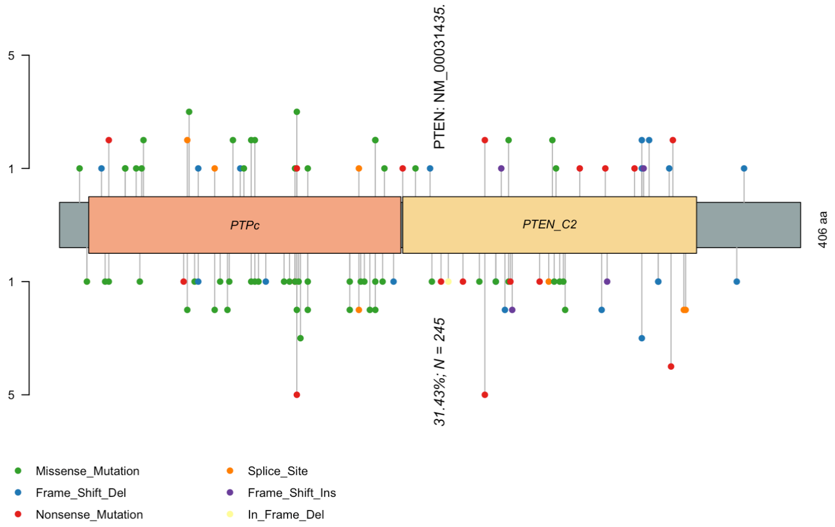

Fig 4. A lollipop plot of PTEN's phosphatase and C2 domain show an increase in mutations among non-survivor GBM patients, especially in the PTPc region. Patients who died from GBM (bottom half) seem to have 3 common nonsense mutations in PTEN that appear at much higher rates than in patients who survived.

We were also interested in potentially identifying any mutation differences between survivors and non-survivors for individual genes. We decided to look at PTEN, as the most commonly mutated gene across all GBM patients and the most common within survivors. As seen in Figure 4, there appears to be distinct differences between the mutation profile of surivor and non-survivor patients. We saw a larger number of mutations in PTEN among non-survivors, especially in the protein tyrosine phosphatase catalytic (PTPc) domain. Addditionally, there was a larger prevelence of nonsense mutations within patients who passed away due to GBM, suggesting that a change in the function of the PTEN protein or inviability of the protein product from a mutated PTEN could impact the survivability of GBM.

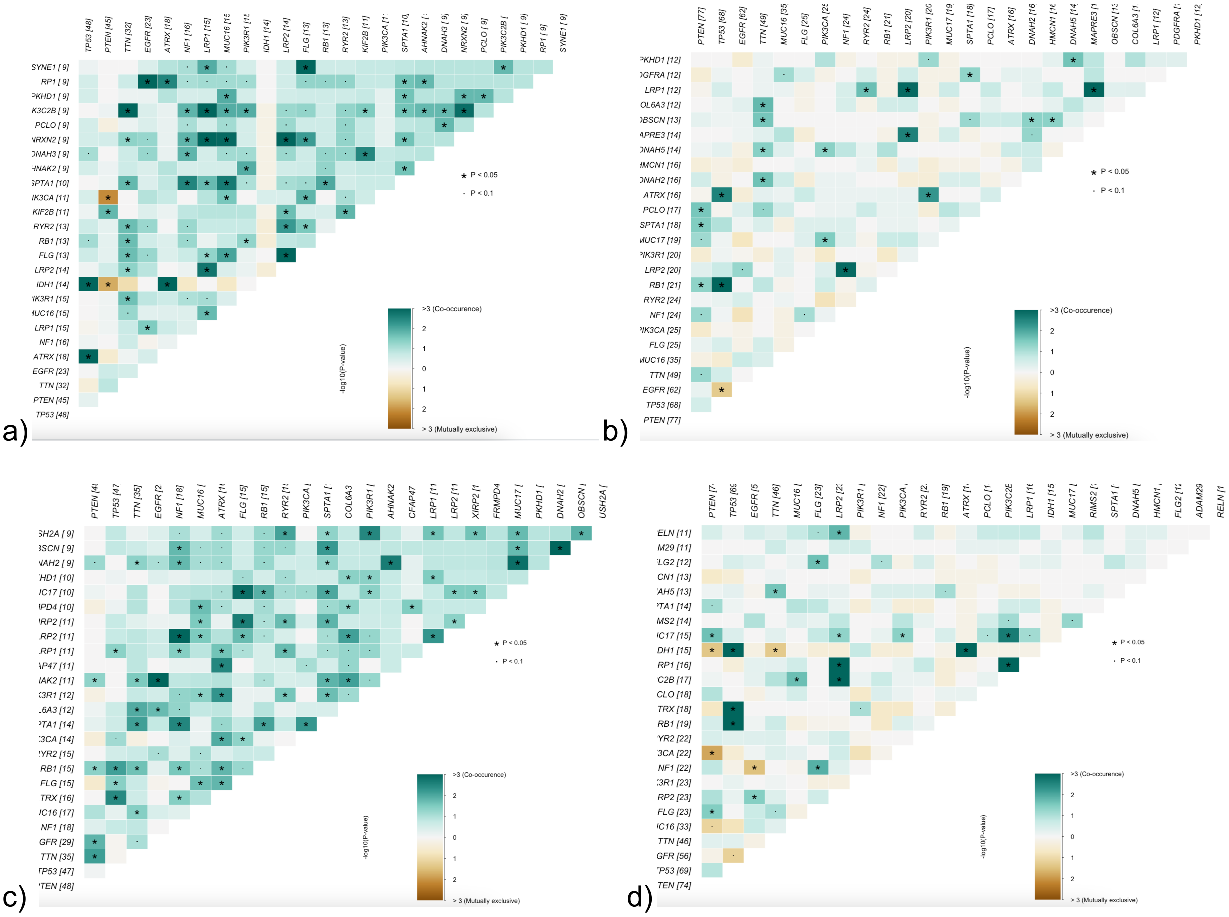

Fig 5. Mutation co-occruance plots of a) survivor patients, b) non-survivor patients, c) female patients, and d) male patients.

Another aspect of the GBM patient mutation profile that we wanted to look at was the co-occruance of common mutations with mutations on other genes. In Figure 5, we can see that patients who survived (Figure 5a) appeared to have a higher rate of mutation co-occurance when compared to patients who died from GBM (Figure 5b). Why this is the case and whether there are downstream effects are questions that may be analyzed further. As an aside, we also noticed a similar difference in the level of mutation correlation between male and female patients. Mutation co occurrence of top 20 genes for female patients (c) and top 20 genes for male patients (d) shows that there appears to be a smaller number of significantly correlated mutation occurrences in male patients compared to female patients. Given that Figure 1b) demonstrates no significant difference between the survival rates of male and female patients, we expected that there wouldn't be a difference in the mutation co-occurance profiles stratified by sex. The fact that we do see one is a point of interest for future work.

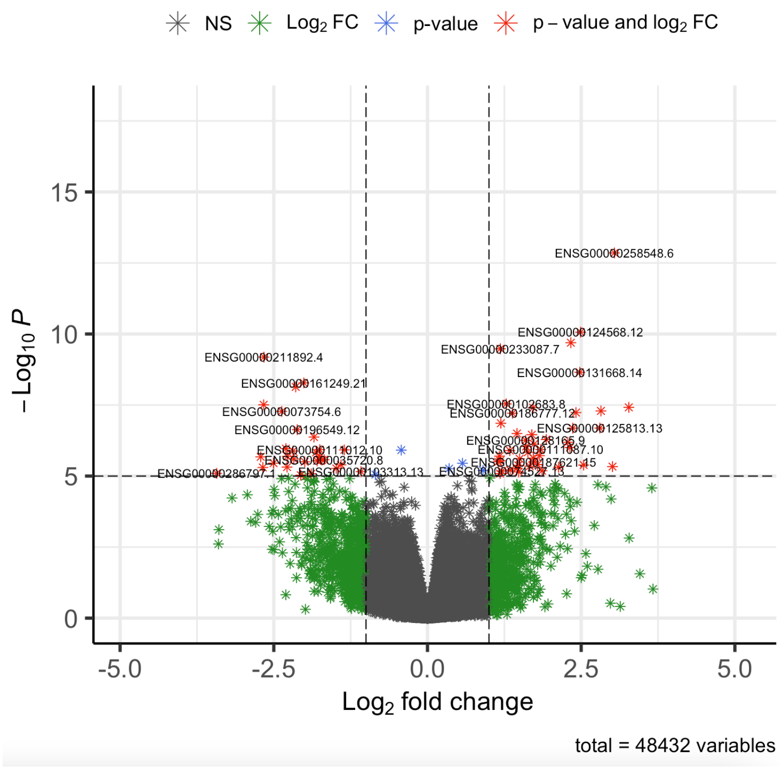

Figure 6. Volcano plot differing by vital status. Red markers demonstrate statistically significant differences in gene expression. A positive log2 fold change indicates higher rate of expression in patients who survived, and a negative log2 fold change indicates lower expression in patients who survived.

The volcano plot in Figure 6 reveals that there are many genes that are statistically-significantly upregulated in survivors when compared to non-survivors, as seen by the red markers with a positive log2 fold change. Many genes are also statistically significantly upregulated in non-survivors when compared to survivors, as seen by the red markers with a negative log2 fold change. PAX1, which corresponds to gene ID ENSG00000125813.13, is the most up-regulated gene for survivors, while AC009315.1, which corresponds to gene ID ENSG00000286797.1, is the most up-regulated gene in non-survivors. The second most up-regulated gene in survivors is LINC00645, which is non-coding and likely has regulator functions. A complete table of statistically significant genes as labeled in the Figure 6 volcano plot can be found in Table 1.

ENSG00000258548.6	LINC00645
ENSG00000124568.12	SLC17A1
ENSG00000233087.7	RAB6D
ENSG00000131668.14	BARX1
ENSG00000102683.8	SGCG
ENSG00000186777.12	ZNF732
ENSG00000125813.13	PAX1
ENSG00000128165.9	ADM2
ENSG00000111087.10	GLI1
ENSG00000187621.15	TCL6
ENSG00000074527.13	NTN4
ENSG00000211892.4	IGHG4
ENSG00000161249.21	DMKN
ENSG00000073754.6	CD5L
ENSG00000196549.12	MME
ENSG00000111012.10	CYP27B1
ENSG00000035720.8	STAP1
ENSG00000286797.1	AC009315.1
ENSG00000103313.13	MEFV

Table 1. Table of genes where RNA expression is statistically significant according to our DESeq2 analysis. Blue signifies up-regulation in survivors and red signfiies up-regulation in non-survivors.

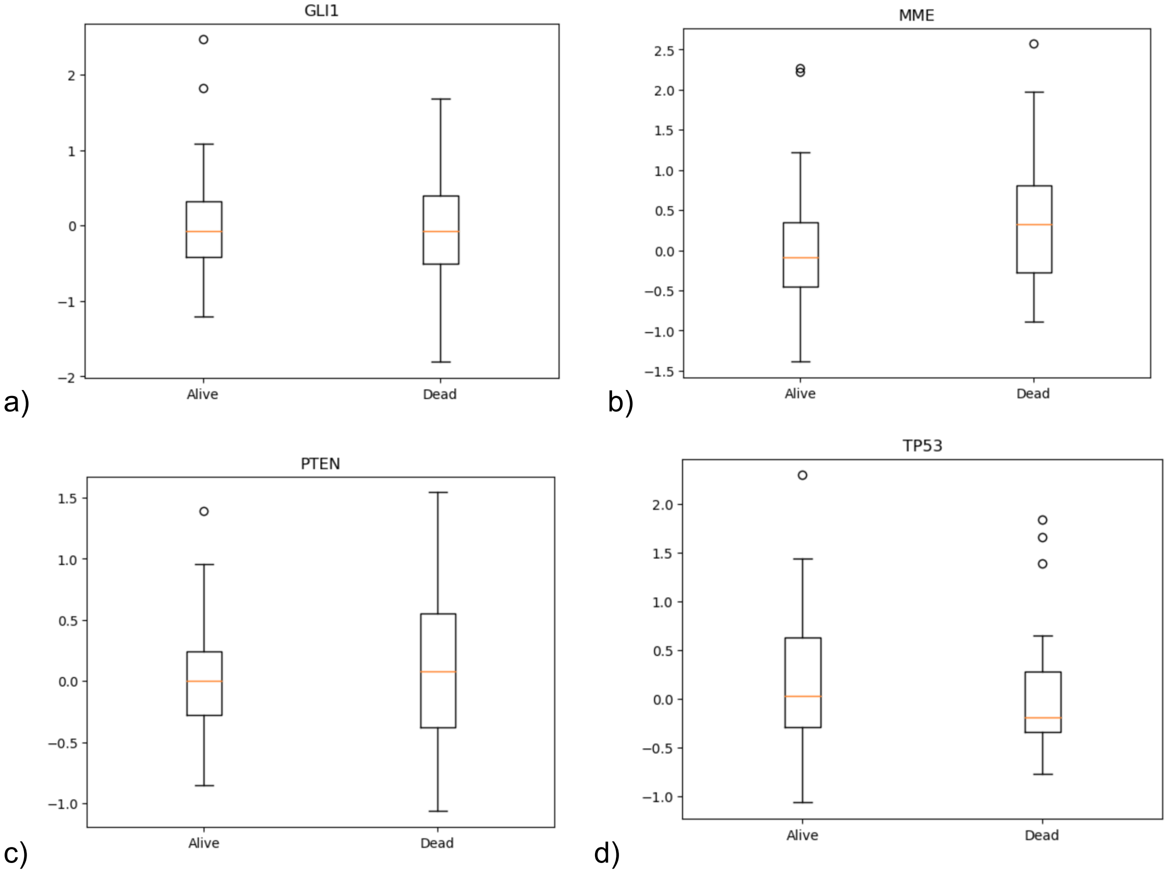

Figure 7. Boxplots of relative protein expression stratified by patient vital status. a) GLI1 and b) MME are up-regulated and down-regulated genes in survivor patients, respectively. c) PTEN and d) TP53 are the top two most commonly mutated genes in both survivor and non-survivor patients.

Finally, we looked at the protein expression of different genes separated by vital status of the patients. We first wanted to see whether or not commonly mutated genes identified in Figure 2 resulted in proteomic-level differences. In Figure 7c), we can see that there doesn't appear to be very much difference in relative expression of the PTEN protein, though the median is slightly higher in patients who passed away. This is interesting given the previously established difference in mutation profile of the gene between survivor and non-survivor patients. TP53 appears to have a little more of a difference in relative expression, as seen in Figure 7c), where non-survivor patients have a smaller spread of TP53 expression than survivor patients do and the non-survivor median is slightly lower. However, the biological significance of this difference is not obvious.

We see a similar difference in range of expression for GLI1 in Figure 7a), though this time survivor patients had a smaller range. GLI1 is of interest because the gene appears to be up-regulated in survivor patients, as established in Figure 6. Of the four genes, MME is the only one where the median relative protein expression appears to be different between survivors and non-survivors. As seen in Figure 7b), MME's median relative protein expression in non-survivors patients is around 0.25, whereas in survivors, the median is approximately -0.1. This difference also lines up with our transcriptomic analysis results, as MME is an up-regulated transcript in non-survivor patients.

Discussion

Overall, our analysis of GBM's multi-omic features demonstrates notable differences in genomic, transcriptomic, proteomic, and clinical properties between survivor and non-survivor GBM patients. During our investigation, we also discovered many non-coding genes were frequently mutated in GBM survivors and non-survivors. In particular, PTEN's increased rate of mutations within the PTPc domain among non-survivor patients supports previous research that has established the PTEN protein's phosphatase activity as vital to regulating cell proliferation and genetic stability (Hopkins et al., 2014). A high rate of mutations in the PTPc domain would be expected to decrease the efficacy of any PTEN phosphatase product, increasing the proliferation of tumor cells and the aggressiveness of GBM. The increased rate of mutations within non-survivor patients' PTEN PTPc domain supports this idea. The lack of expression continuity between many genes such as PTEN, TP53, and GLI1 also suggest that regulatory elements play a much larger role in GBM mortality than our multi-omic analysis can detect. Investigating the effects of the highly up-regulated non-coding genes such as LINC00645 and AC009315.1 in a future investigation could provide further insight into the survival probabilities of Glioblastoma Multiforme patients. LINC00645 in particular has also already been identified in previous studies as a potential biomarker and drug target for GBM due to its role in tumor proliferation (Li et al., 2019).

Some other genes of note that our analysis has identified include PAX1. PAX1 is a member of a family of genes involved in the regulation of neural crest cell apoptosis and proliferation. Different PAX family genes appear to impact cancer development differently, as previous studies have linked overexpression of PAX3 with increased tumorigenesis (Angelopoulou et al. 2019). Based on the upregluation of PAX1 in our analysis, the PAX gene likely has an opposite impact in GBM tumorigenesis. GLI1 is another oncogene identified in our differential expression analysis that has also been previously identified as a potential drug target (Avery et al. 2021). The gene MME, which was the only gene in our analysis that showed continuity across RNA and protein expression, has been shown in previous studies to be correlated with radioresistent GBM, which supports the increased expression rate we see within non-surviving patients (Nguyen et al. 2018).

Our experiment could be improved as we were lacking data in certain demographics, specifically African-American/Black and Asian patients. A future investigation comparing survival rates of different populations and then comparing their genetic profiles could help determine whether some ethnic groups are less impacted by GBM, or whether this was due to a lack of data. This could also be remedied in our investigation by collecting a larger data set, with a more even distribution of patients of various ethnicities and analyzing the new data. More data could also generally increase our confidence in our findings and develop a better understanding of glioblastoma multiforme.

References

Angelopoulou, Efthalia, et al. “Emerging Pathogenic and Prognostic Significance of Paired Box 3 (PAX3) Protein in Adult Gliomas.” Translational Oncology, vol. 12, no. 10, 2019, pp. 1357–1363., https://doi.org/10.1016/j.tranon.2019.07.001.

Avery, Justin T., et al. “GLI1: A Therapeutic Target for Cancer.” Frontiers in Oncology, vol. 11, 2021, https://doi.org/10.3389/fonc.2021.673154.

“The Cancer Genome Atlas (TCGA).” *Genome.gov*, https://www.genome.gov/Funded-Programs-Projects/Cancer-Genome-Atlas.

Förster, Alisa, et al. “Rare germline variants in the E-cadherin gene *CDH1* are associated with the risk of brain tumors of neuroepithelial and epithelial origin.” Acta Neuropathologica, vol. 142, 2021, pp. 191-210., https://doi.org/10.1007/s00401-021-02307-1

Gan, Hui K et al. “The epidermal growth factor receptor variant III (EGFRvIII): where wild things are altered.” *The FEBS journal* vol. 280,21 (2013): 5350-70. doi:10.1111/febs.12393

Hopkins, Benjamin D et al. “PTEN function: the long and the short of it.” Trends in biochemical sciences vol. 39,4 (2014): 183-90. doi:10.1016/j.tibs.2014.02.006

Li, C., Zheng, H., Hou, W. *et al.* Long non-coding RNA linc00645 promotes TGF-β-induced epithelial–mesenchymal transition by regulating miR-205-3p-ZEB1 axis in glioma. *Cell Death Dis* 10, 717 (2019). https://doi.org/10.1038/s41419-019-1948-8.

Lin, Dongdong, et al. “Trends in Intracranial Glioma Incidence and Mortality in the United States, 1975-2018.” Frontiers in Oncology, vol. 11, 2021, https://doi.org/10.3389/fonc.2021.748061.

Nguyen, Ha S et al. “Molecular Markers of Therapy-Resistant Glioblastoma and Potential Strategy to Combat Resistance.” International journal of molecular sciences vol. 19,6 1765. 14 Jun. 2018, doi:10.3390/ijms19061765

Rasheed, Sumbal, et al. “An Insight into the Risk Factors of Brain Tumors and Their Therapeutic Interventions.” Biomedicine &amp; Pharmacotherapy, vol. 143, 2021, p. 112119., https://doi.org/10.1016/j.biopha.2021.112119.

Shen, Yaoqing, et al. “Comprehensive Genomic Profiling of Glioblastoma Tumors, Btics, and Xenografts Reveals Stability and Adaptation to Growth Environments.” *Proceedings of the National Academy of Sciences*, vol. 116, no. 38, 2019, pp. 19098–19108., https://doi.org/10.1073/pnas.1813495116.

Wang, Haiwei, et al. “Analysis of the EGFR Amplification and CDKN2A Deletion Regulated Transcriptomic Signatures Reveals the Prognostic Significance of spats2l in Patients with Glioma.” *Frontiers in Oncology*, vol. 11, 2021, https://doi.org/10.3389/fonc.2021.551160.

Yang, Jr-M, et al. “Characterization of PTEN Mutations in Brain Cancer Reveals That PTEN Mono-Ubiquitination Promotes Protein Stability and Nuclear Localization.” *Oncogene*, vol. 36, no. 26, 2017, pp. 3673–3685., https://doi.org/10.1038/onc.2016.493.