Machine learning and statistical analysis to identify determinants of survival in primary glioblastoma using genome wide expression studies

Dissertation

for the award of the degree
"Doctor rerum naturalium"
of the Georg-August-Universität Göttingen

within the doctoral program Environmental Informatics (PEI)
of the Georg-August University School of Science (GAUSS)

submitted by

Manasa Kalya Purushothama
from Tumkur, Karnataka, India

Göttingen, 2022

Thesis Committee

Prof. Dr. Tim Beißbarth
Department of Medical Bioinformatics,University Medical Center Göttingen (UMG)

Prof. Dr. Stephan Waack
Institute for Computer Science ,Georg-August-University of Göttingen

Prof. Dr. Edgar Wingender
Department of Medical Bioinformatics,University Medical Center Göttingen (UMG)

Members of the Examination Board

1*st* Referee: Prof. Dr. Tim Beißbarth
Department of Medical Bioinformatics,University Medical Center Göttingen (UMG)

2*nd* Referee: Prof. Dr. Stephan Waack
Institute for Computer Science ,Georg-August-University of Göttingen

Further members of the Examination Board

Prof. Dr. Winfried Kurth
Department Ecoinformatics, Biometrics & Forest Growth,Büsgeninstitut, Georg-August University of Göttingen

Prof. Dr. Burkhard Morgenstern
Institute for Microbiology and Genetics, Dept. of Bioinformatics, Georg-August University of Göttingen

Prof. Dr. Wolfgang May
Institute Institute for Informatics, Georg-August University of Göttingen

Date of oral examination: March 4*th*, 2022

Abstract

Glioblastoma (GBM) is the most common and highly aggressive brain tumor. GBM poses unique challenges due to high rates of tumor recurrence (34%) and resistance to treatment. Despite advances in treatment strategies, the median survival of glioblastoma patients has not improved beyond 12 - 15months. Patients who survive less than 12 months are considered as Short-Term Survivors (STS). Despite all the challenges associated with the disease, there are a small group of patients who survive longer than 3 years and are termed as Long-Term Survivors (LTS). Researchers in the area continue to be perplexed by this group of patients since investigations on clinical, radiological, histological, and genetic features have failed to provide consensus on predictors of long-term response to current treatment.The goal of this study is to identify crucial survival factors in GBM and to elucidate the molecular processes that drive poor survival using gene expression profiles. To achieve this, I have used different computational approaches which are discussed in detail in the four chapters of my thesis.

The time-to-event analysis helps to investigate the impact of any clinical or molecular factor on survival and is hence also called survival Analysis. Survival analysis is performed on 2309 GBM patients aggregated from 14 publicly available datasets in Chapter 1 to see impact of various clinical and molecular factors like Age, Gender, Karnofsky Performance Score, IDH mutation status, and MGMT promoter methylation status on survival. A meta-analytic approach is considered to integrate the observations using the random effects model. Age and MGMT promoter methylation status were found to be factors of prognostic importance. This work also signifies the importance of Age as well as quantifies the risk of death that a patient experience based on his age group. Example,patients aged beyond 70 years were found to experience 2.4 times higher risk of death/ poor survival than younger patients (40 - 50 yrs).

In the next approach,I [Kalya et al., 2021a] investigated gene-expression signatures that had an impact on survival Chapter 2. 720 genes were found to impact survival according to univariate cox regression and are reported. The enrichment analysis has revealed beta-catenin network, hypoxia pathway, IL-6 signaling pathways, cell-cell communication, Interferon signaling pathways which are known for their diverse role in GBM biology. Gene-regulatory networks were built using state-of-art promoter analysis and pathway analysis using the GenomeEnhancer pipeline. This analysis revealed 43 master regulators which regulate the signal transduction networks driving prognosis in glioblastoma. Upon the information available in the HumanPSD^{TM} database, we find that some of these targets have actionable drugs reported at multiple stages of clinical trials.

To investigate the molecular mechanism underpinning poor prognosis and short survival in Glioblastoma, gene-regulatory networks were built on the genes differentially upregulated (LogFC > 0.5) in short-survivors of glioblastoma Chapter 3 [Kalya et al., 2021b]. Regulatory networks are built on the network feedback loops of Walking pathways described earlier. 12 Transcription Factors including NANOG, PPARG, FRA-1 and others were found enriched in promoters of these dysregulated genes. Graph analysis of the signal transduction network upstream of these transcription factors revealed five potential master regulators that could explain gene dysregulation in short-survivors: insulin-like growth factor binding protein (IGFBP2), vascular endothelial growth factor A (VEGF-A), its isoform VEGF165, platelet-derived growth factor A (PDGFA), oncostatin M (OSMR), and adipocyte enhancer-binding protein (AEBP1). All of the identified master regulators were elevated in STS, and their expression patterns were computationally verified in two additional independent cohorts. This work proposes a novel mechanism of gene dysregulation by IGFBP2 by modulating a key molecule of tumor invasiveness and progression - FRA-1 transcription factor.

Machine Learning has now become an indispensable tool in GBM research. Chapter 4 [Kalya et al., 2022] evaluates application of 10 ML models to build a classifier which can classify GBM patients into short-term and long-term survivor groups based on their transcriptomic profiles and clinical information (age). A random forest model with an F1 score of 86.4% (Accuracy = 80%, AUC =74%) is proposed (with good external validity). This classification model is deployed as a webtool. The important features are discussed for their biological relevance in the disease using gene ontology analysis, survival analysis, differential expression and by mapping them onto existing databases of gene-disease biomarkers associations. Using this approach, we have identified 199 mRNAexpression based biomarkers that are associated with survival group prediction. Of them, 171 are not reported to be biomarkers of glioblastoma in existing databases.

In conclusion, this work evaluates clinical factors associated with prognosis using a meta-analytic approach. This work proposes 242 gene-expression based biomarkers associated with GBM survival using different computational approaches. Molecules like PDGFA, AEBP1 and VEGF were found to be important in both gene-regulatory network analysis and in Machine Learning models. A novel mechanism of gene dysregulation in short-term survivors via FRA-1 transcription, a key molecule of tumor invasiveness and progression is proposed. The thesis encompasses 2 published research papers and a ready to submit version of the research article.

Contents

Listing of figures

List of Tables

Acronyms

GBM - Glioblastoma

STS - Short-term survivor

LTS - Long-term survivor

KPS - Karnofsky Performance Score

MGMT - Methyl Guanine Methyl Transferase

IDH - Isocitrate Dehydrogenase

OS - Overall Survival

DNA - Deoxyribonucleic acid

mRNA - messengerRNA

TCGA - The Cancer Genome Atlas

CGGA - Chinese Glioma Genome Atlas

GEO - Gene Expression Omnibus

AUC - Area Under the Curve

TMZ - Temozolomide

Look to this day,
for it is life, the very breath of life.
In its brief course lie
all the realities of your existence;
the bliss of growth,
the glory of action,
the splendor of beauty.
For yesterday is only a dream,
and tomorrow is but a vision.
But today, well lived,
makes every yesterday a dream of happiness,
and every tomorrow
a vision of hope.
Look well, therefore, to this day.
(Ancient Sanskrit)

Acknowledgments

Take up one idea. Make that one idea your life.
Think of it, dream of it, live on that idea.
Let the brain, muscles, nerves, every part of your body, be full of that idea,
and just leave every other idea alone.
This is the way to success – Swami Vivekananda

I grew up listening to these words of Swami Vivekananda in Ramakrishna Vivekananda-Ashrama , Tumkur. It is a true man-making workshop. One such idea that has been simmering in my brain since I was a child is to pursue a Ph.D. As a young girl interested in science, the life of Marie Sklodowska-Curie inspired me a lot. It counts as a blessing that I was chosen for a prestigious fellowship in her name. I would like to thank my fellowship program 'GlioTrain'- EU Horizon 2020 Research and Innovation Program under Marie Sklodowska-Curie Actions (MSCA), for funding as well as for the incredible career-oriented training programs given to young researchers like me.

Prof. Tim Beißbarth, my supervisor, deserves my heartfelt appreciation for his unwavering support and boundless guidance during my Ph.D. He patiently mentored me and saw to the conclusion of my thesis.

I would like to whole-heartedly thank my mentor, Dr. Alexander Kel, for believing in me and for this wonderful opportunity. His unwavering support, counsel, and patience over the years have enabled me to sail through my voyage with ease.

I will always be indebted to Prof. Edgar Wingender for his contributions. His scientific knowledge and student-centric approach makes him a great supervisor and I am fortunate enough to work under his supervision. I appreciate his help in fine-tuning my scientific communications as well as his reassuring endorsement letters.

I'd like to express my gratitude to my other supervisor, Prof. Dr. Stephan Waack, for his

unwavering support and availability during this process. Prof. Winfried Kurth, the Dean of Studies for the PEI program, is the most helpful person you could wish for. His enthusiastic participation and guidance aided me in smoothly navigating the administrative formalities.

I'd like to thank Dr. Rakshit Dadarwal, a personal friend, for his timely aid in making this thesis format a success. I'd like to thank my anchor, Darius Wolchwitz, who helped me with more than just university processes during my Ph.D. I'd like to express my gratitude to Dr. Andreas Leha for his prompt assistance with research and for taking me into consideration for a vacant job.

I'd like to express my gratitude to my lovely colleagues at geneXplain, who have made me feel at home when I'm away from my home. I'd like to express my gratitude to Prof. Tim Beißbarth's group, which was incredibly welcoming, warm, and upbeat. I'd want to thank all of my friends especially Jhenkhar Mallikarjun, Rahul Agrawal and Dr. Parth Joshi for their constant support in keeping me in shape through these trying times. I'd like to thank my great GlioTrain family for making all of the seminars, conferences, review meetings, and training activities so unforgettable.

I'd like to express my gratitude to Dr. Holger Michael, geneXplain, and Dr. Alice O'Farrell GioTrain coordinator, for their efforts in designing and executing this massive journey. I'd like to express my gratitude to my current colleagues of the Computational Biology group at Evotec International GmbH, Gottingen for their invaluable assistance and understanding during my thesis submission process.

I'd like to thank my previous employers and research supervisors (Dr. Odity Mukherjee, Dr. Biju Vishwanath and Dr. Ramkrishnan Kannan) at the National Institute of Mental Health and Neurosciences, National Centre for Biological Sciences, Bangalore, for training me, encouraging my research curiosity, and providing me with exciting opportunities to advance my science career.

Dear Dr. Ravi Kumar Nadella, over the past four years, you have been my rock and a safe set of hands. You've been a part of every step of my scientific and personal development. After my brother, no one has contributed as much to my accomplishment as you have. Thank you is such a little phrase to express my gratitude for everything you've taught me. I believe that if there is support as strong as yours, everything and anything can be accomplished. Thank you for your presence. I appreciate your family's unwavering support.

Thank you for always being there, my dearest sisters Divya Santosh, Ramya V kallur, Rashmi

V kallur, and my best brother Rakshit Anantha Krishna, my vital support system. I'd like to thank Dr.Mahesh Murthy and Mrs.Bharathi Mahesh who are always there by my side. Thank you for being my fortunate charm, Harsha, and please continue to lead me through all of my dilemmas. My brother, Mahadev Prasad, is waiting for me in India, ready to indulge in endless chitchat and mischief. Vyshakh Nag's attention and aid were vital during tough times, and I'd like to express my thanks to him. I cherish the compassion and love that I have received from you and your family.

I'd like to thank my gurus, Swami Veereshananda Saraswathi and Swami Nirbhayanada - Saraswathi, for their immeasurable blessings, which are my biggest source of strength. I am grateful to Swami Paramananda for his reassuring words and counsel. I am extremely grateful to siddaguru Pushkara Purvajna for initiating an extraordinary journey. I will treasure the love and enthusiasm that the children of Shree Sharadadevi Satsangha Kendra, Tumkur, have shown me.

The acknowledgment will not be complete without mentioning the two most important people of my life. One, my father figure who planted the seed of dreaming big, discipline, and value of higher education -B.N. Srinivasa Murthy. Another one, a teacher who held my hand and helped me through a slew of challenges, gave me the best help he could. My life would undoubtedly be incomprehensible if it weren't for him. I offer my sincere gratitude to my tough brother - Sriprasad Kalya. I am grateful for the many blessings that I have received from everyone.

ನಿನ್ನ ಒಲುಮೆಯಿಂದ ...

ಎನ್ನ ಪುಣ್ಯಗಳಿಂದ.. ಈ ಪರಿ ಉಂಟೇನೋ ..

ನಿನ್ನದೇ ಸಕಲ ಸಂಪತ್ತು ... ದೇವಾ ।।

My way of expressing gratitude to the divine...

This work and effort is dedicated to my deep self, my inspiration, my parents, my lovely family full of brilliant individuals, and a particular tribute to my wonderful grandparents (S.Subba Rao- Rajarathna and Mohan Murthy) who have bestowed all of their excellent traits upon me and provided me with priceless memories to cherish for the rest of my life.

Nothing in life is to be feared, it is only to be understood. Now is the time to understand more, so that we may fear less.

Marie Curie

0 Introduction

Glioblastoma (GBM) is one of the most common and highly malignant brain tumors of the central nervous system. It corresponds to 54% of all the adult brain tumors with a global incidence of 10 in 10,000. The incidence is found to be higher in men as compared to women, male-to-female ratio of 1.33 for primary GBMs [Wen & Kesari, 2008, Thakkar et al., 2014]. Even though the incidence rate is low, it is a very important public health issue because of its poor prognosis with a median survival of 15 months [Iacob & Dinca, 2009, Thakkar et al., 2014, Gilard et al., 2021]. Despite improvements in the treatment strategies, less than five percent of patients survive beyond 5 years after diagnosis [Verhaak et al., 2010, Hanif et al., 2017, Henson, 2006].

The coming sections give an overview about Clinical Presentation, Diagnosis and Management aspects of Glioblastoma

0.1 Clinical presentation, diagnosis and management

0.1.1 Clinical Presentation

Clinical presentation of GBM depends on the size and location of the tumor. Most of them are located in the supra-tentorial space. Approximately one-fourth of tumors are in the frontal lobe resulting in deficits of mood and executive abilities. The incidence of tumors in other regions are 20% in temporal lobe, 13% in parietal and 3% in occipital lobe [Davis, 2016]. In

terms of symptoms (Figure 1), nausea and headache are some of the nonspecific symptoms which are seen with large sized tumors. Motor deficit, loss of body weight, confusion, speech or visual deficit are expressed in cases with intracranial hypertension. Initial presentation with epilepsy is not uncommon. 15-20% patients in GBM present with seizures and are associated with better prognosis as it helps in faster diagnosis Figure 1.

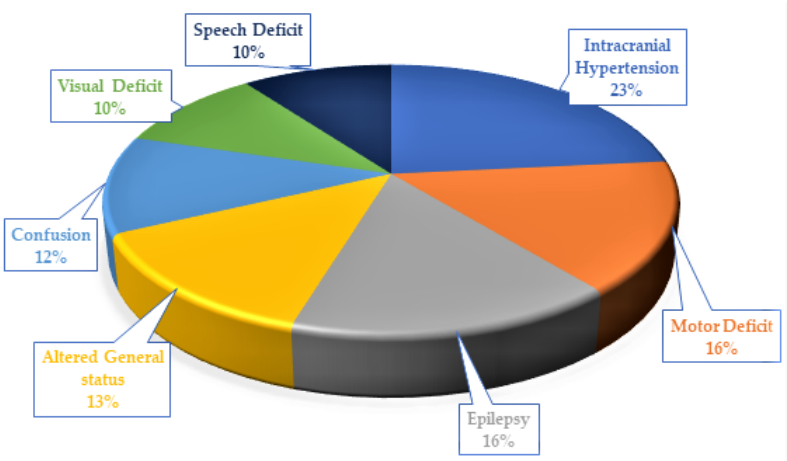

Figure 1: Clinical Presentation of GBM

0.1.2 Diagnosis

In magnetic resonance imaging (MRI) scans, nearly all GBMs show enhancement with gadolinium contrast and show an irregularly shaped mass with a dense ring of enhancement (Figure 2). In T1-weighted images a hypointense centre of necrosis is seen. Necrosis is a hallmark feature of GBM, and presence of necrosis is required for a brain tumor to be grade IV or to be classified as a GBM on the World Health Organization classification system (AANN, 2014). The surrounding vasogeneic edema is seen as a hyperintense signal in the T2-weighted images and fluid-attenuated inversion recovery (FLAIR) images [Ellis & Kurian, 2014, Ellor et al., 2014, Davis, 2016]. Recent multimodal MRI techniques such as diffusion/perfusion sequences give more information on lesions and enable accurate diagnosis. Perfusion weighted imaging (PWI) reveals an increase in cerebral blood flow corresponding to neoangiogenesis and blood brain barrier (BBB) disruption. In addition, the elevated peak of lactate and lipids as well as a decreased peak of myoinositol helps to discriminate glioblastomas from metastasis, lymphoma and brain abscess [Gilard et al., 2021, Hanif et al., 2017].

Figure 2: T1-gadolinium contrast-enhancing tumor of the right frontal lobe [adopted from operativeneurosurgery].

0.1.3 Management

The principal mode of treatment includes a combination of surgical resection of the tumor and/or radiotherapy and/or chemotherapy. The modality of treatment choices depend on the age of the patients, location of the tumor and Karnofsky score. Karnofsky score assesses the patients' ability to conduct everyday chores. Scores vary from 0 to 100 and a higher score means the patient is better able to carry out daily activities.
Surgery is the principal component of standard care in Glioblastoma. Studies show, tumor resection in > 90% of patients with no comorbidities improves treatment outcome and reduced recurrence. However, complete surgical resection is proposed for patients aged less than 70 and having Karnofsky score > 70. Due to the importance of a complete resection on survival, several advanced surgical techniques are developed such as fluorescence-guided surgery, Laser-Interstitial Thermal Therapy (LTT), mass spectrometry-based surgical resection, which can ensure near complete resection of the tumor. Feasibility of surgical resection also depends on the location, for e.g. sites like eloquent cortex, brain stem, or basal ganglia are not amenable to surgical intervention and these patients usually have a worse prognosis [Mrugala, 2013, Hanif et al., 2017].
Radiotherapy kills remaining tumor cells after surgical procedures. It is found to increase life expectancy in patients with high-grade glioblastoma. Radiotherapy is given for six weeks with a total dose of 60 grays. Temozolomide is an alkylating agent administered daily during the RT and then, for six cycles of five consecutive days per month, one month after the end

of the RT. This treatment protocol is famously called the Stupp protocol.
Alkylating agents like Temozolomide, Carmustine (BCNU), and Lomustine (CCNU) are effective in GBM. These drugs work by methylating the guanine nucleotide at the N7 and O6 positions which leads to forming of nicks in the DNA and subsequently blocks the cell cycle at the G2-M boundary and triggers apoptosis [Scott et al., 2011]. These methylation sites can be demethylated by Methyl Guanine Methyl Transferase (MGMT), which is encoded by MGMT gene, whose activity is associated with poor response to Alkylating agents [Hanif et al., 2017].

0.1.4 Prognosis

The median overall survival of GBM patients in population based studies is around 10-12 months even with the advances in the treatment modalities, which indicates an overall poor prognosis. A very small set of patients survive beyond 36 months who are referred to as Long term survivors (LTS). Some of the important clinical based prognostic factors are Age at diagnosis, Gender, Extent of tumor resection, Karnofsky performance status (KPS), chemotherapy, a dose of radiation, tumor location in the brain. Understanding the biological underpinnings which might contribute for long term survival might bring insight to etiopathogenesis and drug development. Several studies have tried to examine the genetic alterations of GBM and develop different classifications which are based on their molecular phenotyping. In the following section, I shall briefly discuss approaches used to classify the GBM tumors using different omics data.

0.2 Genetic and Molecular Pathology

Some of the molecular alterations which are considered to be hallmark in primary GBM include EGFR mutation and amplification, MDM2 overexpression, deletion of p16, and loss of heterozygosity (LOH) of chromosome 10q holding phosphatase and tensin homolog (PTEN) and TERT promoter mutation. As an example molecular phenotyping of GBMs using isocitrate dehydrogenase (IDH) mutation status into IDH mutant and IDH wild type is frequently used; and 90% of cases who are over 55 years of age show IDH wild type. IDH alterations and MGMT hypermethylation are associated with longer survival, whereas Telomerase Reverse Transcriptase promoter (TERTp) variants and chromosome 10 deletion are associated with short survival time. Even though numerous genetic aberrations are known to be reported with GBM, mainly three signaling pathways are found to be playing a significant role in pathogenesis. These three pathways are 1) Receptor Tyrosine Kinase (RAS/PI3K) pathway which is altered in more than 88% of GBMs, 2) P53 pathway which is altered in 87% of cases, and 3) RB signaling pathway which is altered in nearly 80% of the GBM cases [Al-

dape et al., 2015, Hanif et al., 2017]. Using these molecular aberrations and global gene expression several GBM categorisation or typing are proposed. Such classifications which are based on molecular phenotyping can shorten the time from diagnosis to treatment and significantly improve accuracy and testing. In the next section, I shall discuss and review various classifications that are studied in the GBM.

0.2.1 Transcription based classification

Tumor development is highly complex which involves multiple genetic and epigenetic changes. Using microarray or high-throughput sequencing genes associated with GBM can be identified and used as biomarkers for early diagnosis, classification and treatment purposes [Schena et al., 1995, Irizarry et al., 2003, Nutt et al., 2003, Hu et al., 2006].

In a landmark study Phillips et al.2006 [Phillips et al., 2006] has classified GBM into three subtypes: Proneural, Proliferative, and Mesenchymal. The proneural subtype is seen predominantly in the younger age group with better prognosis. The cells are similar to neurons of normal brain tissues and have expression of NCAM (Neural cell adhesion molecule), GABBR1 (Gamma-aminobutyric acid type B receptor subunit 1), and SNAP91 (Clathrin coat assembly protein AP180). The proliferative subtype shows cells that are comparable to stem cells with expression of proliferation markers such as TOP2A (DNA topoisomerase II alpha) and PCNA (Proliferating cell nuclear antigen). The mesenchymal subtype shows overexpression of angiogenesis markers like endothelial PECAM1 (Platelet endothelial cell adhesion molecule), VEGF (Vascular endothelial growth factor), VEGFR1 (Vascular endothelial growth factor receptor 1), and VEGFR2 (Vascular endothelial growth factor receptor 2). These subtypes resemble various stages of neurodevelopment which provides a newer perspective for GBM molecular classification.

Verhaak et al.2010 [Verhaak et al., 2010] have further subdivided GBM into four subgroups Proneural, Neural, Classical and Mesenchymal which is further validated by Wang et al. The proneural subtype is noted to have strong PDGFRA gene expression and frequent IDH1 mutation. The neural subtype includes neural markers such as SYT1 (Synaptotagmin 1), SLC12A5 (Solute carrier family 12 members 5), GABRA1 (Gamma-aminobutyric acid type A receptor alpha1), and NEFL (Neurofilament light polypeptide) and are more susceptible to radiation and chemotherapy. The classical subtype has high expression of neural precursor and stem cell markers and requires aggressive radio and chemotherapy. The mesenchymal subtype has extensive necrosis and inflammation, overexpression of interstitial and angiogenesis genes and has the worst prognosis. Teo et al.2019 [Teo et al., 2019] used six different datasets to validate three GBM subtypes: Proneural/Neural, Classical, and Mesenchymal

Table 1 summarizes all the transcription profiles based Glioblastoma subtypes.

Using large-scale gene expression patterns Park et al.2019 have identified three subtypes that are associated with prognostic prediction: Mitotic, Intermediate and Invasive subtype. The Invasive subtype has much more invasiveness and has poor prognosis than the Mitotic subtype. The methylation of the MGMT gene promoter is linked to the Mitotic subtype, implying that Mitotic subtype patients are more likely to respond to temozolomide" [Park et al., 2019].

0.2.2 Genetic Alteration-Based Subtypes

The large-scale genomics has led to identification of several genetic alterations in the tumor suppressor genes and oncogenes. Some of these genetic alterations are linked to patient survival and can be used as indicators for patient classification. The strongest genetic variant that is linked is Isocitrate Dehydrogenase (IDH) mutation.

Isocitrate Dehydrogenase
In 2016, WHO has divided the GBM into IDH wild type and mutant type because of its prognostic importance. IDH is an enzyme which is involved in citric acid or Krebs cycle. It is important for both oxidative metabolism and oxidative stress response. Krebs cycle occurs both in cytoplasm and mitochondria because of which the enzyme is located in both locations. However, the isoforms are different in different locations. The IDH has three isoforms i.e IDH1-3; of which IDH1 is primarily found in cytoplasm and the other two are found in the mitochondrial matrix. In 2008 Parsons et al discovered a point mutation(R132H) in the IDH1 gene, which is the most prevalent IDH1 mutation found in the gliomas IDH-mutant GBM patients showed a greater overall survival rate and were more sensitive to temozolomide than GBM patients with wild-type IDH [Songtao et al., 2012].

0.2.3 Other mutations

A frequent mutation in Epithelial growth factor receptor (EGFR) gene is EGFRvIII [Gan et al., 2009]. EGFR plays a crucial role in cell proliferation, differentiation, and development. The EGFR gene is located on the short arm of Chromosome 7 and encodes a cell surface tyrosine kinase receptor. EGFRvIII is noted to have absence of 267 amino acids in the extracellular domain which results in the inability of the receptor to bind to its ligand [Hatanpaa et al., 2010]. EGFRvIII is shown to enhance the tumor potential by activating mitotic and anti-apoptotic signalling pathways [Gan et al., 2009] and is associated with poor prognosis.Summary in Table 2

Table 1: Glioblastoma subtypes based on transcription profiles

Phillips et al. (2006)		**Proneural**	**Proliferative**		**Mesenchymal**
	Signature	NCAM, GABBR1, SNAP91	PCNA, TOP2A, EGFR		VEGF, VEGFR1, VEGFR2, PECAM1
	Chromosome Gain/loss	None	Gain on Chr.7, loss on Chr.10		Gain on Chr.7, loss on Chr.10
	Biological process	Neurogenesis	Proliferation		Angiogenesis
Verhaak et al. (2010)		**Proneural**	**Neural**	**Classic**	**Mesenchymal**
	Signature	PDGFRA, OLIG2, DDL3, SOX2, NKX2-2	MBP/MAL, NEFL, SLC12A5, SYT1, GABRA1	EGFR, AKT2, SMO, GAS1, GLI2, NOTCH3, JAG1, LFNG	YKL40, MET, CD44, MERTYK, TRADD, RELB, TNFRSF1A
	Mutated genes	TP53, PI3K, IDH1, PDGFRA		PTEN, CHKN2, PDGFRA	NF-κB, NF1

Phosphatase and tensin homolog (PTEN) gene encodes protein that catalyzes the dephosphorylation of the inositol ring in phosphatidylinositol-3,4,5-trisphosphate (PIP3) to phosphatidylinositol-4,5-bisphosphate (PIP2). This dephosphorylation is an important step in the inhibiting AKT signalling pathway. The PI3K/AKT pathway which is usually dormant in the normal cells, when activated leads to cancer. Loss of PTEN leads to activation of the AKT pathway and is associated with aggressive phenotypes. [Endersby & Baker, 2008]

Table 2: Frequency of mutations in core genes of Glioblastoma across subtypes

	Glioblastoma Subtypes (Frequency of Mutation in %)			
Genes mutated in Glioblastoma	**Proneural**	**Neural**	**Classical**	**Mesenchymal**
TP53	54	21	0	12
NF1	5	16	5	37
EGFR	16	5	0	0
EGFRvIII	3	0	23	3
IDH1	30	5	0	0
PDGFRA	11	0	0	0

In addition, patients with CDK4/MDM2 co-amplification have a median survival rate of 6.6 months after diagnosis in IDH1 wild type GBM, while patients without CDK4/MDM2 co-amplification have a median survival rate of 12.7 months [Abedalthagafi et al., 2018]. A mutation in the TERT promoter has recently been found as an indication of poor prognosis. It is more prevalent in senior people, with around 40% of them having grade II/III glioma, implying that TERT is a crucial pathogenic factor and therapeutic target in glioma [MC & M, 2015, Yuan et al., 2016, Spiegl-Kreinecker et al., 2015].

0.2.4 DNA Methylation-Based Subtypes

DNA methylation is a critical component in facilitating carcinogenesis, as well as a core element of epigenetic modification and an important signaling tool for regulating genomic functions [Koch et al., 2018, Muhammad et al., 2018]. DNA methylation can be used to develop biomarkers for cancer diagnosis and prognosis [Lofton-Day & Lesche, 2003, Gustafsson et al., 2018]. Methylation status of single genes corresponds to expression levels in GBM [Bell et al., 2018, Johannessen et al., 2018]. MGMT promoter methylation is an important prognostic factor and is associated with increased survival [Brennan et al., 2013].

In a study using large-scale methylated sequencing data to characterize GBM, the authors have split the data into six categories based on the level of DNA methylation expression. Cluster M1 through Cluster M6, with Cluster M5 being the G-CIMP subtype. In comparison to the G-CIMP subtype, Cluster M6 is more hypomethylated and has a higher proportion of IDH1 wild-type patients. Cluster M2 was characterized by missense mutations or deletions in MLL (histone-lysine N-methyltransferase 2A) or HDAC (Histone deacetylase) family genes [Brennan et al., 2013]. These findings suggest that GBM can be classified

using a methylation profile. In a recent study, Ma et al. has identified prognostic subtypes using DNA methylation status, and identified three clusters (Cluster 1, Cluster 2, and Cluster 3), each with significantly different survival curves. Cluster 2 offers the best prognosis of all the clusters. The methylation levels in each cluster have specific molecular characteristics. Cluster 3 has shown more TP53 mutations and deletion of wildtype IDH which are related to survival and biological processes in GBM. Using the DNA methylation patterns a new prediction tool is developed for 10 CpGs. These 10CpGs are superior to other molecular markers because they reflect the relationship between the GBM subtypes [Kloosterhof et al., 2013, Paul et al., 2017, Yin et al., 2018].

Methylation is a significant complement to genetic changes and transcription-based classification, allowing for a more thorough classification of GBMs.

0.2.5 Overlap of Subtypes based on different approaches

The subtypes identified using different omics approaches are found to be related and overlapping (Table 1 and 2). The combined analysis of the four transcriptome based subtypes has shown enrichment of mesenchymal subtypes in Cluster M1 and M2, classical subtypes in M3-M4 cluster, Cluster G-CIMP and M6 belong to Proneural subtype [Verhaak et al., 2010]. The Cluster G-CIMP is noted to have increased frequency of MGMT DNA methylation (79 percent of patients with DNA methylation of MGMT in Cluster G-CIMP and 46 percent in non-G-CIMP). The C-CIMP is a distinct and nearly invariable hallmark of IDH1/2 mutant GBMs, and studies have indicated that individuals with this GBM subtype have a better prognosis [Noushmehr et al., 2010, Baysan et al., 2012]. The Proneural subtype is further split into G-CIMP positive and negative groups based on the characteristic of DNA methylation pattern causally connected to IDH1/2 mutation status [Noushmehr et al., 2010]. Summary is given in Table 3

Table 3: Overlap of multiple Glioblastoma subtypes based on different approaches

Phillips	Proneural		Proliferative		Mesenchymal
Verhaak	Proneural		Neural	Classical	Mesenchymal
Brennan	M5 (G-CIMP)	M6		M3,M4	M1,M2
Louis	IDH Mutant	IDH Wildtype			

0.3 Research Question: Surviving glioblastoma despite the odds

While the majority of GBM patients live for less than two years, there is a subpopulation of patients that live for more than three years (36 months) and are referred to as long-term survivors (LTS) [Hwang et al., 2019]. Researchers in the area continue to be perplexed by this group of patients, since investigations on clinical, radiological, histological, and genetic features have failed to provide consensus on predictors of long-term response to current treatment [Hwang et al., 2019].

Glioblastomas are divided in the 2016 CNS WHO (Central Nervous System, World health Organization) into (1) GBM, isocitrate dehydrogenase (IDH)-wildtype (about 90% of cases), which corresponds most frequently with the clinically defined primary or de novo glioblastoma and predominates in patients over 55 years of age; (2) glioblastoma, IDH-mutant (about 10% of cases), (referred to secondary GBM) with a history of prior lower-grade diffuse glioma and preferentially arises in younger patients; and (3) glioblastoma, NOS, a diagnosis that is reserved for those tumors for which full IDH evaluation cannot be performed" [Armocida et al., 2019]. With the vast molecular knowledge gained by omics technology, Glioblastoma are categorized into multiple subtypes based on transcriptional and methylation characteristics. Most recent evidence being 3 subtypes proposed by Teo et al. [Teo et al., 2019] based on transcriptional signatures and Ma et al. [Ma et al., 2020] proposed 3 subtypes based on methylation clusters. These aspects are discussed in previous sections. However, these molecular characterizations failed to explain relationship between long-term survival and membership of one of the four expression-based subclasses [Bi & Beroukhim, 2014]. Researchers have concentrated their efforts on determining the factors that indicate exceptional long-term survival. These patients' sickness path differs significantly from that of the vast majority of GBM patients. "For instance, half of patients who have survived four years will live for another four." [Bi & Beroukhim, 2014].

Here, I review the studies which have discussed clinical, molecular and radiological characteristics associated with Long-term survival in Glioblastoma.

A 14 year retrospective study of 480 GBM patients reported survival advantage of younger age, good KPS score and extent of tumor resection to be good predictors of long-term survival [Mazaris et al., 2014]. A probable association of tumor localization without SVZ contact ($p = 0,05$) was proposed as a significant factor for prolonged survival [Krex et al., 2007]. Another study reported association of LTS to have unilateral tumors and undergo multimodality treatment [Gately et al., 2018]. Long-term survival, on the other hand, remains dismal, and there appears to be little increase in 5-year survival [Poon et al., 2020].

On the grounds of statistical analyses [Armocida et al., 2019] affirm that "volume of the lesion, motor disorder at presentation and/or a Ki67 overexpression had significant survival advantage. This study has also pointed out that performing a standard molecular analysis (IDH, EGFR, p53 and Ki67) is not sufficient to predict the behavior of a GBM in regards to overall survival (OS) [Armocida et al., 2019]. LTS was found associated with a gross total resection (GTR) of tumor correlated with EGFR and p53 mutations with regardless of localization, and poorly correlated to dimension" [Armocida et al., 2019]. Another study [Hwang et al., 2019], has discovered significant changes in "DNA methylation profiles between LTS- and STS- GBMs, which are linked to oncogenic pathways via two separate mechanisms, transcriptional repression and somatic mutation, depending on their genomic position" [Hwang et al., 2019]. Systematic comparison of molecular features ("Abnormality of chromosome 1p/19q, IDH mutation and O6-methylguanine-DNA methyltransferase (MGMT) promoter), clinical and radiological characteristics between LTS and short-term survivors (STS) in the cohort of GBM sheds some light about impact of IDH status in LTS prognosis [Jiang et al., 2021]. "The authors reported that IDH-mut LTS presented a higher rate of 1q/19p co-polysomy than IDH-wt GBM. 1q/19p co-polysomy is previously reported to an independent prognostic factor associated with prolonged survival [Zeng et al., 2017]. IDH-wt LTS had a higher rate of non-local failure than that of IDH-mut LTS. This explains the favorable Progression Free Survival among IDH-wt LTS. The survival analyses demonstrated that IDH-mut LTS showed a trend towards increased survival after receiving re-operation and reirradiation, while the clinical benefits disappeared in the subset of IDH-wt" [Jiang et al., 2021]. However, Gerber et al. [Gerber et al., 2014] had concluded that "survival beyond 4 years does not require IDH mutation and is not dictated by a single transcriptional subclass. In contrast, MGMT methylation continues to have strong prognostic value for survival beyond 4 years. These findings have substantial impact for understanding GBM biology and progression".

A study Ferguson et al. [Ferguson et al., 2021] has built a survival-predictive nomogram based on clinical factors like Age, Sex, KPS Clinical trial participation; Molecular characteristics like PDL1 status (+/-), Tumor mutational burden, MGMT methylation(+/-) ; Radiological features like Extent of Resection (Gross Total Resection, Sub Total Resection, Surgical approach (craniotomy/surgery)), Tumor necrosis Volume, Volumetric EOR based on enhancement, Volumetric EOR based onT2/FLAIR disease, T1-enhancing volume, T2 volume includingT1, T1/T2 volumetric ratio for 80 patients from University of Texas MD Anderson Cancer Center [Ferguson et al., 2021]. The conclusion is that there are an astonishing number of potential factors that might influence the incidence of LTS in GBM.

Deep understanding of clinical and molecular characteristics that determine survival in Glioblastoma is the need of the hour. A complete approach to this problem has the potential to not

only improve prognostic classifications, but also to suggest therapeutic strategies that will help more patients become extraordinary long-term survivors.

0.4 Research Approach

The major goal of the present work is to find key determinants of short and long survival in primary Glioblastoma and to present a mechanistic picture of the signaling pathways that drive prognosis. Gene expression studies provide a snapshot of all transcriptional activity in a biological sample. They help to characterize the state and kind of that specific cell state. Resources like The Cancer Genome Atlas, Chinese Glioma Genome Atlas and some of the datasets of GEO and ArrayExpress not only provide valuable gene expression profiles but also the clinical characteristics of Glioblastoma samples under study.

Figure 3: Computational approaches utilized in the current study to better understand prognosis in Glioblastoma

For the work presented here, data from 20 Glioblastoma studies that are publicly accessible are used. Time-to-event analysis can be used to identify factors which have prognostic value on overall survival in Glioblastoma. At first, a meta-analytic approach is used to investigate prognostic value of clinical and molecular factors that are available for most of these datasets. This includes Age, Gender, Karnofsky Performance Score, IDH mutation status and MGMT promoter methylation status to see if any of these factors have significant con-

tribution to the overall survival using univariate cox regression analysis (mentioned in detail Chapter 1). Upon meta-analysis, age and MGMT status have a significant role in survival.

Time-to-event analysis on gene-expression profiles quantify the prognostic value of a gene in the given disease context. Several studies have demonstrated potential application of a gene panel-derived risk model built on time-to-analysis for predicting GBM prognosis. Zuo et al. [Zuo et al., 2019] developed a six-gene signature risk score model using RNA-seq data from TCGA and CGGA databases, however, it lacked an independent validation. Similarly, Cao et al. [Cao et al., 2019] proposed a 4-gene signature-derived risk score model that can predict prognosis and treatment response in GBM. Yin et al. [Yin et al., 2019] identified a 5-gene signature for prognosis prediction in GBM using TCGA RNA-seq cohort and a dataset from GEO database (GSE7696) [Prasad et al., 2020]. In chapter 2, we have used Time-to-event analysis to identify all the genes of prognostic value. Differentially expressed genes (DEGs) are identified by comparing gene activity spectra of the cellular system of interest and a control cell" [Koschmann et al., 2015]. This approach gives more refined insights on the observed gene expression activity. Most standard transcriptome analysis involve mapping of these differentially expressed genes to Gene ontology categories to decipher Molecular Function, Cellular Composition and biological process. Regulatory and metabolic pathways enriched can also be studied by enrichment analysis on pathway databases like KEGG or Reactome. These conventional approaches which give insights about downstream influences of gene activity are called "downstream analysis". However, they provide a very little clue about the cause of such a dysregulated gene expression. Differential gene expression analysis is exploited at multiple steps in this research work.

A novel approach which explains cause for such a gene dysregulation (Differential gene expression) in a given disease state is built on integrated promoter and pathway analysis. This strategy explains upstream steps of the gene regulation and hence termed as "upstream analysis". Regulation of transcription in eukaryotes is a result of the combined effects of structural properties (how DNA is "packaged") and the interactions of proteins called transcription factors [Cooper, 2000]. Transcription factors are regulatory proteins which upon activation, bind to specific DNA segments like promoters and enhancers and activate (or rarely inhibit) transcription. In turn, activation of transcription factors happens via series of chemical events in the cell upon binding of growth factors to the cell surface receptor. The upstream analysis comprises of both these steps (1) analysis of promoters and enhancers of identified DEGs to identify transcription factors (TFs) involved in the process under study; (2) reconstruction of signaling pathways that activate these TFs and identification of master-regulators on the top of such pathways [Koschmann et al., 2015, Boyarskikh et al., 2018, Stegmaier et al., 2011]. TGFβ and HIF1A were proposed to be upstream regulators [Tejero et al.,

2019], a causal relationship between innate immune cell infiltration and mesenchymal transdifferentiation in glioblastoma [Schmitt et al., 2021], predicted activation of integrin-linked kinase (ILK) signaling, actin cytoskeleton signaling, and lysine demethylase 5B (KDM5B) in Cancer-stem like cells migration [Verano-Braga et al., 2018] were some of the applications of this concept in Glioblastoma research. In this work, gene-regulatory networks are used to understand the causal mechanisms associated with gene dysregulation and to identify drivers of prognosis in Glioblastoma.

Machine Learning (ML) approaches are frequently used in glioblastoma research which is evident from an increased number of publications over the last decade [Valdebenito & Medina, 2019]. ML helps in identifying patterns, predicting outcomes or comprehending the relationships of complicated biochemical networks using large amounts of high dimensional data [Valdebenito & Medina, 2019]. Several studies have used machine learning approaches for classification of GBM using various features (Table 4). A novel stemness-based classifier built on data of 906 glioblastoma patients is suggested to have appealing implications in discriminating the prognosis, immunotherapy and temozolomide responses [Wang et al., 2021]. iGlioSub –a ML application built on integrated gene-expression and methylation profiles of 304 glioblastoma patients shows promising performance (AUC > 95%) in classifying samples into classical, mesenchymal and Proneural subtypes of glioblastoma [Ensenyat-Mendez et al., 2021]. An online survival predictor is developed based on demographic; socioeconomic, clinical, radiographic features obtained for 20,821 patients available from SEER registry [Senders et al., 2020]. Table 4 summarizes Machine learning approaches applied in Glioblastoma research.

Table 4: Machine learning applications in Glioblastoma research.

Publication	Method	No. of samples for training	Clinical endpoint (recurrence/ OS/TFS/)	Parameters (clinical/genomic/ methyl)	Performance
Zihao Wang et al.,2021	logistic regression	N = 376 (training)	Stemness Subtype(I/II) Predictor	Gene-expression	AUC = 0.9599, accuracy= 92.96%
Gregory P way et al., 2017	logistic regression	N= 321	NF1 wild-type and NF1 inactivated	Gene-expression	AUC= 0.77
Miquel Ensenyat-Mendez et al., 2021	Random Forest	N=234	GBM subtype-specific classifiers	Gene-expression	AUC (classical) =90.5 ± 2.1
	Random Forest	N=126		DNA methylation	AUC (classical) = 94.2 ± 1.4%
	nearest shrunken centroid (NSC)	N=234(gene-expression), N=126(DNA methylation) in training		Gene-expression & DNA methylation	AUC (classical) =97.5 ± 1.0%,
Yu-Hang Zhang1 et al., 2020	SVM	N=343 training	GBM subtype-specific classifiers	DNA methylation	Accuracy = 85.2%
	RF				Accuracy = 87.7%
	RIPPER				Accuracy = 95.4%
Tine Geldof et al., 2020	classification tree	N=2472	TMZ treatment response	Clinico-pathological	AUC = 67%
Jokey T senders et al., 2020	Random Forest	N = 20,821 (total)	Overall Survival	demographic, socioeconomic, clinical, and radiographic features	C-index = 0.69
	Bagged Decision trees				C-index =0.67
	SVM				C-index =0.7

0.5 Structure of the thesis

The thesis includes 4 chapters. Each chapter is a manuscript.

Chapter 1: Predictors of survival outcome in Glioblastoma: A meta-analysis of individual patient data (Status: Manuscript unpublished)

Chapter 2: Master regulators associated with poor prognosis in Glioblastoma (Status: Published work)

Chapter 3: IGFBP2 is a potential master-regulator driving the dysregulated gene network responsible for short survival in Glioblastoma (Status: Published work)

Chapter 4: Machine Learning-based Survival Group Prediction in Glioblastoma (Status: Published work

Each chapter includes "Availability of software, data and materials"and "Declaration of my contributions". Corresponding manuscript/publication is given after this section.

One never notices what has been done; one can only see what remains to be done.

Marie Curie

1

Predictors of survival outcome in Glioblastoma: A meta-analysis of individual patient data

Manuscript unpublished

Availability of software, data and materials

The dataset analyzed in the current study, supplementary files and plots are available in the GitHub project here - Predictors of survival outcome in Glioblastoma: A meta-analysis of individual patient data [*]

Declaration of my contributions

I conceptualized and executed this work and Dr. Alexander Kel participated in finalising pipeline for data analysis. To the best of our knowledge, all the publicly available datasets of Glioblastoma multiforme which had both gene expression and clinical information were

[*] https://github.com/genexplain/Manasa_KP_et_al_GBM_Survival_Predictors

collected and datasets were analysed. Prof. Edgar Wingender has participated extensively in improving the manuscript and Prof. Tim Beißbarth has supervised the work, read the manuscript, and approved it for correction.

Predictors of survival outcome in Glioblastoma: a meta-analysis of individual patient data

Manasa Kalya[1,2], Alexander Kel[2,3], Edgar Wingender[2], Tim Beißbarth[1]

1. Department of Medical Bioinformatics, University Medical Center Göttingen, 37099 Göttingen, Germany
2. geneXplain GmbH, 38302 Wolfenbüttel, Germany
3. Institute of Chemical Biology and Fundamental Medicine SBRAS, 630090, Novosibirsk, Russia

· Correspondence: Alexander E. Kel: alexander.kel@genexplain.com

Keywords: Survival Analysis, IDH, MGMT, Karnofsky Performance Score, Meta-analysis

Abstract

Glioblastoma is the most aggressive brain tumor with a poor median survival of ~15months. Several factors have been implicated in their role in treatment response and survival. Using a rigorous meta-analytic method, we investigated the determinants of survival in 14 separate GBM investigations. The research focuses on cohorts that include not only clinical data but also transcriptomics data from the related patient tumors. We only looked at primary GBMs, which are malignant tumors that started in the brain and didn't spread elsewhere. The biology of malignant brain tumors in children and adolescents (under the age of four years) might differ dramatically from that of adult-onset GBM patients. Overall, the current study looked at characteristics that influence prognosis in individuals with primary adult glioblastoma and each of these factors are discussed.

1 Introduction

Survival Analysis also called "time-to-event analysis" is an important approach in oncology research. It involves fundamental aspects of clinical management and drives decision-making around treatment strategies. The goal is to estimate the time for an individual or a group of individuals to experience an event of interest (e.g. time to disease remission, progression, or death). Patients with Glioblastoma have a poor prognosis with a poor median overall survival (OS) period of ~15months. Individual heterogeneity in the survival rates is undoubtedly observed and several prognostic factors have been found in recent years. To better understand the factors affecting the prognosis of glioma, we performed this retrospective study on patient cohorts collected from public databases.

Several clinical factors such as Age, Gender, Extent of tumor resection, Karnofsky performance status (KPS), chemotherapy, a dose of radiation, tumor location in the brain affect response to therapy and thereby survival.

1. Age at diagnosis– Age of the patient at the time he was diagnosed with the disease. Age is an important risk factor and is one of the strongest single prognostic factors for the outcome

(Paszat et al., 2001; Bauchet et al., 2010; Fabbro-Peray et al., 2019; Ius et al., 2020). Studies show median survival of patients older than 70 years is significantly lower than the younger patients (Ladomersky et al., 2020; Straube et al., 2020).

2. Karnofsky Performance Score (KPS) - In GBM, the Karnofsky Performance Scale (KPS) score is widely used to stratify a patient's prognosis and identify suitable therapy (GBM)(Lamborn et al., 2004; Gorlia et al., 2008). It is an established method of assessing cancer patients' ability to conduct everyday chores. Scores vary from 0 to 100 on the Karnofsky Performance Status scale. Higher scores imply that the patient is better at carrying his daily activities. Low preoperative KPS is known to be associated with shorter survival. Some studies use postoperative KPS scores for prediction as surgical resection can have a dramatic effect on a patient's functional status. Recent studies have reported postoperative KPS scores to have superior predictive capabilities for survival than preoperative KPS (Straube et al., 2020).

2. MGMT status- Cells deficient in O(6)-methylguanine-DNA methyltransferase (MGMT) were found to show increased sensitivity to temozolomide (TMZ) (Jovanović et al., 2019). MGMT gene methylation was an independently significant prognostic factor for both OS and Progression-Free Survival (Gorlia et al., 2008). In contrast to what has been documented in major reviews and studies with larger series of patients, MGMT methylation was not identified to be a prognostic factor or predictive of the TMZ response. They reported that "no association was detected between methylation of MGMT promoter and molecular markers such as ATRX, IDH, p53, and Ki67. These results indicate that MGMT methylation did not influence in patient survival in their cohort" (Egaña et al., 2020).

3. IDH status – The recurrent mutations in the isocitrate dehydrogenase (IDH1) gene is common in most Glioblastoma multiforme (GBM) cells and is associated with a better prognosis (Amelot et al., 2015) (Songtao et al., 2012).IDH1/2 mutations may result in genome-wide epigenetic changes in human gliomas. These mutations also reduce the capacity of cells to produce NADPH, which in turn increases the vulnerability to oxidative stress, making the tumor cells more susceptible to irradiation and chemotherapy.

4. Gender - A stratified analysis of GBM patient data obtained from the SEER database showed that male patients always had the lowest 5 year- cancer-specific survival rate across localized cancer stages and different age subgroups and is proposed to have a prognostic value (Tian et al., 2018).

In this work, we have explored 14 independent GBM studies to study the predictors of survival using a systematic meta-analytic approach. The work concentrates on the cohorts that not only have clinical information but also transcriptomics data of the corresponding patient tumors. We restricted our analysis to the primary GBM, meaning malignant tumors which primarily originated from the brain and nowhere else. The biology of malignant brain tumors in pediatric or younger(<4oyrs) can significantly differ from that of the adult-onset

one we included on adult GBM patients. Overall, the current study explored factors affecting prognosis in primary adult glioblastoma patients.

2 Results

2.1 Sample Characteristics

The median age of 1708 GBM patients from 14 different cohorts is 58years (Min = 40, max =90). Their median survival is 13.7 months. Of these patients, 244 patients are aged above 70years and 414 patients are between the age of 40-50yrs. ~56% of the patients are Male and 33.8% are Female. **Table 1a and Table 1b.** Age group was sub divided into Group A(40-50yrs), B(51-60), C(61-70) and D(71-90 yrs).

Table1: a) Distribution Age, Karnofsky score and Overall survival for data integrated from 14 individual studies b) Number of samples in each category across 14 individual studies

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max	NA's
Age at diagnosis	**40**	**50.57**	**58**	**58.6**	**66**	**89**	**41**
Karnofsky score	**0**	**70**	**80**	**77.36**	**90**	**100**	**1044**
Overall Survival	**0**	**7.02**	**13.7**	**19.94**	**23.14**	**299.42**	**61**

Gender	**Male: 973**
	Female: 579
	NA's: 156
MGMT Promoter Methylation status	**Methylated: 365**
	Unmethylated: 388
	NA's: 955
IDH mutation Status	**Wildtype: 814**
	Mutant: 69

	NA's: 825
Age Group	**Group_A: 414**
	Group_B: 552
	Group_C: 457
	Group_D: 244
	NA's : 41

2.2 Univariate Analysis

Clinical factors such as Age, Gender, Karnofsky performance, MGMT status, IDH status were studied for their impact on overall survival in all 14 datasets individually using univariate survival analysis. **Figure 1** depicts KM plots describing the overall survival according to the single clinical factor under investigation. The log-rank test helps compare the groups within each clinical variable (e.g., Male and Female) for their impact on overall survival.

The log-rank test for difference in overall survival gives a p-value of p = 2e-11, indicating that the age groups differ significantly in survival in the TCGA_GBM cohort.

Univariate cox regression analysis is performed to evaluate simultaneously the effect of several factors on survival. The coefficient beta, Hazard ratio, corresponding Standard Error, and value for every clinical factor in every dataset are obtained from univariate Cox regression Table 2. According to these observations, Gender was not found to have a significant ($p < 0.05$) impact on overall survival in any of the individual cohorts, MGMT methylation status was found significant in its impact in at least 4 individual datasets.

The estimates of univariate Cox regression – Hazard ratio, Standard Error, and value are considered.

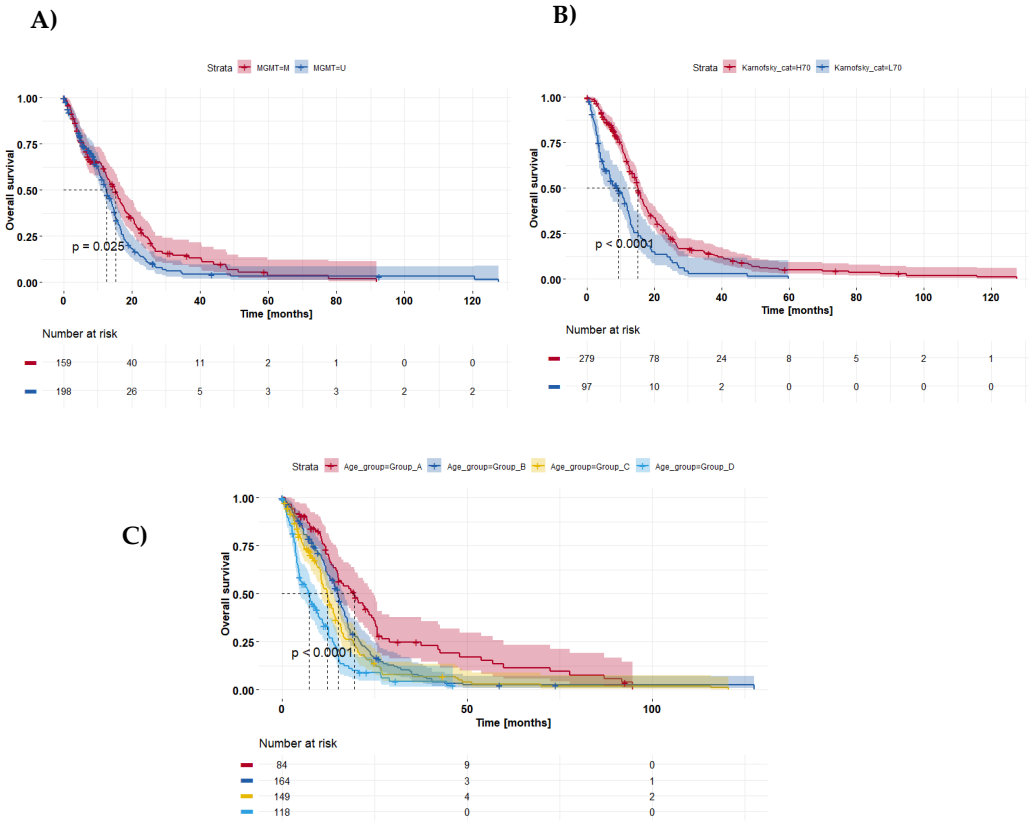

Figure1. KM plots depicting the impact of clinical variables **A)** MGMT methylation status (M-Methylated/U-Unmethylated), **B)** Karnofsky category (>70 H70, <70 L70), **C)** Age groups [A:(40-50), B:(51-60), C:(61-70), D:(71-90)] years of age on overall survival according to TCGA_GBM microarray data of 560 patients.

Table 2. Univariate cox regression analysis of the clinical variables understudy in all 14 datasets under study.

Dataset_Name	Clinical_Variable	beta	p-value	SE	HR
Zhao Z et al.,2017	Gender_Cat	0.10	0.54	0.16	1.10
	age_at_diagnosis	0.00	0.88	0.01	1.00

	MGMT	**-0.05**	**0.77**	**0.18**	**0.95**
	IDH_status	**-0.78**	**0.00**	**0.27**	**0.46**
	AgeGroup_B	**-0.34**	**0.08**	**0.20**	**0.71**
	AgeGroup_C	**-0.21**	**0.32**	**0.21**	**0.81**
	AgeGroup_D	**0.46**	**0.21**	**0.36**	**1.58**
TCGA_GBM	**Gender_Cat**	**-0.18**	**0.07**	**0.10**	**0.83**
	age_at_diagnosis	**0.04**	**0.00**	**0.00**	**1.04**
	MGMT	**0.27**	**0.03**	**0.12**	**1.31**
	IDH_status	**-1.01**	**0.01**	**0.38**	**0.37**
	Karnofsky	**-0.02**	**0.00**	**0.00**	**0.98**
	AgeGroup_B	**0.42**	**0.01**	**0.15**	**1.52**
	AgeGroup_C	**0.60**	**0.00**	**0.15**	**1.82**
	AgeGroup_D	**1.11**	**0.00**	**0.16**	**3.03**
GlioTrain	**Gender_Cat**	**-0.29**	**0.16**	**0.21**	**0.75**
	age_at_diagnosis	**0.01**	**0.35**	**0.01**	**1.01**
	MGMT	**1.10**	**0.00**	**0.22**	**3.00**
	AgeGroup_B	**-0.06**	**0.82**	**0.26**	**0.94**
	AgeGroup_C	**0.17**	**0.48**	**0.24**	**1.19**
	AgeGroup_D	**-0.09**	**0.93**	**1.02**	**0.91**
Gusev Y et al.,2018	**Gender_Cat**	**-0.12**	**0.51**	**0.18**	**0.89**
	age_at_diagnosis	**0.00**	**0.88**	**0.01**	**1.00**
	Karnofsky	**0.00**	**0.56**	**0.00**	**1.00**

	AgeGroup_B	0.11	0.59	0.20	1.12
	AgeGroup_C	0.16	0.48	0.23	1.18
	AgeGroup_D	-0.07	0.77	0.24	0.93
Murat A et al.,2008	Gender_Cat	0.00	0.99	0.29	1.00
	age_at_diagnosis	0.05	0.02	0.02	1.05
	MGMT	1.50	0.00	0.31	4.48
	AgeGroup_B	0.03	0.91	0.31	1.04
	AgeGroup_C	0.42	0.22	0.35	1.53
	AgeGroup_D	1.02	0.33	1.04	2.76
Kawaguchi A, et al.,2013	Gender_Cat	0.33	0.43	0.42	1.39
	age_at_diagnosis	0.05	0.02	0.02	1.05
	Karnofsky	-0.04	0.01	0.02	0.96
	AgeGroup_B	2.13	0.05	1.10	8.44
	AgeGroup_C	2.45	0.02	1.07	11.53
	AgeGroup_D	2.48	0.03	1.15	11.95
Gravendeel LA et al.,2009	Gender_Cat	-0.33	0.10	0.20	0.72
	age_at_diagnosis	0.05	0.00	0.01	1.05
	IDH_status	-0.44	0.12	0.29	0.64
	AgeGroup_B	0.25	0.35	0.27	1.28
	AgeGroup_C	0.71	0.01	0.27	2.03
	AgeGroup_D	1.53	0.00	0.34	4.64
Lee Y*et al.,2008	Gender_Cat	0.07	0.68	0.17	1.07

	age_at_diagnosis	0.03	0.00	0.01	1.03
	AgeGroup_B	0.50	0.02	0.21	1.65
	AgeGroup_C	0.72	0.00	0.25	2.04
	AgeGroup_D	0.83	0.00	0.27	2.28
Sturm D et al.,2012	Gender_Cat	0.38	0.40	0.45	1.46
	age_at_diagnosis	0.02	0.66	0.06	1.02
	AgeGroup_B	0.68	0.12	0.43	1.96
	Gender_Cat	0.38	0.40	0.45	1.46
Vital AL et al.,2010	age_at_diagnosis	0.02	0.66	0.06	1.02
	AgeGroup_B	0.32	0.79	1.19	1.38
	AgeGroup_C	-0.19	0.79	0.71	0.83
	AgeGroup_D	0.31	0.70	0.80	1.36
	age_at_diagnosis	0.00	0.88	0.03	1.00
Puchalski R.B. et al. ,2018	MGMT	1.35	0.02	0.57	3.84
	AgeGroup_B	-0.55	0.40	0.65	0.57
	AgeGroup_C	-0.28	0.64	0.60	0.76
	AgeGroup_D	-0.08	0.91	0.72	0.92
	age_at_diagnosis	0.02	0.25	0.02	1.02
Joo KM et al.,2013	Gender_Cat	-0.31	0.37	0.34	0.74
	AgeGroup_B	-0.27	0.52	0.42	0.76
	AgeGroup_C	0.13	0.77	0.43	1.13
	AgeGroup_D	1.03	0.09	0.60	2.81

	age_at_diagnosis	**0.04**	**0.02**	**0.02**	**1.04**
Ducray F. et al.,2010	**MGMT**	**-0.34**	**0.62**	**0.68**	**0.71**
	AgeGroup_B	**0.36**	**0.46**	**0.48**	**1.43**
	AgeGroup_C	**0.56**	**0.26**	**0.50**	**1.75**
	AgeGroup_D	**1.69**	**0.01**	**0.63**	**5.40**
	age_at_diagnosis	**0.03**	**0.06**	**0.01**	**1.03**
Freije WA. et al.,2004	**Gender_Cat**	**0.57**	**0.11**	**0.36**	**1.77**
	AgeGroup_B	**0.95**	**0.03**	**0.45**	**2.59**
	AgeGroup_C	**0.37**	**0.46**	**0.50**	**1.45**
	AgeGroup_D	**0.93**	**0.09**	**0.55**	**2.54**

2.3. Identifying sources of heterogeneity in the data:

Baujat plot helps to detect sources of heterogeneity in the meta-analytic data. The plot shows the contribution of each study to the overall Q-test statistic for heterogeneity on the horizontal axis versus the influence of each study (defined as the standardized squared difference between the overall estimate based on an equal-effects model with and without the ith study included in the model) on the vertical axis.

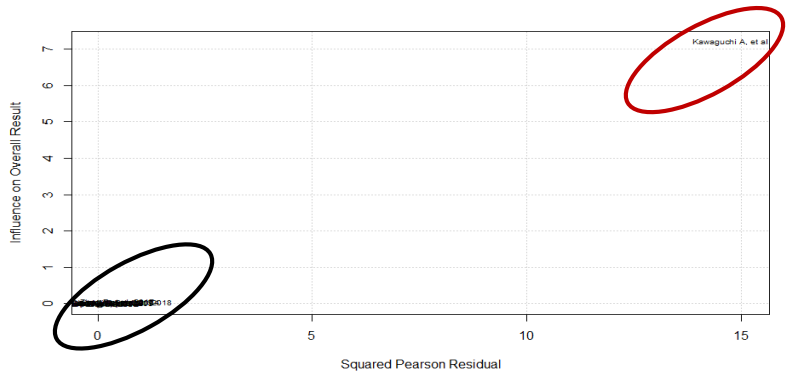

Figure 3. Baujat plot for meta-analytic data of Age groups(B/C/D). The plot shows Kawaguchi A, et al.,2013 as the dataset contributing to high overall heterogeneity and has a high influence on the overall result.

2.4 Publication Bias:

Publication bias was evaluated by Test for plot asymmetry and Kendall's Tau both were insignificant (p= 0.72 & p=0.59 respectively), indicating no publication bias. Figure 5A

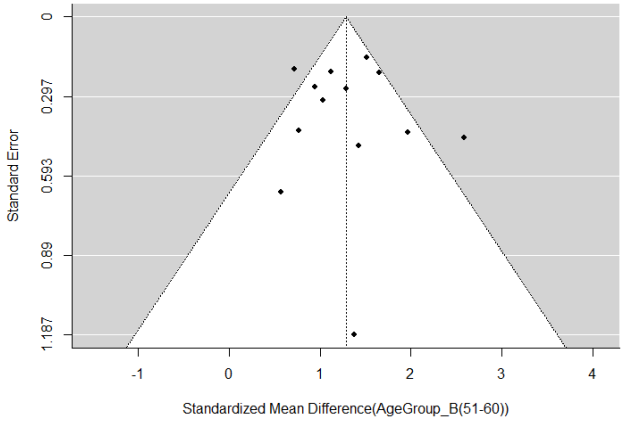

Figure 4: A Funnel plot for visualizing publication bias in meta-analytic data used in the study for Age group B (51-60yrs)

2.5. Impact of Age on Overall Survival in GBM

In 14 studies (References), survival was defined as the time from diagnosis till death or the end of follow-up, and it indicates OS. Due to heterogeneity (I2 >50%), we have used the Random Effect model for meta-analyses. We find that all clinical variables were significant in their impact on OS. We find that age group D (70-90yrs) were having high and significant HR of 2.47 with 95% CI of (1.646-3.294) and is higher than age group C [HR=1.43, 95% CI = 1.15-1.72] and group B [HR = 1.284, 95%CI =1.016-1.55]. In the Fixed effect model, I^2 was 23.58% and H^2 was 1.31. Gender and Karnofsky had ~1 HR indicating no higher influence upon Overall survival. Karnofsky score and IDH status information were available in fewer cohorts (K=4 & 3 respectively). The heterogeneity statistic I2 can be biased in such small meta-

analyses. For Gender and IDH status, the Fixed effect models were also implemented due to low heterogeneity (I2 >50%). In the Fixed effect model, I^2 was 23.58% and H^2 was 1.31. **Table 3.**

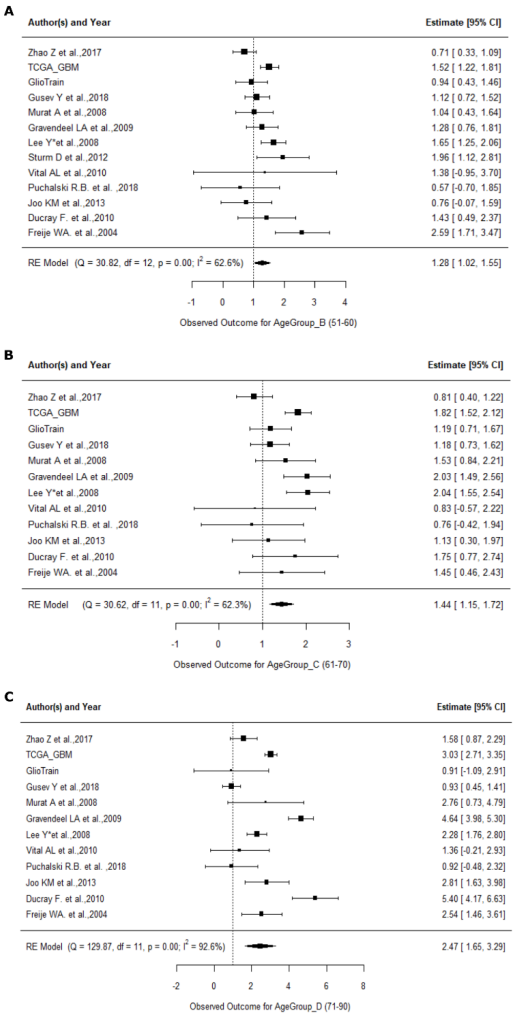

Figure 5: Forest plot for visualizing meta-analyses results for Age groups (B, C &D) or individual studies together with their 95% Confidence intervals. The Q and I2 statistic along with the p.value for the Random Effect model is given at the bottom left of the plot. The polygon added at the bottom of the plot depicts a summary estimate based on the model (with the center of the polygon corresponding to the estimate and the left/right edges indicating the confidence interval limits.

Table 3. Meta-analysis on survival outcomes in 14 GBM cohorts using RandomEffect model

Clinical Factors	No. of Data sets	Degrees of Freedom	tau^2	tau	I^2	H^2	Test for Heterogeneity	Model Results:					
								Estimate	SE	Zval	pval	95% CI	Significance
Age_at_diagnosis	k = 14	df=13	0.000 (SE = 0.000)	0.016	66.53%	2.99	Q(df = 13) = 41.429, p-val < .001	1.026	0.006	173.319	<.001	1.014-1.037	***
AgeGroup_B	k=14	df=13	0.1296 (SE = 0.0933)	0.36	62.63%	2.68	Q(df = 12) = 30.8170, p-val = 0.0021	1.2846	0.1371	9.3719	<.0001	1.016-1.5533	***
AgeGroup_C	k=12	df=11	0.1348 (SE = 0.1012)	0.3672	62.26%	2.65	Q(df = 11) = 30.6206, p-val = 0.0013	1.4378	0.1443	9.9656	<.0001	1.1550-1.7205	***
Age_Group_D	k=12	df=11	1.7783 (SE = 0.8979)	1.3335	92.65%	13.6	Q(df = 11) = 129.8737, p-val < .0001	2.4707	0.4205	5.8751	<.0001	1.6464-3.294	***
Gender	k=12	df=11	0.0054 (SE = 0.0184)	0.0733	10.60%	1.12	Q(df = 11) = 14.3938, p-val = 0.2120	0.9591	0.0652	14.7149	<.0001	0.8313-1.0868	***
MGMT	k=6	df=5	2.375 (SE = 1.597)	1.541	97.73%	44.02	Q(df = 5) = 158.825, p-val < .001	2.388	0.649	3.678	<.001	1.114-3.660	***
Karnofsky	k=3	df=2	0.000 (SE = 0.000)	0.014	84.55%	6.47	Q(df = 2) = 11.868, p-val = 0.003	0.982	0.009	107.239	<.001	0.964-1.00	***
IDH_Status	k=4	df=4	0 (SE = 0.091)	0	0.00%	1	Q(df = 3) = 0.398, p-val = 0.941	0.508	0.173	2.926	0.003	0.168-0.848	**

3. Discussion

Despite improvements in treatment strategies over the last decade, the median survival in Glioblastoma has not significantly improved. Nevertheless, there is a small group of GBM patients who respond to standard care and survive beyond 36 months, clinically termed as Long-term survivors. It is been a long-standing effort to understand these extreme survivors and what factors favor their prognosis both at clinical and at the gene level information.

In this chapter, we have conducted a retrospective study collecting clinical data from 2309 patients with Glioblastoma obtained from 14 independent studies. The criteria for study selection is that they have clinical as well as gene-expression data available. To achieve more conclusive observations specific to primary glioblastoma, we have restricted the study to explore adult (>40yrs) primary glioblastoma. It is noteworthy that the biology of younger and pediatric glioblastoma is different than that of adult GBM. We have explored clinical predictors of overall survival in Glioblastoma based on a systematic meta-analytic approach.

We find a significant association of age and MGMT status with overall survival. Further stratification of age groups as A/B/C & D (40-50, 51-60,61-70 & 71-90) gave us a clearer picture of the impact of age on prognosis. Most authors agree that age is an important prognostic factor, however, there is no standard age-group classification/information used to evaluate risk in GBM patients. Here we report that patients with age group of (51-60yrs) experience 20% more risk & age group C(61-70yrs) experience 40% more risk of death respectively compared to younger age group (40-50yrs), group A. It is important to mention that the older age group – group D (71-90yrs) has 2.4times higher risk of death compared to group A. Thus, giving an amplitude of risk experienced at each age category.

Methylation of the O (6)-Methylguanine-DNA methyltransferase (MGMT) promoter is predictive for treatment response in glioblastoma patients and is directly associated with temozolomide drug sensitivity. Methylation of the MGMT promoter in GBM patients is associated with significantly higher survival rates if treated with radiotherapy and TMZ. With a meta-analytic approach on available data from 6 cohorts, we report that MGMT unmethylated patients exhibit a significant 2.3-fold increased risk of death.

Earlier studies have reported an increased risk of death in Males than females. In the current study, we find that the risk of death is ~1 for both the gender categories. The information on Karnofsky and IDH status were available for fewer cohorts making it difficult to conclude on our observations of risk of hazard, As a result of the data's unavailability, the chapter has not explored the impact of other important clinical variables like tumor size, tumor location, tumor resection, and recurrence aspects which are strongly associated with the existing clinical factors like Karnofsky score. Multivariate analysis of individual survival predictors will help to understand the association of such multiple factors on survival factors which are

not discussed in the current chapter. Sensitivity analysis is also not discussed in the current chapter as there were few prominent and influential studies removing them

The knowledge gained in this chapter is later used in chapter 4, where we explore the machine learning approach integrating both clinical and transcriptomic information of tumors from GBM patients.

4. Methods

The analysis is restricted to all the GBM studies used in this research work. It mainly deals with the studies containing information on overall survival and availability of gene expression data.

4.1 Patients

This study involves Primary GBM cohorts publicly available like REMBRANDT, TCGA, CGGA, and from the GEO database. The clinical information – Overall Survival(OS), age, gender, Karnofsky_Score, MGMT_status (methylated or unmethylated), and IDH_mutation status (wild type/mutated) are extracted from the sample details provided in the respective studies. Karnofsky score is later classified into 2 categories as 'Karnofsky_Category' variable (>70 or <70). In Gravendeel La et Al.,2009 (REMBRANDT), the age of the patient is given in a range. The median of the age range is considered as the age of the patient for this analysis. The characteristics of all included studies are summarized in Table 1. This encompasses ~14 studies containing information on 2309 patients with a diagnosis of Glioblastoma. To make the study more homogeneous, pediatric samples are excluded, reducing the sample size to 1708 adult(>40yrs) primary GBM tumors. 'Survival in months' is taken as a time variable and the survival status (Death/Alive) of the patients at the end of each study is considered as an event.

4.2 Ethics statement

The analysis did not involve interaction with human subjects or the use of personal identifying information. All the data is in the public domain.

Table 4. Studies and sample characteristics considered for the meta-analytic approach

Study	Age	Gender	Karnofsky	MGMT status	IDH status
TCGA,2008 (N=528)	(N=519)	(N=517)	-	(N=347)	(N=402)
	Range	M =314		U= 177	WT= 372
	(10.9-89.3)	F= 203		M=170	M= 30

	Median=59.4				
Zhao Z et Al.,2017		**(N=388)**			**(N=228)**
(N=388)	**Range (8-79)**	**M =235**	**-**	**-**	**WT= 175**
	Median=49	**F= 153**			**M= 53**
Gusev Y et Al.,2018		**(N=165)**	**(N=108)**		
(N=220)	**Range (42-87)**	**M =106**	**Range (20-100)**	**-**	**-**
	Median=57	**F= 59**	**≤40 =9**		
			41-69=22		
			≥70 = 77		
Gliotrain, 2020		**(N=133)**	**(N=133)**	**(N=130)**	**(N=133)**
(N=133)	**Range (24.17 - 70.35)**	**M =89**	**Range (70-100)**	**U= 57**	**WT= 133**
(unpublished)	**Median= 57.92**	**F= 44**	**<70 = 0**	**M=73**	**M= 0**
			≥70 = 133		
Murat A et Al.,2008		**(N=80)**		**(N=78)**	
(N=80)	**Range (41.7-70.3)**	**M =59**	**-**	**U= 34**	**-**
	Median=52.25	**F= 21**		**M=44**	
Kawaguchi A, et Al.,2013		**(N=50)**	**(N=29)**		
(N=32)	**Range (18-80)**	**M =34**	**Range (40-100)**	**-**	**-**
	Median=59	**F= 16**	**<70 = 9**		
			≥70 = 20		
Gravendeel La et Al.,2009		**(N=159)**			**(N=128)**
(N=159)	**Range (40.58-80.65)**	**M =108**	**-**	**-**	**WT=95**
	Median=55.39	**F= 51**			**M= 33**
Lee Y et Al.,2008		**(N=192)**	**-**	**-**	**-**

(N=191)	Range (41-86)	M =117			
	Median=54	F= 74			
Sturm D et Al.,2012		(N=109)	-	-	(N=)
(N=136)	Range (41-75)	M =57			WT= 114
	Median=31.5	F= 52			M= 22
Vital Al et Al.,2010		(N=26)	(N=26)	-	-
(N=26)	Range (30-84)	M =13	Range (50-90)		
	Median=67	F= 13	<70 = 11		
			≥70 = 15		
IVYGAP 2018		-	-	(N=36)	-
(N=37)	Range (17-70)			U= 23	
	median =60			M=13	
Joo et al., 2013		(N=57)	-	-	-
(N=57)	Range (40-76)	M =31			
	Median=51	F= 26			
Ducray et al.,2010		-	-	(N=16)	-
(N=48)	Range (78.7)			U=12	
	Median=58.2			M=4	
Freije WA. et al.,2004	(N=43)	(N=43)	-	-	-
(N=43)	Range (40-82)	M=19			
	Median=54	F=24			

4.3 **Statistical analysis**

Univariate Analysis

The Wilcoxon test was used to compare patient characteristics for overall survival in GBM for variables that were either continuous or ordered categorical (e.g., Karnofsky_Category). Age was used as a continuous variable in the analysis.

To see if there was statistical evidence of differences between the groups' survival curves, univariate cox proportional hazards regression model was used. To calculate the groups' hazard ratios, exact conditional maximum likelihood estimates were employed, and Fisher 95% percent confidence intervals were created for significance testing between the groups' hazard ratios. The results were statistically significant at P 0.05. The hazard is modelled with the equation:

$$h(t) = h(0)t + \exp(b1x1 + b2x2 + \cdots .. bkxk)$$

Where, $h_0(t)$ represents underlying hazard

$b_1, b_2 .. b_k$ are parameters to be estimated

$x_1, x_2 \ldots x_k$ are risk factors(covariates)

The hazard is the chance that at any given moment, the event will occur, given that it hasn't already done so. The hazard ratio (HR) is a measure of the relative hazard in two groups i.e. ratio of the hazard for one group compared to another.

$$Hazard\ Ratio = \frac{Hazard\ of\ group\ A}{Hazard\ of\ group\ B}$$

0< HR <1: group A are at a decreased hazard compared to group B.

HR =1: The hazard is the same for both groups

HR >1: group A are at increased hazard compared to group B

a HR of 0.5 means a halving of hazard and a HR of 2 means doubling of hazard

Kaplan-Meier curves are the graphical representation of the survival function estimated from the data under study. Non-parametric Log-rank test is used here to compare two groups. The log rank test compares the total number of events observed with the number of events we would expect assuming that there is no group effect. KM plots starts at 1 at time 0, where all patients are alive and event free. It is a step function the curve steps down each time an event occurs, and so tails off to 0. Poor survival is reflected by a curve that drops relatively rapidly.

All statistical analyses utilized R software (Survival, Survminer and Metafor packages).

Meta-Analysis:

Meta-analyses are used to determine the strength of the evidence for a condition or treatment. One goal is to see if there is an effect; another is to see if the effect is positive or negative, and, ideally, to get a single summary assessment of the effect.

A meta-analysis' findings can increase the precision of effect estimates, answer problems not addressed by individual research, resolve disagreements emerging from seemingly contradictory studies, and develop new hypotheses.

The investigation of heterogeneity, in particular, is critical for the formulation of new theories.

The direct method of meta-analysis is to use the estimates of $lnHR$ and its variance obtained from univariate cox regression model. Pooling is done using metafor package in R.

An estimate of the log hazard ratio and variance pooled across studies can be calculated:

$$\ln(HR) = \frac{\sum_{i=1}^{k} \frac{\ln(\text{HRi})}{Var[\ln(HRi)]}}{\sum_{i=1}^{k} \frac{1}{Var[\ln(HRi)]}}$$

$$Var[\ln(HR)] = \left[\sum_{i=1}^{k} \frac{1}{Var[\ln(HRi)]}\right]^{-1}$$

$$Standard\ Error = \sqrt{Variance}$$

Standard error for each study

We have tested Fixed-Effect and Random effect statistical models to aggregate all the data. Fixed Effect models assume that the explanatory variable has a fixed or constant relationship with the response variable across all observations. All the observations in the model have pre-determined categories and the inferences (patients' response). Heterogeneity measures (Q and I2) will give insights on what model should be considered.

The null hypothesis is to test that all treatment effects are zero. The alternate is to say that the effects are heterogenous.

When the overall null hypothesis is rejected, the next step is to test whether all effects are equal, that is, whether the effects are homogeneous. Alternative is that at least one effect is different, that is, that the effects are heterogeneous

chi-square test was designed to test the null hypothesis that all treatment effects are equal. This hypothesis is tested using Cochran's Q test which is given by

$$Q = \sum_{I=1}^{k} w_i \, (\theta_i - \theta)^2$$

The test is conducted by comparing Q to a x^2_{k-1} distribution.

In the result table of meta-analysis;

- Cochran's Q: This is the computed chi-square value for Cochran's Q statistic.
- DF: For this test, the degrees of freedom is equal to the number of studies minus one
- Pvalue: This is the significance level of the test. If this value is less than the specified value of alpha (usually 0.05), the test is statistically significant and the alternative is concluded. If the value is larger than the specified value of alpha, no conclusion can be drawn other than that you do not have enough evidence to reject the null hypothesis.

Baujat plot a diagnostic plot to detect sources of heterogeneity in meta-analytic data. The plot shows the contribution of each study to the overall Q-test statistic for heterogeneity on the horizontal axis versus the influence of each study (defined as the standardized squared difference between the overall estimate based on an equal-effects model with and without the i^{th} study included in the model) on the vertical axis.

The I^2 indicates the level of heterogeneity. It can take values from 0% to 100%. If I^2 ≤ 50%, studies are considered homogeneous, and a fixed effect model of meta-analysis can be used. If I^2 > 50%, the heterogeneity is high, and one should use **random effect model** for meta-analysis.

Forest plots provide a graphical display of the observed effect, confidence interval, and usually also the weight of each study. They also display the pooled effect we have calculated in a meta-analysis.

4.4 Publication Bias

A potential stumbling block is relying solely on the corpus of published research, which can lead to inflated results due to publication bias, as studies with poor or negligible results are less likely to be published.

Publication bias (the association of publication probability with the statistical significance of study results) may lead to asymmetrical funnel plots and is evaluated by regression and correlation tests for symmetry in funnel plot.

Availability of software, data and materials: All the datasets analyzed in the current study are available from previous publications. The data and the results of the analysis performed using the Genome Enhancer in geneXplain platform are available here.

https://github.com/genexplain/Manasa_KP_et_al_GBM_Survival_Predictors

Conflicts of Interest: The authors Manasa Kalya Purushothama, and Tim Beißbarth are from Department of Medical Bioinformatics, University Medical Center Göttingen, Alexander Kel and Edgar Wingender are employees of geneXplain GmbH.

Author Contributions:

AK is involved in providing resources, supervision, and manuscript reviewing. MKP has conceptualized the work, performed data collection, data analysis, interpreted results, and written the manuscript. EW and TB were involved in the supervision of work and in reviewing the draft.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No 766069.

Acknowledgements: It is my pleasure to acknowledge Ravi Kumar Nadella who has read and given comments on every version of this manuscript.

References

Amelot, A., De Cremoux, P., Quillien, V., Polivka, M., Adle-Biassette, H., Lehmann-Che, J., et al. (2015). IDH-Mutation Is a Weak Predictor of Long-Term Survival in Glioblastoma Patients. *PLoS One* 10, e0130596. doi:10.1371/JOURNAL.PONE.0130596.

Bauchet, L., Mathieu-Daudé, H., Fabbro-Peray, P., Rigau, V., Fabbro, M., Chinot, O., et al. (2010). Oncological patterns of care and outcome for 952 patients with newly diagnosed glioblastoma in 2004. *Neuro. Oncol.* 12, 725–735. doi:10.1093/NEUONC/NOQ030.

Ducray, F., de Reyniès, A., Chinot, O., Idbaih, A., Figarella-Branger, D., Colin, C., et al. (2010). An ANOCEF genomic and transcriptomic microarray study of the response to radiotherapy or to alkylating first-line chemotherapy in glioblastoma patients. *Mol. Cancer* 9. doi:10.1186/1476-4598-9-234.

Egaña, L., Auzmendi-Iriarte, J., Andermatten, J., Villanua, J., Ruiz, I., Elua-Pinin, A., et al. (2020). Methylation of MGMT promoter does not predict response to temozolomide in patients with glioblastoma in Donostia Hospital. *Sci. Reports 2020 101* 10, 1–11. doi:10.1038/s41598-020-75477-9.

Fabbro-Peray, P., Zouaoui, S., Darlix, A., Fabbro, M., Pallud, J., Rigau, V., et al. (2019). Association of patterns of care, prognostic factors, and use of radiotherapy–temozolomide therapy with survival in patients with newly diagnosed glioblastoma: a

French national population-based study. *J. Neurooncol.* 142, 91–101. doi:10.1007/S11060-018-03065-Z/TABLES/4.

Freije, W. A., Castro-Vargas, F. E., Fang, Z., Horvath, S., Cloughesy, T., Liau, L. M., et al. (2004). Gene expression profiling of gliomas strongly predicts survival. *Cancer Res.* 64, 6503–6510. doi:10.1158/0008-5472.CAN-04-0452.

Gorlia, T., van den Bent, M. J., Hegi, M. E., Mirimanoff, R. O., Weller, M., Cairncross, J. G., et al. (2008). Nomograms for predicting survival of patients with newly diagnosed glioblastoma: prognostic factor analysis of EORTC and NCIC trial 26981-22981/CE.3. *Lancet. Oncol.* 9, 29–38. doi:10.1016/S1470-2045(07)70384-4.

Gravendeel, L. A. M., Kouwenhoven, M. C. M., Gevaert, O., De Rooi, J. J., Stubbs, A. P., Duijm, J. E., et al. (2009). Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res.* 69, 9065–9072. doi:10.1158/0008-5472.CAN-09-2307.

Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J. C., Fine, H., and Madhavan, S. (2018). Data descriptor: The REMBRANDT study, a large collection of genomic data from brain cancer patients. *Sci. Data* 5. doi:10.1038/sdata.2018.158.

Ius, T., Somma, T., Altieri, R., Angileri, F. F., Barbagallo, G. M., Cappabianca, P., et al. (2020). Is age an additional factor in the treatment of elderly patients with glioblastoma? A new stratification model: an Italian Multicenter Study. *Neurosurg. Focus* 49, E13. doi:10.3171/2020.7.FOCUS20420.

Joo, K. M., Kim, J., Jin, J., Kim, M., Seol, H. J., Muradov, J., et al. (2013). Patient-specific orthotopic glioblastoma xenograft models recapitulate the histopathology and biology of human glioblastomas in situ. *Cell Rep.* 3, 260–273. doi:10.1016/J.CELREP.2012.12.013.

Jovanović, N., Mitrović, T., Cvetković, V. J., Tošić, S., Vitorović, J., Stamenković, S., et al. (2019). The Impact of MGMT Promoter Methylation and Temozolomide Treatment in Serbian Patients with Primary Glioblastoma. *Medicina (Kaunas).* 55.

doi:10.3390/MEDICINA55020034.

Kawaguchi, A., Yajima, N., Tsuchiya, N., Homma, J., Sano, M., Natsumeda, M., et al. (2013). Gene expression signature-based prognostic risk score in patients with glioblastoma. *Cancer Sci.* 104, 1205–1210. doi:10.1111/CAS.12214.

Ladomersky, E., Zhai, L., Lauing, K. L., Bell, A., Xu, J., Kocherginsky, M., et al. (2020). Advanced Age Increases Immunosuppression in the Brain and Decreases Immunotherapeutic Efficacy in Subjects with Glioblastoma. *Clin. Cancer Res.* 26, 5232–5245. doi:10.1158/1078-0432.CCR-19-3874.

Lamborn, K. R., Chang, S. M., and Prados, M. D. (2004). Prognostic factors for survival of patients with glioblastoma: recursive partitioning analysis. *Neuro. Oncol.* 6, 227–235. doi:10.1215/S1152851703000620.

Lee, Y., Scheck, A. C., Cloughesy, T. F., Lai, A., Dong, J., Farooqi, H. K., et al. (2008). Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Med. Genomics* 1, 52. doi:10.1186/1755-8794-1-52.

McLendon, R., Friedman, A., Bigner, D., Van Meir, E. G., Brat, D. J., Mastrogianakis, G. M., et al. (2008). Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* 455, 1061–1068. doi:10.1038/nature07385.

Murat, A., Migliavacca, E., Gorlia, T., Lambiv, W. L., Shay, T., Hamou, M. F., et al. (2008). Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *J. Clin. Oncol.* 26, 3015–3024. doi:10.1200/JCO.2007.15.7164.

Paszat, L., Laperriere, N., Groome, P., Schulze, K., Mackillop, W., and Holowaty, E. (2001). A population-based study of glioblastoma multiforme. *Int. J. Radiat. Oncol. Biol. Phys.* 51, 100–107. doi:10.1016/S0360-3016(01)01572-3.

Puchalski, R. B., Shah, N., Miller, J., Dalley, R., Nomura, S. R., Yoon, J. G., et al. (2018). An

anatomic transcriptional atlas of human glioblastoma. *Science* 360, 660–663. doi:10.1126/SCIENCE.AAF2666.

Songtao, Q., Lei, Y., Si, G., Yanqing, D., Huixia, H., Xuelin, Z., et al. (2012). IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. *Cancer Sci.* 103, 269–273. doi:10.1111/J.1349-7006.2011.02134.X.

Straube, C., Kessel, K. A., Antoni, S., Gempt, J., Meyer, B., Schlegel, J., et al. (2020). A balanced score to predict survival of elderly patients newly diagnosed with glioblastoma. *Radiat. Oncol.* 15, 1–11. doi:10.1186/S13014-020-01549-9/FIGURES/4.

Sturm, D., Witt, H., Hovestadt, V., Khuong-Quang, D. A., Jones, D. T. W., Konermann, C., et al. (2012). Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. *Cancer Cell* 22, 425–437. doi:10.1016/J.CCR.2012.08.024.

Tian, M., Ma, W., Chen, Y., Yu, Y., Zhu, D., Shi, J., et al. (2018). Impact of gender on the survival of patients with glioblastoma. *Biosci. Rep.* 38. doi:10.1042/BSR20180752.

Vital, A. L., Tabernero, M. D., Castrillo, A., Rebelo, O., Tão, H., Gomes, F., et al. (2010). Gene expression profiles of human glioblastomas are associated with both tumor cytogenetics and histopathology. *Neuro. Oncol.* 12, 991–1003. doi:10.1093/NEUONC/NOQ050.

Zhao, Z., Zhang, K. N., Wang, Q., Li, G., Zeng, F., Zhang, Y., et al. (2021). Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients. *Genomics. Proteomics Bioinformatics* 19, 1–12. doi:10.1016/J.GPB.2020.10.005.

Have no fear of perfection; you'll never reach it.

Marie Curie

2

Master regulators associated with poor prognosis in Glioblastoma

This work is published in Biochemistry (Moscow), Supplement Series B: Biomedical Chemistry in November 2021

Kalya, M., Beißbarth, T. & Kel, A.E. Master Regulators Associated with Poor Prognosis in Glioblastoma Multiforme. Biochem. Moscow Suppl. Ser. B 15, 263–273 (2021). https://doi.org/10.1134/S1990750821040077 (link to the article) [*]

Availability of software, data and materials

The dataset analyzed in the current study, supplementary files and plots are available in the GitHub project here - Master regulators of poor prognosis in Glioblastoma [†]

[*] https://doi.org/10.1134/S1990750821040077

[†] https://github.com/genexplain/Manasa_KP_et_al_Master_regulators_of_poor_prognosis_in_Glioblastoma

Declaration of my contributions

This worked was conceptualized by Dr. Alexander Kel. I collected the raw data and clinical information for all GBM samples analysed. I took care of result interpretation and manuscript writing. Reviewing was done in stipulated time with the active participation of Dr. Alexander Kel. Prof. Edgar Wingender and Prof. Tim Beißbarth supervised the work and approved for submission.

ISSN 1990-7508, Biochemistry (Moscow), Supplement Series B: Biomedical Chemistry, 2021, Vol. 15, No. 4, pp. 263–273. © Pleiades Publishing, Ltd., 2021.
Russian Text © The Author(s), 2021, published in Biomeditsinskaya Khimiya.

Master Regulators Associated with Poor Prognosis in Glioblastoma Multiforme

M. Kalya[a, b], T. Beißbarth[a] , and A. E. Kel[b, c, *]

[a] *Department of Medical Bioinformatics, University Medical Center Göttingen, 37099 Göttingen, Germany*
[b] *geneXplain GmbH, Wolfenbüttel, 38302 Germany*
[c] *Institute of Chemical Biology and Fundamental Medicine SBRAS, Novosibirsk, 630090 Russia*
**e-mail: alexander.kel@genexplain.com*

Received April 25, 2021; revised May 5, 2021; accepted May 11, 2021

Abstract—Glioblastoma multiforme is a highly malignant brain tumor with average survival time of 15 months. Less than 2% of the patients survive beyond 36 months. To understand the molecular mechanism responsible for poor prognosis, we analyzed GBM samples of TCGA microarray (n = 560) data. We identified 720 genes that have a significant impact upon survival based on univariate cox regression. We applied the Genome Enhancer pipeline to analyze potential mechanisms of regulation of activity of these genes and to build gene regulatory networks. We identified 12 transcription factors enriched in the promoters of these genes including the key molecule of GBM—*STAT3*. We found that *STAT3* has significant differential expression across extreme survivor groups (short-term survivors– survival < 12 months and long-term survivors—survival > 36 months) and also has significant impact on survival. In the next step, we identified master regulators in the signal transduction network that regulate the activity of these transcription factors. Master regulators are filtered based on their differential expression across extreme survivor groups and impact on survival. This work validates our earlier report on master regulators *IGFBP2*, *PDGFA*, *OSMR* and *AEBP1* driving short survival. Additionally, we propose *CD14*, *CD44*, *DUSP6*, *GRB10*, *IL1RAP*, *FGFR3* and *POSTN* as master regulators driving poor survival. These master regulators are proposed as promising therapeutic targets to counter poor prognosis in GBM. Finally, the algorithm has prioritized several drugs for the further study as potential remedies to conquer the aggressive forms of GBM and to extend survival of the patients.

Keywords: glioblastoma, gene regulatory networks, master regulators, upstream analysis, STAT3, survival, short term survivors, transcription factors

DOI: 10.1134/S1990750821040077

INTRODUCTION

Glioblastoma multiforme (GBM) is the most common, highly malignant primary brain tumor [1] Despite huge developments in treatment strategies, there are as little as 2% of patients who actually respond to standard care and survive beyond 36 months (3 years), known as long-term survivors (LTS) [2]. Patients who survive less than 12 months are called short-term survivors (STS). The patient group with survival between 12 months to 36 months are called mid-term survivors (MTS).

Analysis of differentially expressed genes (DEGs) is an important and established in-silico strategy to identify potential molecules of cellular state transitions. Decreased expression of the *CHI3L1*, *FBLN4*, *EMP3*, *IGFBP2*, *IGFBP3*, *LGALS3*, *MAOB*, *PDPN*, *SERPING1* and *TIMP1* genes have been reported to be associated with prolonged survival [3–6] based on gene expression analysis of extreme survivor groups (STS & LTS). In our earlier work we reported ~200 genes differentially expressed between 113 STS and 58 LTS using publicly available datasets [7]. However, in that analysis we have excluded the MTS group which forms a majority of the patient population. In this work using the univariate Cox regression analysis on the entire TCGA (n = 560) dataset we identified ~720 genes that are associated with survival.

Reconstruction of the disease-specific regulatory networks can help identify potential master regulators of the respective pathological process. We used the Genome Enhancer (https://genexplain.com/genome-enhancer/), a multi-omics analysis tool, to reconstruct the regulatory network using the top 300 genes sorted by increasing FDR value and under the FDR cutoff of 0.05 that are identified using cox regression analysis. In this analysis, the first step is to analyze promoters and enhancers of genes for the transcription factors (TFs) involved in their regulation and, thus, important for the process under study; (2) re-constructing the signalling pathways that activate these TFs and identifying master regulators at the top of such pathways [8–10].

We applied the Genome Enhancer tool for the top 300 genes which had maximal impact on survival. At the 1st step, we identified important transcription factors enriched at the promoters of genes under study. Of them, *NR3C1* (GR) and *STAT3* were found to have highest regulatory scores signifying their role in controlling the expression of genes that encode master regulators. *STAT3* is an important GBM regulator, which induces cell proliferation, glioma stem cell maintenance, tumor invasion, angiogenesis, and immune evasion [11]. Next, we identified master regulators which regulate activity of these TFs. Out of them, 4 master regulators, namely; *IGFBP2*, *PDGFA*, *AEBP1*, *OSMR* were reported to drive poor prognosis in GBM in our earlier published work [7]. Here, we report *POSTN*, *CD14*, *CD44*, *DUSP6*, *FGFR3*, *GRB10* and *IL1RAP* to be playing critical roles in driving poor survival in GBM.

This work aims to explain the gene regulatory network in GBM which drives poor survival. The identified master regulators can point ways to block a pathological regulatory cascade. Suppression of components of the regulatory network may stop the pathological process.

MATERIALS AND METHODS

Data Collection

The raw gene expression profiles for Glioblastoma multiforme patients and their corresponding clinical information for GBM patients were collected from TCGA legacy archive [12]. The dataset contains 560 samples. 540 samples belonging to 526 patients have survival information. Duplicates are not removed in the study. There are 271 short-term survivors (STS; survival < 12 months), 240 mid-term survivors (MTS; 12 months < survival < 36 months) and 49 long-term survivors (LTS; survival > 36 months) with GBM, respectively. Sample information and cleaned datasets are given in Github supplementary materials (Tables S1-A, S1-B).

Affymetrix Microarray Data Pre-Processing

The raw data files (CEL format) of U133 Affymetrix microarray were preprocessed using RMA algorithm in R (affy package) for background correction, quality check and normalization to obtain log2 transformed expression values [13]. Batch correction of the pooled expression data for various data collection centers was performed using empirical Bayes framework is performed [14]. This batch corrected file is used for further analysis (Supplementary materials, Fig. S1). Multiple Affymetrix ids were summarized to genes ids by choosing the maximum out of probe intensities of multiple probes belonging to a single gene. The final expression matrix comprised 13914 probes and 560 samples.

Identification of Differentially Expressed Genes

The differential gene expression analysis between STS and LTS groups of GBM, from the batch corrected TCGA-GBM dataset was performed using Limma [15] with FDR cutoff of 5%. The analysis revealed 191 genes that are significantly differentially expressed more than 0.5 fold (DEGs) (*adj*. *p*-value < 0.05)

Impact on Survival

Survival and Survminer libraries in R were used to perform univariate survival analysis. Univariate Cox regression for survival analysis was performed using coxph function to calculate hazard ratio (HR) and FDR value corrected for multiple testing. We identified 720 genes with FDR cutoff of 0.05 which were used in the further upstream analysis in Genome Enhancer. KM plots are generated using 50% non-overlapping upper and lower quantiles based on median of expression values.

Databases Used in the Study

Transcription factor binding sites in promoters and enhancers of genes under study were analyzed using known DNA-binding motifs described in the TRANSFAC® library, release 2019.3 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transfac) [16].The master regulator search uses the TRANSPATH® database, release 2019.3 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transpath) [17]. A comprehensive signal transduction network of human cells is built by Genome Enhancer software based on reactions annotated in TRANSPATH®. The Ensembl database build 99.38 (http://www.ensembl.org) [18] was used for gene IDs representation.

Analysis of Pathway Enrichment

To explore the biological importance of gene signatures, the pathway enrichment analysis is performed using Binomial distribution to compute *p*-value and using Benjamin-Hochberg procedure to compute adjusted *p*-value. The pathway enrichment of 720 genes was done by mapping the input genes to canonical pathways in TRANSPATH® and Reactome databases.

Genome Enhancer

The approaches mentioned above helps us in understanding the impact of the genes under study in GBM biology. To understand the reason behind this dysregulation, the Genome Enhancer tool of geneXplain is used. This incorporates an automated pipeline for the previously published "upstream analysis" [8, 9] and the advanced approach "walking path-

ways" [10]. The genes which had significant (FDR < 0.05) impact on survival are used in this analysis. The workflow works in 2 steps: (1) analyzing promoters and enhancers of the genes for the transcription factors (TFs) involved in their regulation and, thus, important for the process under study; (2) re-constructing the signalling pathways that activate these TFs and identifying master regulators at the top of such pathways. For the first step, the database TRANSFAC® is employed together with the TF binding site identification algorithms MATCH™ and CMA [19, 20]. The second step involves the signal transduction database TRANSPATH® and special graph search algorithms. The tool also generates a visualization output of selected master regulators and also maps the log2FC and p-values to color the nodes on the created regulatory network.

2.6. Drug Prioritisation

We seek for the optimal combination of molecular targets (key elements of the regulatory network of the cell) that potentially interact with pharmaceutical compounds from a library of approved drugs (more than 9200 drugs) and pharmaceutically active known chemical compounds (2507 compounds), using information about known drugs from HumanPSD™ [21] and predicting potential drug compounds using PASS program.

We select drugs from the HumanPSD™ database that have at least one target. Next, we prioritise these drugs using "Drug rank" that is the sum of two other ranks: rank by 'Target activity score" ($T\text{-}score_{PSD}$) and rank by "Disease activity score" ($D\text{-}score_{PSD}$).

"Target activity score" ($T\text{-}score_{PSD}$) is calculated as follows:

$$T\text{-}score_{PSD} = -\frac{|T|}{|T| + w(|AT| - |T|)} \sum_{t \in T} \log_{10}\left(\frac{rank(t)}{1 + \max Rank(T)}\right),$$

where T is set of all targets related to the compound intersected with input list, $|T|$ is number of elements in T; AT and $|AT|$ are the set of all targets related to the compound and number of elements in i; w is weight multiplier, $rank(t)$ is rank of given target, $maxRank(T)$ equals $max(rank(t))$ for all targets t in T.

We use the following formula to calculate "Disease activity score" ($D\text{-}score_{PSD}$):

$$D\text{-}score_{PSD} = \begin{cases} \sum_{d \in D} \sum_{p \in P} phase(p,d) \\ 0,\ D = \emptyset \end{cases},$$

where D is the set of selected diseases, and if D is an empty set, $D\text{-}score_{PSD} = 0$. P is a set of all known phases for each disease, $phase(p, d)$ equals to the phase number if there are known clinical trials for the selected disease on this phase and zero otherwise.

For prioritization of active chemical compounds using PASS, we are using a precomputed database which was built by applying PASS software (a chemo-informatics approach based on SAR/QSAR) to library of which performs analysis of the structures of 2507 chemical compounds of known drugs (from HumanPSD). PASS predicts potential pharmacological activities of those substances, their possible side and toxic effects, as well as the possible mechanisms of action (targets). All biological activities are expressed as probability values for a substance to exert this activity (Pa).

So, we select chemical compounds with at least 2 targets (corresponding to the predicted activity-mechanisms) from our list of targets for which PASS predicted $Pa > 0.3$. Next, we prioritize these compounds in the similar way as before using PASS-based "Drug rank" that is the sum of two other ranks: rank by PASS-based "Target activity score" ($T\text{-}score$) and rank by PASS-based "Disease activity score" ($D\text{-}score$) that are calculated as follows. PASS-based "Target activity score":

$$T\text{-}score(s) = -\frac{|T|}{|T| + w(|AT| - |T|)} \times \sum_{m \in M(s)} Pa(m) \sum_{g \in G(m)} IAP(g) optWeight(g),$$

where $M(s)$ is the set of activity-mechanisms for the given structure (which passed the chosen threshold for activity-mechanisms Pa); $G(m)$ is the set of targets (converted to genes) that corresponds to the given activity-mechanism (m) for the given compound; $pa(m)$ is the probability to be active of the activity-mechanism (m), $IAP(g)$ is the invariant accuracy of prediction for gene from $G(m)$; $optWeight(g)$ is the additional weight multiplier for gene. T is set of all targets related to the compound intersected with input list, $|T|$ is number of elements in T, AT and $|AT|$ are set of all targets related to the compound and number of elements in it, w is weight multiplier set by a user.

PASS-based "Disease activity score":

$$D\text{-}score(s) = \max_{m \in M(s,g)}(Pa(m)),$$

where $S(g)$ is the set of structures for which target list contains given target, $M(s,g)$ is the set of activity-mechanisms (for the given structure) that corresponds to the given gene, $Pa(m)$ is the probability to be active of the activity-mechanism (m), $IAP(g)$ is the invariant accuracy of prediction for the given gene.

RESULTS AND DISCUSSION

Identification of Genes Having Impact on Survival

The univariate Cox regression analysis revealed 720 genes, which had a significant impact on survival. Among the genes with the highest hazard ratio there are: *PDCD1LG2* (HR = 2.1), *PPA3* (HR = 1.8),

SIGLEC9 (HR = 1.7) and with the lowest hazard ratio *MLNR* (HR = 0.28), *ZNF208* (HR = 0.35), and *NEUROG1* (HR = 0.37). Survival impact of all 720 genes is given in supplementary materials (Table S2-A**).**

Pathway Enrichment

The pathway enrichment analysis of 720 genes was done by mapping the input genes to canonical pathways in TRANSPATH® and Reactome databases. We have revealed 35 TRANSPATH short chains and pathways and 22 Reactome pathways enriched with the proteins encoded by the genes analysed. Among the revealed signaling pathways we found: "beta-catenin network", "hypoxia pathway", IL-6 signaling pathways and chains, "cell-cell communication", "Interferon signaling". It is known that many of these pathways play an important role in several stages of tumor progression. Full list of enriched pathways is given in supplementary materials (Tables S2-B, S2-C**).** In the Fig. 1 below we focus our attention on two most statistically significant chains involved in STAT3 signaling and in hypoxia regulation with several genes that are characterized by extreme GBM survival hazard ratio values.

Analysis of Enriched Transcription Factor Binding Sites and Composite Modules

In the next step, analysis of transcription factor binding sides on the promoters (–1000 bp upstream of transcription start site (TSS)) of the 720 genes is performed using the TF binding motif library of the TRANSFAC® database. CMA method is applied to identify the transcription factors that through their cooperation provide a synergistic effect and thus have a great influence on the gene regulation process. Using CMA, we identified two modules of transcription factors controlling the expression of the genes: module1: *MYOGNF1, JUN, NFATC2, AP2, LEF1, TFAP2A, HOXA10*; module2: *E2A, CEBPA, GR, MYOGNF1, MZF1, IK, NF1C, STAT3* (supplementary materials, Fig. S2 and Table S2-D**).** These two models together perform a reasonably good discrimination of promoters of the 720 genes from the promoters of housekeeping genes (Wilcoxon test p-value = 3.7×10^{-33}; AUC = 0.74 (which is significantly higher than expected for a random set of regulatory regions, Z-score = 4.44).

Out of them, we pay our attention to three TFs that are involved in controlling expression of the revealed genes: *JUN*, *STAT3,* and *GR***.** JUN is a protooncogene that plays a critical role in cell proliferation and malignant transformation with its levels reported to be elevated in GBM. Glucocorticoid receptor (GR) is reported to promote stem cells-like phenotype and resistance to chemotherapy [22]. STAT3 is a very important glioblastoma related transcription factor. Persistent activation of *STAT3* induces cell proliferation, anti-apoptosis, glioma stem cell maintenance, tumor invasion, angiogenesis, and immune evasion [11, 23]. Here, we report that *STAT3* is significantly differentially expressed in short-term survivors (log2FC = 0.403427421, *adj. p*-value = 0.00129). *STAT3* also had a significant impact on survival with HR = 1.4 (p-value = 0.0015 with FDR = 0.009) Fig. 2. *STAT3* is suggested to be a therapeutic target to control tumorigenesis by shaping tumor immune microenvironment [11].

Finding Master Regulators in Networks

Master regulators of the revealed transcription factors were identified using signal transduction database TRANSPATH® and the Genome Enhancer algorithm that searches for key-nodes in the global signal transduction network upstream of transcription factors as it is described previously [8, 9] and filters identified key-nodes by the criteria of presence of positive feedback loops [10] ("walking pathways" approach), requiring that the key-node proteins should be expressed by the genes that are found to be under the regulatory control of the same key-nodes.

As the result of such master-regulator analysis of the revealed 720 genes associated with the GBM survival we identified 43 potential feedback-loop-controlled master regulators. The full list of identified master regulators is given in supplementary materials (Table S2-E). These master regulators map mainly to such signaling pathways as: beta-catenin network, EGF pathway, stress-associated pathway and Cytokine Signaling. We have constructed a heatmap of the expression values of these 43 master-regulator genes (see Fig. S3 in the supplementary materials). It splits them into two subgroups of genes—in average with higher expression in the STS group (*POSTN, IGFBP2, FGFR3* and others) and with average higher expression in LTS group (*CASP9, PARD3, APEX1*). Still, we can see very high variability of the expression of these genes in each group.

Identification of Perspective Drug Targets

The identified master regulators can be considered as key candidates for therapeutic targets as they have a master effect on regulation of intracellular pathways that activate the pathological process of our study. In order to select the most prospective drug targets we, first of all, filtered all found master regulators based on their differential expression (log2FC > 0.5) between short-term and long-term survivors. Full results of Limma analysis are given in supplementary materials (Table S2-F). We identified several master regulators out of which 16 had significant differential expression between extreme survivor groups (full list is given in supplementary materials, Table S2-G**)**. Of them, 11 master regulators which had higher differential expressions (log2FC > 0.5) across extreme survivor

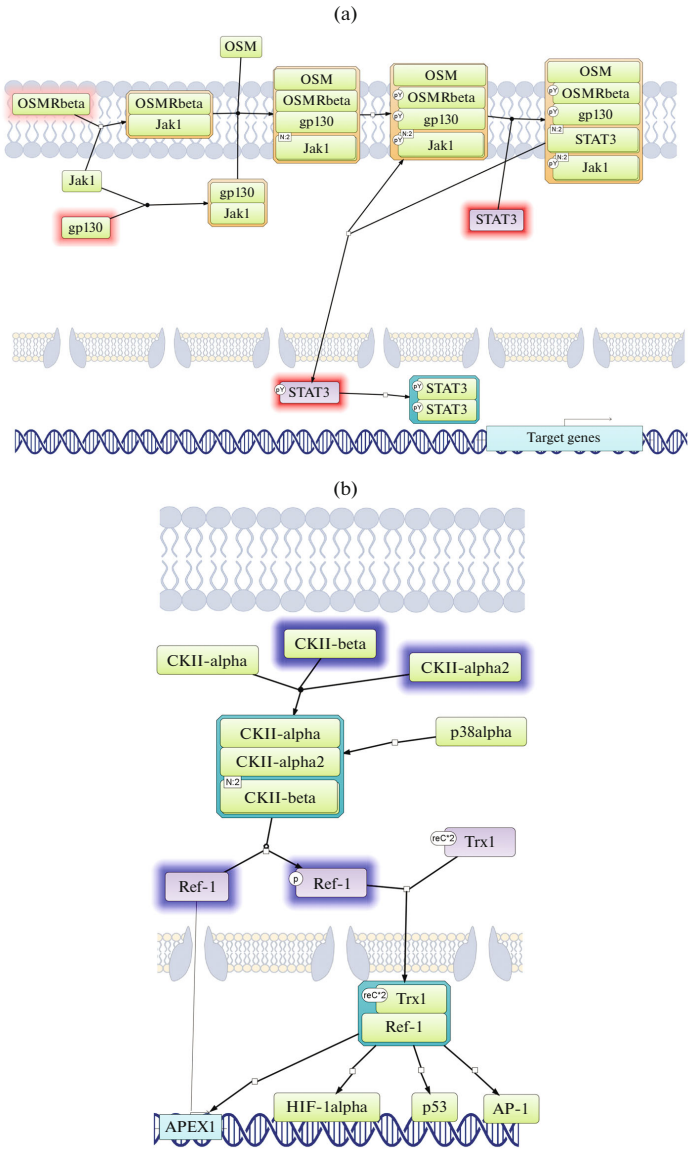

Fig. 1. Two top signal transduction chains (from TRANSPATH) significantly enriched by gene with extreme GBM survival hazard ratio values. (a) *OSM* → *STAT3* chain with three gene products with increased hazard ratios (red shadows around nodes) (p-value $< 6.5 \times 10^{-4}$). (b) *CHII* → *AP-1* chain with three gene products with decreased hazard ratios (blue shadows) (p-value $< 6.5 \times 10^{-4}$). The node shade coloring is done according to the Cox regression values. The color version is available in the electronic version of the article.

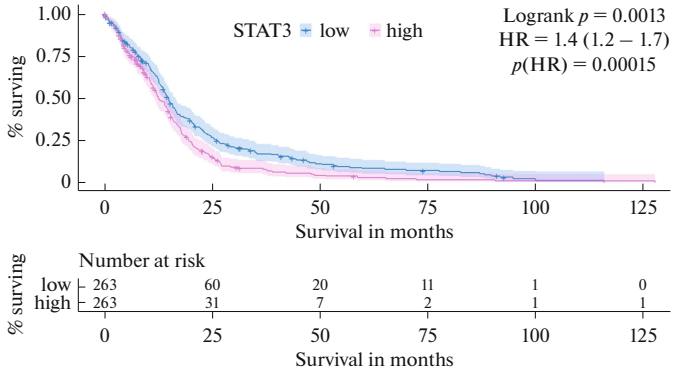

Fig. 2. The Kaplan-Meier plot to depict impact of *STAT3* on survival using 526 samples of TCGA-GBM microarray data for STAT3 transcription factor. Hazard ratio (HR) and statistical significance (p(HR)) according to Cox survival estimates are mentioned.

groups were considered for reconstructing gene-regulatory networks. Below, we characterize these genes from the point of view of their involvement in neoplasm pathology.

Periostin (POSTN), a secreted extracellular matrix protein is reported to play a major role in GBM progression, invasiveness and a potential role in the clinical response to angiogenic therapy [25]. Growth factor receptor-binding protein 10 (GRB10), known substrate of mTOR has been suggested as a major downstream effector of PI3K-AKT signalling with tumor promoting effects in prostate cancer [26]. It is reported to have high expression in mesenchymal subtype but lower expression in G-CIMP tumor subtypes of GBM [27]. Fibroblast growth factor receptor gene (*FGFR*) aberrations have been implicated in tumor development and progression and include *FGFR* overexpression, amplification, mutations, splicing isoform variations, and *FGFR* translocations in many cancers [28]. Nonetheless, *FGFR* expression changes in astrocytes can lead to malignant transformation and GBM progression due to the activation of mitogenic, migratory, and antiapoptotic responses [28]. *IGFBP2* is considered as one of the strongest biomarkers of aggressive behavior in GBM [29, 30] and also a prognostic marker for survival [30, 31]. *IGFBP2* along with *AEBP1 (ACLP), PDGFA* and *OSMR* are reported to be master regulators driving short survival in GBM and are discussed in detail in our earlier work [7].

CD14, which modulates cellular and humoral immune response by interacting directly with T and B cells, is suggested to play a major role in immunodepletion which contributes to the grim prognosis of GBM [32]. *CD44* overexpression predicts poor survival in GBM, plays a role in GBM progression and its role as a therapeutic target are reported earlier [33, 34]. *DUSP6*, encoding dual specificity phosphatase 6, is reported to be overexpressed in GBM and is suggested to play a vital role in epithelial-mesenchymal transition and tumor progression [35].

Next, since we observed very high variability of expression values between different samples, some important drug targets can be found among genes, whose expression is specifically high in a subset of ST samples only. Therefore, we have looked also at all other potential feedback-loop-controlled master regulators, paying particular attention to their well known or PASS predicted role as drug targets in GBM, in other neoplasms or in other related diseases.

The identified master regulators that may govern pathology associated genes were checked for druggability potential (*Drugability score*) using HumanPSD™ [21] database of gene-disease-drug assignments and PASS software [22] for prediction of biological activities of chemical compounds on the basis of a (Q)SAR approach. The *Drugability score* represents the number of drugs that are known to act on the corresponding target either according to the information extracted from medical literature (from HumanPSD™ database) or according to cheminformatics predictions of compounds activity against the examined target (from PASS software). So, we did the further selection of the drug targets using the *Drugability score* and have added three additional targets: *APEX1, MAPK8,* and *PTK2B*. These three targets, although have relatively low hazard ratio, but are characterized by quite high *Drugability score*.

APEX Nuclease (Multifunctional DNA Repair Enzyme) 1 is a DNA repair enzyme which is reported to positively correlate with altered MGMT status, signatures of Temozolomide treatment resistance, GBM recurrence and polarize towards immune-suppressive microenvironment in GBM [36]. Downregulation of *APEX1* is said to enhance sensitivity to Temozolomide

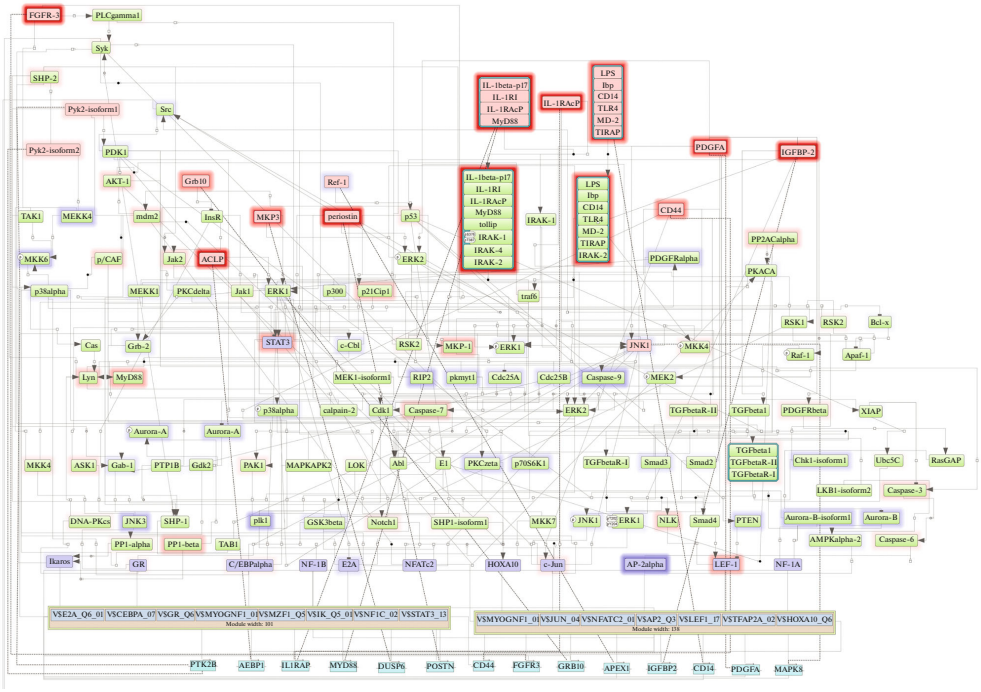

Fig. 3. Signal transduction and gene regulatory network of master regulators (red nodes) regulating two transcription factor modules (purple nodes) found as enriched in the promoters of genes under study. The dotted lines from genes to the encoding by them signaling proteins represent the transcription and translation processes (positive feedback loops). Network is constructed using master regulators which have significant log2FC > 0.5. *OSMR* is included to validate the previously reported drivers of short-survival [7]. Three additional master-regulators (*APEX1*, *MAPK8, PTK2B*) as known drug targets are also added to the diagram. The outside box filling is based on differential gene expression and is filled red when upregulated (logFC > 0.1) and blue if downregulated (log2FC < –0.1) in the current study. The color version is available in the electronic version of the article.

treatment in resistant GBM cell lines [37]. *MAPK8* and *MAPK* signaling is reported to be activated in Temozolomide resistant GBM cell lines. MAPK8 also enhances cell proliferation and inhibits apoptosis [38]. Expression of *PTK2B* (encoding PYK2, Protein Tyrosine Kinase 2 Beta) is suggested to play a critical role in migratory behavior of tumor cells thus leading to more aggressiveness in GBM [39]. In addition to these known targets, we can reveal two components of the beta-catenin pathway shown in Fig. 1b—*CSNK2A2* and *CSNK2B* as potential new targets, whose potential in therapy of aggressive forms of GBM is still to be confirmed [40].

The diagram of the master regulator network with positive feedback loops is shown in Fig. 3. The identified master regulators may potentially act as targets for therapeutic interventions.

Prioritisation of Potential Drugs

Finally, we ranked the drugs that are known or PASS predicted as active on the identified targets. The ranking of the drugs was done by *Drug rank* which is a sum of partial ranks computed on the basis of the target role as the potent master-regulator of the network mechanism (*Target activity score*), on the basis of the target *Druggability score* and *Disease activity score* (See the Materials and Methods section).

As a result, we identified and ranked the following drugs presented in Tables 1 and 2.

Among the top prioritized drugs, we can see several drugs that are known to be used or going through clinical trials on Glioma, Glioblastoma, Neoplasms of Central Nervous System, other Neoplasms. The algorithm has also proposed several drugs that can be suggested for further studies as repurposing candidates. Several of the identified drugs are currently going through various studies to confirm their potential role

Table 1. Most promising treatment candidates selected for the identified drug targets on the basis of literature curation in the HumanPSD™ database

Drug	Target names	Status (provided by Drugbank)	Clinical trials (Phase)	Target activity score	Disease activity score	Drug rank
Palifermin	FGFR3	Biotech, approved	Brain Abscess (1, 3, 4); Neoplasms (1, 2, 3); Leukemia (1, 2, 3); Multiple Myeloma (1, 2, 3, 4), Mucosis (3)	0.1115	0	16
Pazopanib	FGFR3	Small molecule, approved	**Glioma** (1, 2); Neoplasms (1, 2, 3, 4), **Central Nervous System Neoplasms** (2, 3)	0.0820	9	17
Leflunomide	PTK2B	Small molecule, approved, investigational	Arthritis (1, 2, 3, 4); Psoriatic (1); **Central Nervous System Neoplasms** (2, 3)	0.0874	0	20
Lenvatinib	FGFR3	Small molecule, approved	Adenocarcinoma (1, 2), Neoplasms (1, 2, 3)	0.0899	3	21
Nintedanib	FGFR3	Small molecule, approved	Adenocarcinoma (1, 2); Neoplasms (1, 2, 3), Pulmonary Fibrosis (3, 4)	0.0753	0	23
XL999	FGFR3	Small molecule, investigational	Lung neoplasms (1,2); Neoplasms (1, 2); Brain Abscess (2)	0.1267	0	24
Ponatinib	FGFR3	Small molecule, approved	Leukemia (1, 2, 3); Neoplasms (1, 2, 3)	0.0569	2	27
Hyaluronic acid	CD44	Small molecule, approved	Arthritis (1, 2, 3, 4); Osteoarthritis (1, 2, 3, 4); Glaucoma (4)	0.0009	0	28
Lucanthone	APEX1	Small molecule, approved, investigational	**Glioblastoma** (2), Neoplasms (2)	0.0755	2	29
Genistein	PTK2B	Small molecule, investigational	Carcinoma (1); Neoplasms (1, 2, 3); Bone Diseases (3, 4)	0.0559	0	30
Pyrazolanthrone	MAPK8	Small molecule, experimental		0.1050	0	30
Flavopiridol	CDK8	Small molecule, experimental, investigational	Carcinoma (1), Lymphoma (1, 2)	0.0488	0	35
Adenosine triphosphate	NAE1	Small molecule, approved, nutraceutical	Neoplasms (1), Pain (1); Alzheimer Disease (2)	0.0027	0	36

in treating Glioblastoma, including leflunomide (clin trial:NCT00003293), nintedanib [41], pamidronate [42], palifermin (in cell lines) [43]. In another study, meta-analysis showed that etoposide and teniposide improves survival in high-grade glioma [44]. Recently, enhanced efficiency of GBM treatment by doxorubicin was reported by combination with standard therapy [45].

CONCLUSIONS

The current work focuses on analysing TCGA-GBM microarray data (n = 560) to identify master regulators driving the expression of the genes which had significant impact on survival in GBM. We have reported important genes of survival and have computationally proposed gene regulatory networks driving poor prognosis using the "Genome Enhancer" pipeline. Out of the transcription factors reported, *STAT*3 was found to be important in terms of its regulatory scores, differential expression in extreme survivor groups as well as its impact on survival. Along with the earlier reported *IGFBP2, PDGFA, OSMR* and *AEBP1*, we propose 7 more master regulators, which can potentially act as therapeutic targets. We propose, *STAT3* and other transcription factors are in positive feedback loop with these master regulators to drive pathological self-enhancing processes leading to poor survival in GBM. On the basis of reconstructed master-regulatory signal transduction and gene regulatory network and chemoinformatic tool PASS we have identified the most promising drug targets and priori-

Table 2. Prospective drugs, predicted by PASS software to be active against the identified drug targets

Drug	Target names	Target activity score	Disease activity score	Drug rank
2,5,7-Trihydroxynaphthoquinone	*MAPK8*, *DUSP5*, *DUSP6*, *DUSP3*	0.0480	0.38	8
Teniposide	*APEX1*, *CASP9*	0.0443	0.378	12
Daunorubicin	*STAT3*, *APEX1*	0.0506	0.29	14
Doxorubicin	*STAT3*, *APEX1*	0.0423	0.308	18
Epirubicin	*STAT3*, *APEX1*	0.0423	0.308	18
Idarubicin	*STAT3*, *APEX1*	0.0426	0.293	18
Pyrazolanthrone	*MAPK8*, *PTK2B*, *FES*	0.0359	0.42	19
Etoposide	*APEX1*, *CASP9*	0.0343	0.422	19
Alendronate	*DUSP5*, *FGFR3*, *DUSP6*, *DUSP3*	0.0445	0.248	22
Pamidronate	*DUSP5*, *FGFR3*, *DUSP6*, *DUSP3*	0.0411	0.255	26
Fluorouracil	*FGFR3*, *PTK2B*, *FES*	0.0226	0.537	31
Oxybenzone	*MAPK8*, *DUSP5*, *DUSP6*, *DUSP3*	0.0463	0.202	36

tized drugs that can be potentially used for treating high grade GBM.

ACKNOWLEDGMENTS

The authors are grateful to E. Wingender for help in discussing this work.

FUNDING

This project has received funding from the European Union's Horizon 2020 research and the innovation program under the Marie Skłodowska-Curie grant agreement no 766069.

COMPLIANCE WITH ETHICAL STANDARDS

This article does not contain any research involving humans or the use of animals as objects.

CONFLICT OF INTEREST

Authors M. Kalya and A.E. Kel are employees of geneXplain.

SUPPLEMENTARY INFORMATION

*Availability of software, data and materials***:** The dataset analyzed in the current study, supplementary files and plots are available in the GitHub project here: https://github.com/genexplain/Manasa_KP_et_al_ Master_regulators_of_poor_prognosis_in_Glioblastoma

Supplementary materials are available in the electronic version of the article at the journal website (pbmc.ibmc.msk.ru).

REFERENCES

1. Wen, P.Y. and Kesari, S., *N. Engl. J. Med.*, 2008, vol. 359, pp. 492–507. https://doi.org/10.1056/NEJMra0708126
2. Krex, D., Klink, B., Hartmann, C., von Deimling, A., Pietsch, T., Simon, M., Sabel, M., Steinbach, J.P., Heese, O., Reifenberger, G., Weller, M., Schackert, G., German Glioma Network, *Brain*, 2007, vol. 130, no. 10, pp. 2596–2606. https://doi.org/10.1093/brain/awm204
3. de Vega, S., Iwamoto, T., and Yamada, Y., *Cell. Mol. Life Sci.*, 2009, vol. 66, pp. 1890–1902. https://doi.org/10.1007/s00018-009-8632-6
4. Bi, W.L. and Beroukhim, R., *Neuro-Oncology*, 2014, vol. 16, no. 9, pp. 1159–1160. https://doi.org/10.1093/neuonc/nou166
5. Reifenberger, G., Weber, R.G., Riehmer, V., Kaulich, K., Willscher, E., Wirth, H., Gietzelt, J., Hentschel, B., Westphal, M., Simon, M., Schackert, G., Schramm, J., Matschke, J., Sabel, M.C., Gramatzki, D., Felsberg, J., Hartmann, C., Steinbach, J.P., Schlegel, U., Wick, W., Radlwimmer, B., Pietsch, T., Tonn, J.C., von Deimling, A., Binder, H., Weller, M., Loeffler, M., German Glioma Network, Int. J. Cancer, 2014, vol. 135, vol. 8, pp. 1822–1831. https://doi.org/10.1002/ijc.28836
6. Franceschi, S., Mazzanti, C.M., Lessi, F., Aretini, P., Carbone, F.G., la Ferla, M., Scatena, C., Ortenzi, V., Vannozzi, R., Fanelli, G., Pasqualetti, F., Bevilacqua, G., Zavaglia, K., and Naccarato, A.G., Oncology Lett., 2015, vol. 10, no. 6, pp. 3599–3606. https://doi.org/10.3892/ol.2015.3738
7. Kalya, M.P., Kel, A., Wlochowitz, D., Wingender, E., and Beißbarth, T., *Front. Genet.*, 2021, vol. 12, https://doi.org/10.3389/fgene.2021.670240
8. Koschmann, J., Bhar, A., Stegmaier, P., Kel, A.E., and Wingender, E., *Microarrays* (Basel), 2015, vol. 4, no. 2, pp. 270–286. https://doi.org/10.3390/microarrays4020270

9. Boyarskikh, U., Pintus, S., Mandrik, N., Stelmashenko, D., Kiselev, I., Evshin, I., Sharipov, R., Stegmaier, P., Kolpakov, F., Filipenko, M., and Kel, A., *BMC Med Genomics*, 2018, vol. 11, 12. https://doi.org/10.1186/s12920-018-0330-5
10. Kel, A., Boyarskikh, U., Stegmaier, P., Leskov, L.S., Sokolov, A.V., Yevshin, I., Mandrik, N., Stelmashenko, D., Koschmann, J., Kel-Margoulis, O., Krull, M., Martínez-Cardús, A., Moran, S., Esteller, M., Kolpakov, F., Filipenko, M., and Wingender, E., *BMC Bioinformatics*, 2019, vol. 20, 119. https://doi.org/10.1186/s12859-019-2687-7
11. Chang, N., Ahn, S.H., Kong, D.-S., Lee, H.W., and Nam, D.-H., *Mol. Cell. Endocrinol.*, 2017, vol. 451, pp. 53–65. https://doi.org/10.1016/j.mce.2017.01.004
12. Grossman, R.L., Heath, A.P., Ferretti, V., Varmus, H.E., Lowy, D.R., Kibbe, W.A., and Staudt, L.M., *N. Engl. J. Med.*, 2016, vol. 375, pp. 1109–1112. https://doi.org/10.1056/NEJMp1607591
13. Gautier, L., Cope, L., Bolstad, B.M., and Irizarry, R.A., *Bioinformatics*, 2004, vol. 20, no. 3, pp. 307–315. https://doi.org/10.1093/bioinformatics/btg405
14. Leek, J.T., Johnson, W.E., Parker, H.S., Jaffe, A.E., and Storey, J.D., *Bioinformatics*, 2012, vol. 28, no. 6, pp. 882-883. https://doi.org/10.1093/bioinformatics/bts034
15. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K., *Nucleic Acids Res.*, 2015, vol. 43, no. 7, e47. https://doi.org/10.1093/nar/gkv007
16. Wingender, E., Dietze, P., Karas, H., and Knüppel, R., *Nucleic Acids Res.*, 1996, vol. 24, no. 1, pp. 238–241. https://doi.org/10.1093/nar/24.1.238
17. Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A., and Wingender, E., *Nucleic Acids Res.*, 2003, vol. 31, no. 1, pp. 97–100. https://doi.org/10.1093/nar/gkg089
18. Aken, B.L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., Fernandez Banet, J., Billis, K., García Girón, C., Hourlier, T., Howe, K., Kähäri, A., Kokocinski, F., Martin, F.J., Murphy, D.N., Nag, R., Ruffier, M., Schuster, M., Tang, Y.A., Vogel, J.-H., White, S., Zadissa, A., Flicek, P., and Searle, S.M.J., *Database*, 2016, vol. 2016, baw093. https://doi.org/10.1093/database/baw093
19. Kel, A.E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., and Wingender, E., *Nucleic Acids Res.*, 2003, vol. 31, no. 13, pp. 3576–3579. https://doi.org/10.1093/nar/gkg585
20. Waleev, T., Shtokalo, D., Konovalova, T., Voss, N., Cheremushkin, E., Stegmaier, P., Kel-Margoulis, O., Wingender, E., and Kel, A., *Nucleic Acids Res.*, 2006, vol. 34, Suppl. 2, W541–W545. https://doi.org/10.1093/nar/gkl342
21. Wingender, E., Hogan, J., Schacherer, F., Potapov, A.P., and Kel-Margoulis, O., *In Silico Biol.*, 2007, vol. 7, pp. S17–S25.
22. Filimonov, D.A., Druzhilovskiy, D.S., Lagunin, A.A., Gloriozova, T.A., Rudik, A.V., Dmitriev, A.V., Pogodin, P.V., and Poroikov, V.V., *Biomedical Chemistry: Research and Methods*, 2018, vol. 1, no. 1, e00004. https://doi.org/10.18097/BMCRM00004
23. Kostopoulou, O.N., Mohammad, A.-A., Bartek, J., Winter, J., Jung, M., Stragliotto, G., Söderberg-Nauclér, C., and Landázuri, N., *Int. J. Cancer*, 2018, vol. 142, no. 6, pp. 1266–1276. https://doi.org/10.1002/ijc.31132
24. Jahani-Asl, A., Yin, H., Soleimani, V.D., Haque, T., Luchman, H.A., Chang, N.C., Sincennes, M.-C., Puram, S.V., Scott, A.M., Lorimer, I.A.J., Perkins, T.J., Ligon, K.L., Weiss, S., Rudnicki, M.A., and Bonni, A., *Nat. Neurosci.*, 2016, vol. 19, pp. 798–806. https://doi.org/10.1038/nn.4295
25. Emini, E., Ramos-Moreno, T., Stefani, R.F., Svensson, A., and Bengzon, J., *J. Stem Cell Res. Ther.*, 2018, vol. 8, 414. https://doi.org/10.4172/2157-7633.1000414
26. Khan, M.I., Johani, A.A., Hamid, A., Ateeq, B., Manzar, N., Adhami, V.M., Lall, R.K., Rath, S., Sechi, M., Siddiqui, I.A., Choudhry, H., Zamzami, M.A., Havighurst, T.C., Huang, W., Ntambi, J.M., and Mukhtar, H., *FASEB J.*, 2019, vol. 33, no. 3, pp. 3198–3211. https://doi.org/10.1096/fj.201800265RR
27. Smith, A.A., Huang, Y.-T., Eliot, M., Houseman, E.A., Marsit, C.J., Wiencke, J.K., and Kelsey, K.T., *Epigenetics*, 2014, vol. 9, no. 6, pp. 873–883. https://doi.org/10.4161/epi.28571
28. Jimenez-Pascual, A. and Siebzehnrubl, F.A., *Cells*, 2019, vol. 8, no. 7, 715. https://doi.org/10.3390/cells8070715
29. Holmes, K.M., Annala, M., Chua, C.Y.X., Dunlap, S.M., Liu, Y., Hugen, N., Moore, L.M., Cogdell, D., Hu, L., Nykter, M., Hess, K., Fuller, G.N., and Zhang, W., *Proc. Natl. Acad. Sci. USA*, 2012, vol. 109, no. 9, pp. 3475–3480. https://doi.org/10.1073/pnas.1120375109
30. Phillips, L.M., Zhou, X., Cogdell, D.E., Chua, C.Y., Huisinga, A., Hess, K.R., Fuller, G.N., and Zhang, W., *J. Pathol.*, 2016, vol. 239, no. 6, pp. 355–364. https://doi.org/10.1002/path.4734
31. McDonald, K.L., O'Sullivan, M.G., Parkinson, J.F., Shaw, J.M., Payne, C.A., Brewer, J.M., Young, L., Reader, D.J., Wheeler, H.T., Cook, R.J., Biggs, M.T., Little, N.S., Teo, C., Stone, G., and Robinson, B.G., *J. Neuropathol. Exper. Neurol.*, 2007, vol. 66, no. 5, pp. 405–417. https://doi.org/10.1097/nen.0b013e31804567d7
32. Deininger, M.H., Meyermann, R., and Schluesener, H.J., *Acta Neuropathol.*, 2003, vol. 106, pp. 271–277. https://doi.org/10.1007/s00401-003-0727-9
33. Bradshaw, A., Wickremsekera, A., Tan, S.T., Peng, L., Davis, P.F., and Itinteang, T., *Front. Surg.*, 2016, vol. 3, 21. https://doi.org/10.3389/fsurg.2016.00021
34. Si, D., Yin, F., Peng, J., and Zhang, G., *CMAR*, 2020, vol. 2020, no. 12, pp. 769–775. https://doi.org/10.2147/CMAR.S233423

35. Zuchegna, C., di Zazzo, E., Moncharmont, B., and Messina, S., *BMC Res. Notes*, 2020, vol. 13, 374. https://doi.org/10.1186/s13104-020-05214-y
36. Hudson, A.L., Parker, N.R., Khong, P., Parkinson, J.F., Dwight, T., Ikin, R.J., Zhu, Y., Chen, J., Wheeler, H.R., and Howell, V.M., *Front. Oncol.*, 2018, vol. 8, 314. https://doi.org/10.3389/fonc.2018.00314
37. Montaldi, A.P., Godoy, P.R.D.V., and Sakamoto-Hojo, E.T., *Mutat. Res. Genet. Toxicol. Environ. Mutagen.*, 2015, vol. 793, pp. 19–29. https://doi.org/10.1016/j.mrgentox.2015.06.001
38. Xu, P., Zhang, G., Hou, S., and Sha, L., *Biomedicine Pharmacotherapy*, 2018, vol. 106, pp. 1419–1427. https://doi.org/10.1016/j.biopha.2018.06.084
39. Lipinski, C.A., Tran, N.L., Menashi, E., Rohl, C., Kloss, J., Bay, R.C., Berens, M.E., and Loftus, J.C., *Neoplasia*, 2005, vol. 7, no. 5, pp. 435–445. https://doi.org/10.1593/neo.04712
40. Zheng, Y., McFarland, B.C., Drygin, D., Yu, H., Bellis, S.L., Kim, H., Bredel, M., and Benveniste, E.N., *Clin. Cancer Res.*, 2013, vol. 19, no. 23, pp. 6484–6494. https://doi.org/10.1158/1078-0432.CCR-13-0265
41. Muhic, A., Poulsen, H.S., Sorensen, M., Grunnet, K., and Lassen, U., *J. Neurooncol.*, 2013, vol. 111, pp. 205–212. https://doi.org/10.1007/s11060-012-1009-y
42. Jarry, U., Chauvin, C., Joalland, N., Léger, A., Minault, S., Robard, M., Bonneville, M., Oliver, L., Vallette, F.M., Vié, H., Pecqueur, C., and Scotet, E., *OncoImmunology*, 2016, vol. 5, no. 6, e1168554. https://doi.org/10.1080/2162402X.2016.1168554
43. Brake, R., Starnes, C., Lu, J., Chen, D., Yang, S., Radinsky, R., and Borges, L., *Mol. Cancer Res.*, 2008, vol. 6, no. 8, pp. 1337–1346. https://doi.org/10.1158/1541-7786.MCR-07-2131
44. Leonard, A. and Wolff, J.E., *Anticancer Res.*, 2013, vol. 33, pp. 3307–3315.
45. Norouzi, M., Yathindranath, V., Thliveris, J.A., Kopec, B.M., Siahaan, T.J., and Miller, D.W., *Sci. Rep.*, 2020, vol. 10, 11292. https://doi.org/10.1038/s41598-020-68017-y

I was taught that the way of progress was neither swift nor easy.

Marie Curie

3

IGFBP2 is a potential Master regulator of poor prognosis in GBM

This work is published in Frontiers in Genetics in June, 2021

Kalya M, Kel A, Wlochowitz D, Wingender E, Beißbarth T. IGFBP2 Is a Potential Master Regulator Driving the Dysregulated Gene Network Responsible for Short Survival in Glioblastoma Multiforme. Front Genet. 2021 Jun 15;12:670240. DOI: 10.3389/fgene.2021.670240. PMID: 34211498; PMCID: PMC8239365 (link to the article) [*]

Availability of software, data, and materials

The datasets analyzed in the current study, software pipeline for data analysis, supplementary information are available here: IGFBP2 regulatory networks in Gliobastoma. [†]

[*] https://www.frontiersin.org/articles/10.3389/fgene.2021.670240/full

[†] https://github.com/genexplain/Manasa_KP_et_al_IGFBP2_regulatory_networks_in_Gliobastoma

Declaration of my contributions

This work was conceptualized by Dr. Alexander Kel. I collected datasets, merged, performed data analysis and data interpretation. I also prepared original draft of this publication. Dr. Alexander Kel has extensively participated in parameter tuning in upstream analysis workflow and validating methods used in the study. Darius Wlochowitz has given extensive inputs regarding the data analysis pipeline and has helped me improve the draft of the manuscript. Prof. Edgar Wingender has participated extensively in improving the manuscript and Prof. Tim Beißbarth has supervised the work, read the manuscript, and approved it for correction.

frontiers in Genetics

ORIGINAL RESEARCH
published: 15 June 2021
doi: 10.3389/fgene.2021.670240

IGFBP2 Is a Potential Master Regulator Driving the Dysregulated Gene Network Responsible for Short Survival in Glioblastoma Multiforme

Manasa Kalya[1,2], Alexander Kel[2,3], Darius Wlochowitz[1], Edgar Wingender[2] and Tim Beißbarth[1]*

[1] Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany, [2] geneXplain GmbH, Wolfenbüttel, Germany, [3] Institute of Chemical Biology and Fundamental Medicine SB RAS, Novosibirsk, Russia

Only 2% of glioblastoma multiforme (GBM) patients respond to standard therapy and survive beyond 36 months (long-term survivors, LTS), while the majority survive less than 12 months (short-term survivors, STS). To understand the mechanism leading to poor survival, we analyzed publicly available datasets of 113 STS and 58 LTS. This analysis revealed 198 differentially expressed genes (DEGs) that characterize aggressive tumor growth and may be responsible for the poor prognosis. These genes belong largely to the Gene Ontology (GO) categories "epithelial-to-mesenchymal transition" and "response to hypoxia." In this article, we applied an upstream analysis approach that involves state-of-the-art promoter analysis and network analysis of the dysregulated genes potentially responsible for short survival in GBM. Binding sites for transcription factors (TFs) associated with GBM pathology like NANOG, NF-κB, REST, FRA-1, PPARG, and seven others were found enriched in the promoters of the dysregulated genes. We reconstructed the gene regulatory network with several positive feedback loops controlled by five master regulators [insulin-like growth factor binding protein 2 (IGFBP2), vascular endothelial growth factor A (VEGFA), VEGF165, platelet-derived growth factor A (PDGFA), adipocyte enhancer-binding protein (AEBP1), and oncostatin M (OSMR)], which can be proposed as biomarkers and as therapeutic targets for enhancing GBM prognosis. A critical analysis of this gene regulatory network gives insights into the mechanism of gene regulation by IGFBP2 *via* several TFs including the key molecule of GBM tumor invasiveness and progression, FRA-1. All the observations were validated in independent cohorts, and their impact on overall survival has been investigated.

Keywords: glioblastoma, master regulators, upstream analysis, IGFBP2, FRA-1, FOSL1, short term survivors, transcription factors

OPEN ACCESS

Edited by:
Alessandro Laganà, Icahn School of Medicine at Mount Sinai, United States

Reviewed by:
Balaji Banoth, St. Jude Children's Research Hospital, United States
Daniela Albrecht-Eckardt, BioControl Jena GmbH, Germany

****Correspondence:***
Alexander Kel
alexander.kel@genexplain.com

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *20 February 2021*
Accepted: *06 April 2021*
Published: *15 June 2021*

Citation:
Kalya M, Kel A, Wlochowitz D, Wingender E and Beißbarth T (2021) IGFBP2 Is a Potential Master Regulator Driving the Dysregulated Gene Network Responsible for Short Survival in Glioblastoma Multiforme. Front. Genet. 12:670240. doi: 10.3389/fgene.2021.670240

INTRODUCTION

Glioblastoma multiforme (GBM) is the most common, highly malignant primary brain tumor (Wen and Kesari, 2008). Despite huge developments in treatment strategies, GBM poses unique treatment challenges due to tumor recurrence (34%) and drug resistance leading to poor survival rates of less than 15 months even after advanced chemoradiotherapy (Krex et al., 2007). As few as

2% of patients respond to standard therapy and survive beyond 36 months (Krex et al., 2007; Das et al., 2011), clinically called long-term survivors (LTS). Another group termed short-term survivors (STS) are those who survive less than 12 months (Shinawi et al., 2013). The factors that determine the long survival are not well understood.

Though several factors like age, gender, Karnofsky Performance Score, the extent of tumor resection, radiotherapy, and chemotherapy are associated with survival and treatment response (Scott et al., 1999; Lee et al., 2008; Sonoda et al., 2009; Zhang et al., 2012), it is evident from recent research that certain molecular signatures can be connected with treatment response and thereby survival. Promoter methylation of the gene MGMT, mutations in the genes IDH1/2, and loss of heterozygosity in chromosome 1p/19q have been confirmed to be highly informative (Krex et al., 2007; Das et al., 2011; Zhang et al., 2012; Han et al., 2014; Reifenberger et al., 2014; Franceschi et al., 2015; Chen et al., 2016). Furthermore, CHI3L1, FBLN4, EMP3, IGFBP2, IGFBP3, LGALS3, MAOB, PDPN, SERPING1, and TIMP1 gene expression has repeatedly been reported to be decreased in LTS patients (De Vega et al., 2009; Bi and Beroukhim, 2014; Han et al., 2014; Franceschi et al., 2015). A better characterization of these extreme survival groups at the molecular level will likely shed important light on the biological aspects that drive their malignancy and survival.

With the advent of gene expression profiling and remarkable developments in high-throughput technologies, it is possible to gain deeper molecular insights into disease biology. Databases like Gene Expression Omnibus—GEO (Barrett et al., 2013), Array Express (Athar et al., 2019), and The Cancer Genome Atlas—TCGA (Grossman et al., 2016) serve as open platforms for retrieval of high-quality multi-omics data to search for new markers in cancer research. The analysis of differentially expressed genes (DEGs) is already an important and established *in silico* strategy to identify potential drivers of cellular state transitions. For a more refined analysis, annotation of DEGs, using *a priori* known biological categories from the Gene Ontology (GO; Ashburner et al., 2000) and pathway databases, e.g., TRANSPATH®(Krull et al., 2003), KEGG (Kanehisa et al., 2020), PANTHER (Thomas et al., 2003), and Reactome (Jassal et al., 2020), has proven to be an effective hypothesis-driven approach in cancer research. Moreover, with the advent of state-of-the-art promoter analysis, it is now possible to establish gene regulatory networks computationally that can be used to understand the causes of gene dysregulation and for identification of causal master regulators driving them. In this regard, we applied the Genome Enhancer[1], a multi-omics analysis tool that makes use of the open-source programming environment BioUML (Kolpakov et al., 2019) and incorporates an automated pipeline for the previously published "upstream analysis" (Koschmann et al., 2015; Boyarskikh et al., 2018) and the "walking pathways" (Kel et al., 2019) approach. There are two major steps that constitute this strategy: (1) analysis of the promoters of DEGs to identify relevant transcription factors (TFs): this is done with the help of the TRANSFAC® database (Matys et al., 2006) and the binding site identification algorithms, MATCH™ (Kel et al., 2003, 2006) and CMA (Waleev et al., 2006); (2) reconstruction of signaling pathways that activate these TFs and identification of master regulators on the top of such pathways: for this, the signaling pathway database TRANSPATH® (Krull et al., 2003) has been employed in conjunction with special graph search algorithms that identify positive feedback loops (Kel et al., 2019).

In this study, we applied the upstream analysis to publicly available datasets of GBM from the GEO database to understand the gene-regulatory networks contributing to short survival in GBM. This regulatory network revealed a set of 12 TFs binding to the regulatory regions of the genes of interest and five master regulators regulating them, namely, (a) vascular endothelial growth factor A (VEGFA), a mediator of angiogenesis (Xu et al., 2013) and a promoter of stem-like cells in GBM; (b) PDGF, a highly amplified gene and key player of tumorigenesis (Martinho and Reis, 2011); (c) oncostatin M (OSMR), which orchestrates feed-forward signaling with EGFR and STAT3 to regulate tumor growth (Jahani-As et al., 2016); (d) adipocyte enhancer-binding protein (AEBP1), which plays a key role in pathogenesis through NF-κB activation (Majdalawieh et al., 2020); and (e) IGFBP2.

Insulin-like growth factor binding protein 2, a well-established molecule of interest in GBM (Yao et al., 2016), was found to be more highly expressed in STS and to have an impact on overall survival. IGFBP2 expression is said to be higher in all four (classical, mesenchymal, proneural, and neural) GBM subtypes (Lindström, 2019). It also drives gene programs for immunosuppression in the mesenchymal subtype and is suggested as an immunotherapeutic target (Liu et al., 2019). In non-mesenchymal subtypes (classical, proneural, and neural), it modulates cell proliferation (Phillips et al., 2016; Cai et al., 2018). It has also been found to be a marker of tumor aggressiveness and a prognostic marker for survival (Lindström, 2019). However, the molecular mechanism by which IGFBP2 affects disease progression and patient prognosis is not fully understood.

This work focuses on understanding gene regulatory networks that drive short survival in GBM and their master regulators, which we suggest as biomarkers and therapeutic targets. Later, we critically discuss the role of IGFBP2 in the gene regulatory network.

RESULTS

Identification of Differentially Expressed Genes

Identifying DEGs gives us insight into the biological semantics of a cellular state and helps to identify promising biomarkers of various disease states. The differential gene expression analysis between STS and LTS groups of GBM, from the batch-corrected GSE dataset, was performed using linear models for microarray data (LIMMA) (Ritchie et al., 2015) with FDR cutoff of 5%. The analysis revealed 957 genes that are significantly differentially expressed (DEGs) (adjusted p-value < 0.05). Furthermore, the analysis revealed 115 significantly upregulated (log2FC > 0.5) and 83 significantly downregulated [log2FC $< (-0.5)$] genes.

[1]https://genexplain.com/genome-enhancer/

The top five upregulated and downregulated genes and their corresponding log2FC are shown in **Table 1** and the full list is given in **Supplementary Table 1-A**.

Functional Annotation of Differentially Expressed Genes

Functional annotation was performed to investigate the biological roles of these DEGs. As shown in **Supplementary Figure 1A**, the top GO biological processes are extracellular structure and matrix organization with 30 DEG hits. **Supplementary Figure 1B** shows the results for GO cellular component enrichment, which revealed dysregulation of genes that encode proteins for the extracellular matrix and synaptic membranes. The important enriched molecular function GO terms are channel activity and transmembrane transporter activity (**Supplementary Figure 1C**). The disruption in extracellular matrix organization is one of the important signatures in glioblastoma treatment response dealing with invasiveness and malignancy (De Vega et al., 2009). Deeper biological insights are required in this aspect. It is interesting to see enrichment of genes known to be involved in glioma (**Figure 1A**). Gene signature enrichment based on hallmark gene sets of MSigDB clearly signifies the enrichment of epithelial-to-mesenchymal transition depicted in **Figure 1B**. The process of epithelial-to-mesenchymal transition plays a very important role in GBM survival by driving tumor invasiveness and drug resistance (Iwadate, 2016). Important pathways like Aurora signaling, G2/M phase transition, and TGF-β pathway are found to be enriched according to TRANSPATH® (**Table 2**). The full list of enrichment results can be found in **Supplementary Table 1-B**.

TABLE 1 | The list of the top five significantly upregulated and downregulated genes in STS identified in the GSE dataset.

Gene symbol	Description	Log2FC	*p*-Value	Adjusted *p*-value
Upregulated genes				
CHI3L1	Chitinase-3-like 1	1.371	9.73E−05	0.013
PDPN	Podoplanin	1.241	7.88E−07	0.002
MEOX2	Mesenchymal homeobox 2	1.159	6.45E−04	0.028
IGFBP2	Insulin-like growth factor binding protein 2	1.149	4.87E−05	0.010
COL6A2	Collagen type VI alpha 2 chain	1.0479	5.79E−05	0.011
Downregulated genes				
KLRC2	Killer cell lectin-like receptor C2	−1.2187	3.63E−04	0.022
KLRC1	Killer cell lectin-like receptor C1	−1.2187	3.63E−04	0.022
FUT9	Fucosyltransferase 9	−1.0709	1.15E−04	0.014
DPP10	Dipeptidyl peptidase-like 10	−1.02781	2.97E−05	0.008
GABRB3	Gamma-aminobutyric acid type A receptor subunit beta3	−0.96352	6.73E−05	0.011

Identifying the Master Regulators of Dysregulated Gene Networks

Reconstruction of the disease-specific regulatory networks can help to identify potential master regulators that may serve as mechanism-based biomarkers or as therapeutic targets to block a specific pathological regulatory cascade. Using the promoter analysis as a first step, we analyzed enrichment of TF binding sites in promoters of upregulated genes of STS using DNA-binding motifs from the TRANSFAC® library. Two hundred seventy-four TFs (**Supplementary Table 1-C**) enriched for CCKR signaling, interleukin signaling, PDGF signaling, and WNT signaling were found to have their binding sites enriched; full enrichment results can be found in **Supplementary Table 1-D**.

Next, we applied the Composite Module Analyst (CMA) and identified two modules involving 12 TF binding site combinations that regulate the expression of the genes of interest. CMA revealed the following modules comprising clustering binding sites for the following TFs: module 1: HNF3B, NANOG, NFKAPPAB, TAF1, TCF4, and FRA-1; module 2: PPARG, TAL1, REST, POU6F1, FOSJUN, and PBX. The modules and their significance are depicted in **Supplementary Figure 2**. Differential expression statistics for the 12 TFs are given in **Supplementary Table 2**. Among them, FRA-1 TF (also known as FOSL1) was found to be *p*-value significant and upregulated in STS of GBM (log2FC = 0.023, *p*-value = 0.008, adjusted *p*-value (0.093) (**Supplementary Table 2**).

Figure 2 validates the predicted cluster of TF binding sites from the composite modules identified in the promoter of IGFBP2 gene. We can see that binding sites for the TFs – c-Fos/c-Jun, Nanog, Tal-1, and HNF3/FoxA1 in this cluster can be confirmed by publicly available ChIP-seq data of the GTRD database (Kolmykov et al., 2021). In addition, binding site of FRA-1 can be confirmed by a cluster of mapped reads of independent publicly available ChIP-seq data (FRA1 track in **Figure 2**) (full map is shown in the **Supplementary Figure 4**).

Finally, we reconstructed signaling network that activates the TFs revealed by CMA analysis and thereby identifying the top regulators in these networks using the TRANSPATH® database. With this approach, we identified five important master regulators that are plausible drivers of short survival in GBM: IGFBP2, VEGFA/VEGF165, platelet-derived growth factor A (PDGFA), AEBP1, and OSMR. All the master regulators were found to be significantly upregulated in STS. The genes that encode the master regulator proteins are controlled by the TFs revealed by CMA in their promoters, which maintains the multiple positive feedback loops in the system. It should be underlined here that, in such networks with positive feedback loops, the identified key TFs, such as FRA-1, are both upstream of their target genes, among them the IGFBP2, as well as downstream from the master regulator proteins, one of them the IGFBP2 protein. The regulatory network reconstructed with six master regulators is shown in **Figure 3**, and the master regulators and their log2FC in STS are listed in **Table 3**. Since

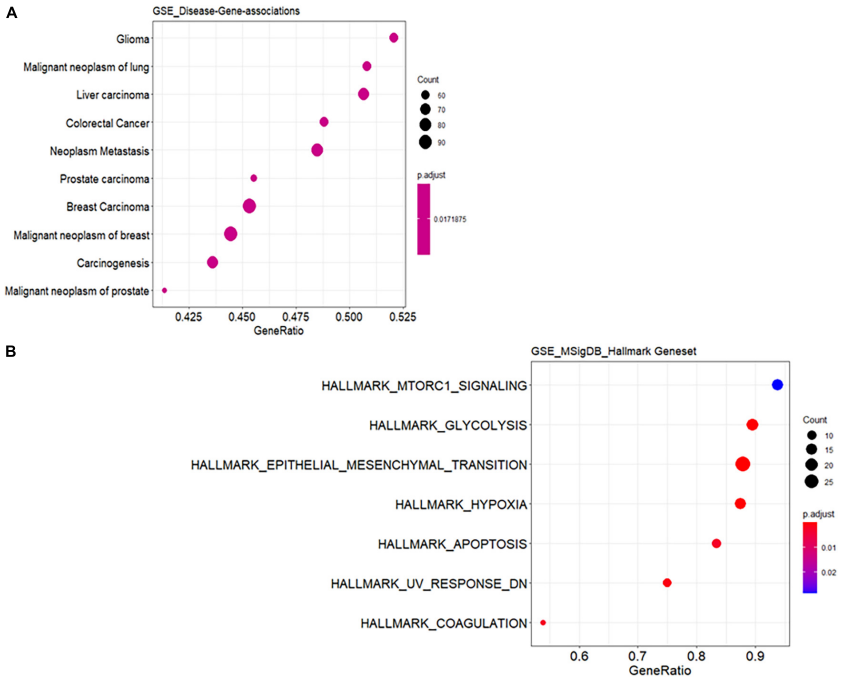

FIGURE 1 | Functional enrichment analysis of differentially expressed genes (DEGs). **(A)** Enrichment for known disease gene networks in different diseases. *Y*-axis represents enriched ontology categories and *X*-axis represents gene ratio. Gene ratio is count/set size. The "count" is the number of genes that belong to a given gene set, while "set size" is the total number of genes in the gene set. *Y*-axis is sorted based on leading edge. Leading edge is a subset of genes that contributes most to the enrichment score. The dots are sized based on gene ratio and are colored according to their adjusted *p*-value. **(B)** Enrichment for hallmark gene sets in the molecular signature database similar to **(A)**.

TABLE 2 | Pathway enrichment using the TRANSPATH® pathway (2019.3) for differentially expressed genes.

ID (TRANSPATH)	Title	Group size	Expected hits	Nominal *p*-value	ES	Rank at max	NES	FDR	Number of hits
CH000001004	Aurora-A cell cycle regulation	68	67.262	0	0.422	8,347	4.138	0	68
CH000000919	Cyclosome regulatory network	77	76.164	0	0.349	7,336	3.728	0	77
CH000000694	G2/M phase (cyclin B: Cdk1)	66	65.284	0	0.375	6,641	3.587	0	66
CH000000879	Caspase network	83	82.099	0	0.333	8,414	3.523	0	83
CH000000711	TGFbeta pathway	153	151.340	0	0.232	8,431	3.346	0	151

VEGF165 is a splice variant of VEGFA, only the latter will be considered further on.

Validating the Expression of Master Regulators in Other Cohorts

The expression patterns of the master regulators identified above have been validated in two different cohorts: (A) TCGA-GBM microarray data (Grossman et al., 2016) and (B) GSE16011 (Gravendeel et al., 2009). The expression patterns were similar, and there is a significant upregulation of all master regulators except for VEGFA (GSE16011: adjusted *p*-value = 0.069 and TCGA-GBM: adjusted *p*-value = 0.075) (**Supplementary Tables 1-E, 1-F**). The differential expression values are given in **Table 4**.

Validating the Master Regulators in the TCGA-GBM Cohort

The TCGA-GBM microarray data containing 271 STS and 49 LTS is used to validate the above-identified drivers of short survival. The data is preprocessed and adjusted for batch effects

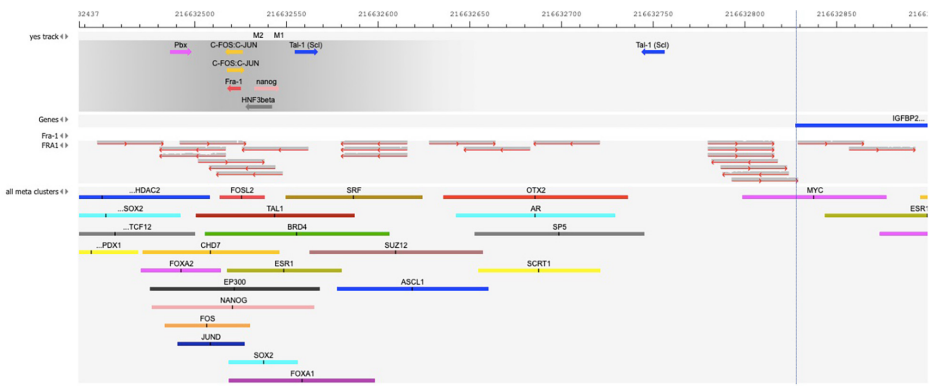

FIGURE 2 | Map of the cluster of transcription factor (TF) binding sites of the composite model identified within the promoter of IGFBP2 gene [–1000 to + 100 bp relative to TSS]. The position of the TSS (the beginning of the first exon on IGFBP2 gene) is shown by the vertical dotted line. "Yes track" represents the cluster of identified TF binding sites of the composite model within the promoter. The direction of the arrows gives the orientation of the PWMs. The names of TFs binding to these sites are shown above the arrows. The track "FRA1" represents the mapped reads of the FRA1 (also called FOSL1) ChIP-seq data of GEO, GSM803382. The reads were mapped on the hg38 human genome using Subread aligner (Liao et al., 2013) with default parameters. The track "all meta clusters" shows all known meta-clusters in this region from the GTRD database that represent the overlapping fragments of peaks for one particular TF from several ChIP-seq experiments. The name of TF is shown above each meta-cluster. Several predicted TF binding sites in the composite model are confirmed in independent ChIP-seq experiments: several overlapping reads of FRA1 ChIP-seq data in the "FRA1" track and FOSL2 meta-cluster in the GTRD confirm the predicted site for Fra-1; FOS and JUN meta-clusters in the GTRD confirm the predicted c-Fos/c-Jun binding sites; NANOG meta-cluster confirms the predicted Nanog binding site; TAL1 meta-cluster confirms the predicted Tal-1 binding site; FOXA2 and FOXA1 meta-clusters of the GTRD confirm the HNF3beta binding site.

(**Supplementary Figure 3**), and a differential gene expression analysis is performed. Same cutoffs for log2FC and adjusted *p*-value are used. We identified 171 genes upregulated in STS of GBM (log2FC > 0.5 and adjusted *p*-value < 0.05) (full list in **Supplementary Table 1-E**). Forty-nine of them were in common between the GSE dataset and TCGA-GBM; the full differential gene expression analysis results are given in **Supplementary Table 1-G**. Composite models selected by the CMA algorithm across the two datasets were expected to vary. We identified a model that includes a set of 16 TFs (**Supplementary Table 3**) and 12 master regulators upstream of them (**Supplementary Table 4**) regulating the signal transduction and gene regulatory network in STS.

As a result, the TCGA-GBM dataset validates IGFBP2, AEBP1 (ACLP), and PDGFA as master regulators driving the dysregulated gene network in STS. We also found that binding sites for FRA-1 TF are statistically significantly enriched at the regulatory regions of the dysregulated genes including IGFBP2 in the TCGA-GBM cohort (**Supplementary Table 5**).

Impact of Master Regulators on Survival in GBM

Univariate survival analysis was used to study the impact of these master regulators and the TFs they regulate on the overall survival in GBM based on TCGA-RNA-seq data. Patients are split into non-overlapping 50% upper and lower quantiles. Additionally, cox regression for the univariate survival analysis is performed, and hazard ratio (HR) and corresponding *p*-values are shown in **Figure 4**. Univariate survival Cox regression analysis on other microarray datasets is given in **Supplementary Table 1-I**. All master regulators were found to have a significant impact upon survival except VEGFA. FRA-1 (FOSL1) was found to have a significant HR.

Master Regulator Expression Patterns Across GBM Subtypes

Based on the regulatory landscape of GBM, there are four subtypes—classical, mesenchymal, proneural, and neural (Verhaak et al., 2010). There is a significant level of intertumoral as well as intra-tumoral heterogeneity within each of them (Verhaak et al., 2010; Bradshaw et al., 2016). Molecular subtypes of GBM in the GSE dataset is given in **Supplementary Table 6**. DEGs between STS and LTS within each subtype are given in **Supplementary Table 7**. The expression patterns of master regulators across subtypes and across survival groups are depicted as boxplot in **Supplementary Figure 5**. None of the master regulators were found to be significantly differentially expressed between survivor groups in any subtypes.

DISCUSSION

Gene regulatory networks represent the causal regulatory relationships between TFs and their gene targets, which enables us to discover dysregulated genes in certain biological states (Marbach et al., 2012). Comparative studies of STS and LTS of

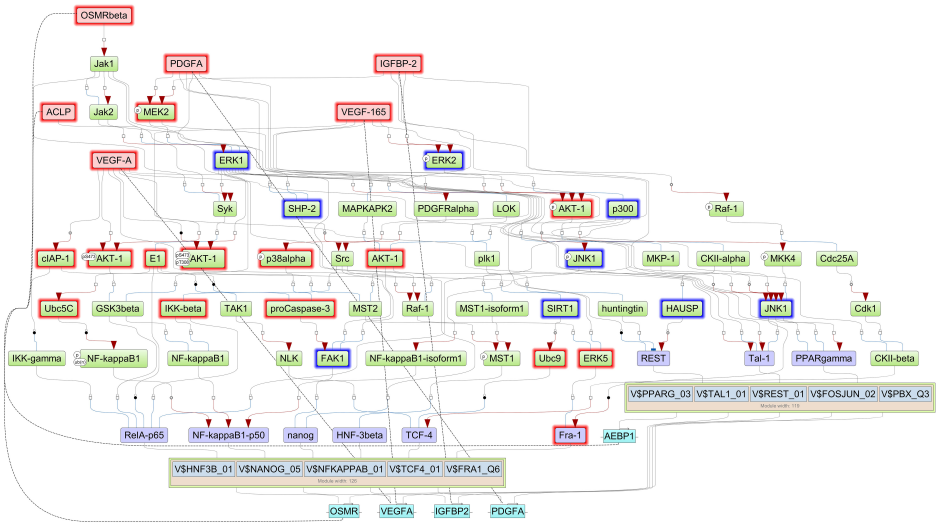

FIGURE 3 | Signal transduction and gene regulatory network of six master regulators (red nodes) regulating two transcription factor modules (purple nodes) enriched in promoters of highly upregulated genes of short-term survivors (STS). The dotted lines from genes to such signaling proteins represent the transcription and translation processes (positive feedback loops). The outside box filling is based on log2FC < (−0.2) and is filled red when upregulated (log2FC > 0.2 and *p*-value < 0.05) and filled blue when downregulated (log2FC < 0.2 and *p*-value < 0.05) in the current study.

GBM showed that gene expression programs executed across survival groups vary significantly. In the light of these findings, we sought to apply an upstream analysis approach to gain an insight about gene regulatory networks driving the short survival.

In the promoter analysis, we identified a set of 12 TFs in composite clusters that are enriched in the promoter regions of dysregulated genes in STS (upregulated in STS). For several of these TFs, a connection to GBM has previously been established. The TFs NANOG and REST are critical for self-renewal and maintenance of oncogenic signatures in glioblastoma stem-like cells (Kamal et al., 2012; Bradshaw et al., 2016); PPARG has emerged as a promising therapeutic target as its agonists increased median survival in GBM patients (Ellis and Kurian, 2014); NF-κB is implicated in several processes like invasion, epithelial–mesenchymal transition (Yamini, 2018), resistance to radiotherapy (Avci et al., 2020), and maintenance of cancer stem-like cells (da Hora et al., 2019); and FRA-1/FOSL1 has been reported to be important in maintenance/progression of malignant glioma (Debinski and Gibo, 2005). FRA-1 along with JUN-B modulates a malignant feature of GBM by regulating the expression of the metalloproteinases like MMP-2 and MMP-9 (Kesari and Bota, 2011). Among these 12 TFs, we found that FRA-1 has a significant impact upon survival and has a higher expression in STS. Debinski and Gibo (2005) hypothesized that any AP1-stimulating signals like epidermal growth factor (EGF), leukemia inhibitory factor, OSMR, or FGF-2 can positively regulate FRA-1. VEGF-D is regulated by FRA-1 (supporting the feedback loop found in our work) and is a known prognostic factor in other aggressive cancers (Debinski et al., 2001; Azar et al., 2014).

A graph analysis of the signal transduction network upstream of these TFs identified five potential master regulators that

TABLE 3 | Table of the master regulators identified, their description, log2FC in STS, and number of transcription factors regulated.

Molecule name	Gene description	HGNC gene symbol	Log2FC in STS	Number of TFs regulated
IGFBP2	Insulin-like growth factor binding protein 2	IGFBP2	1.149	9
ACLP	AE-binding protein 1	AEBP1	0.782	9
VEGFA	Vascular endothelial growth factor A	VEGFA	0.778	9
VEGF165	Vascular endothelial growth factor A	VEGFA	0.778	9
OSMRbeta	Oncostatin M receptor	OSMR	0.634	8
PDGFA	Platelet-derived growth factor subunit A	PDGFA	0.529	9

TABLE 4 | Expression of the master regulators identified across survival groups (STS and LTS, respectively) and across three datasets (GSE,GSE16011 and TCGA-GBM microarray).

Master regulator	GSE		GSE16011		TCGA	
	Log2FC (STS vs LTS)	Adjusted *p*-value	Log2FC (STS vs LTS)	Adjusted *p*-value	Log2FC (STS vs LTS)	Adjusted *p*-value
IGFBP2	1.149	4.87E−05	2.030	4.598E−04	1.098	5.00E−06
AEBP1	0.782	7.75E−05	1.723	0.001	0.971	3.96E−06
PDGFA	0.529	4.55E−04	1.680	4.709E−09	0.825	2.07E−05
VEGFA	0.778	5.20E−04	0.884	0.069	0.500	0.0752
OSMR	0.634	8.65E−04	1.957	4.24E−05	0.486	0.0318

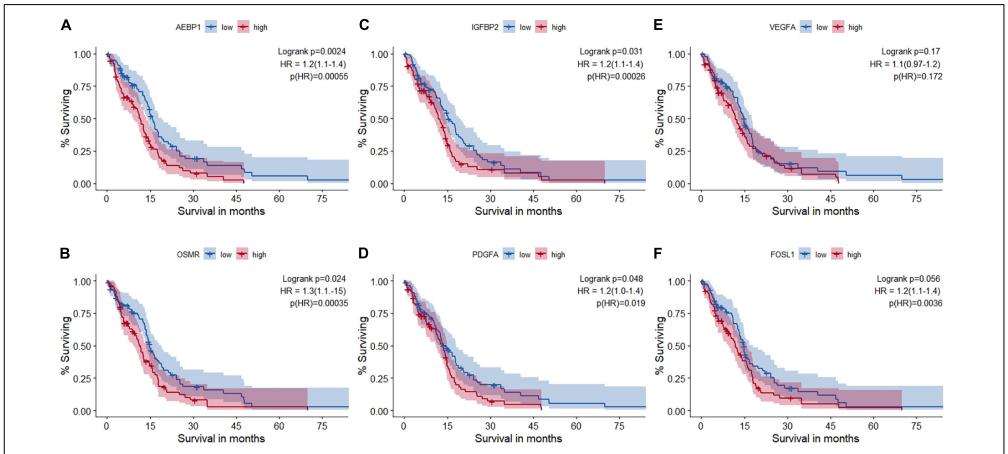

FIGURE 4 | Survival analysis using RNA-seq data of The Cancer Genome Atlas glioblastoma (TCGA-GBM) cohort. **(A)** AEBP1, **(B)** OSMR, **(C)** IGFBP2, **(D)** PDGFA, **(E)** VEGFA, and **(F)** FRA-1 (FOSL1). Out of the five master regulators, all but VEGFA (AEBP1, OSMR, PDGFA, and IGFBP2) had a statistically significant impact on survival. Hazard ratio (HR) and statistical significance [*p* (HR)] according to Cox survival estimates are mentioned.

might explain gene dysregulation in STS, namely, insulin-like growth factor binding protein 2 (IGFBP2), VEGFA, its isoform VEGF165, PDGFA, OSMR, and AEBP1. All the identified master regulators were upregulated in STS, and their expression patterns were validated computationally in two other independent cohorts. We found that the expression of all master regulators, with the exception of VEGFA, was correlated with overall survival in the GBM patients. IGFBP2, AEBP1, and PDGFA master regulators driving short survival were validated as master regulators of short survival in the TCGA-GBM microarray cohort. Out of them, IGFBP2 had higher expression in STS. The IGFBP2 is said to be one of most potential glioma oncogenes and functions as a hub of oncogenic signaling pathways by regulating pro-tumorigenic signals of tumor initiation and progression. Earlier studies have suggested IGFBP2 to drive EMT and as a potential therapeutic target in mesenchymal GBM (Yamini, 2018; Liu et al., 2019). It is established that exogenous IGFBP2 promotes proliferation, invasion, and chemoresistance to temozolomide in glioma cells *via* integrin β1 by promoting ERK phosphorylation and nuclear translocation (Schütt et al., 2004; Yau et al., 2015). IGFBP2 is considered as one of the strongest biomarkers of aggressive behavior in GBM (Holmes, 2012; Phillips et al., 2016) and also a prognostic marker for survival (McDonald et al., 2007; Phillips et al., 2016).

Here, we propose that IGFBP2 can be a potential regulator of FRA-1 TF. IGFBP2-induced RAF/MAPK signaling can activate FRA-1 (**Figure 3**). It has been shown earlier that IGFBP2 and FRA-1 regulate transcription of VEGF (Debinski et al., 2001; Azar et al., 2011, 2014), which is the second most dysregulated master regulator in our network. Enhanced ERK signaling, triggered by these master regulators, may lead to mitogen-induced FRA-1 transcription (Adiseshaiah et al., 2005) as well as its protection from proteasomal degradation (Vial and Marshall, 2003). The gene regulatory network deduced here suggests that FRA-1 mediates a positive feedback loop where it activates transcription of master regulator genes in cooperation with other TFs, which in turn cause an increase in FRA-1 activity. Promoters of the genes of all five master regulators reported in the study contain potential binding sites for FRA-1. Experimental evidences that IGFBP2 can drive GBM invasion by enhancing MMP2 expression

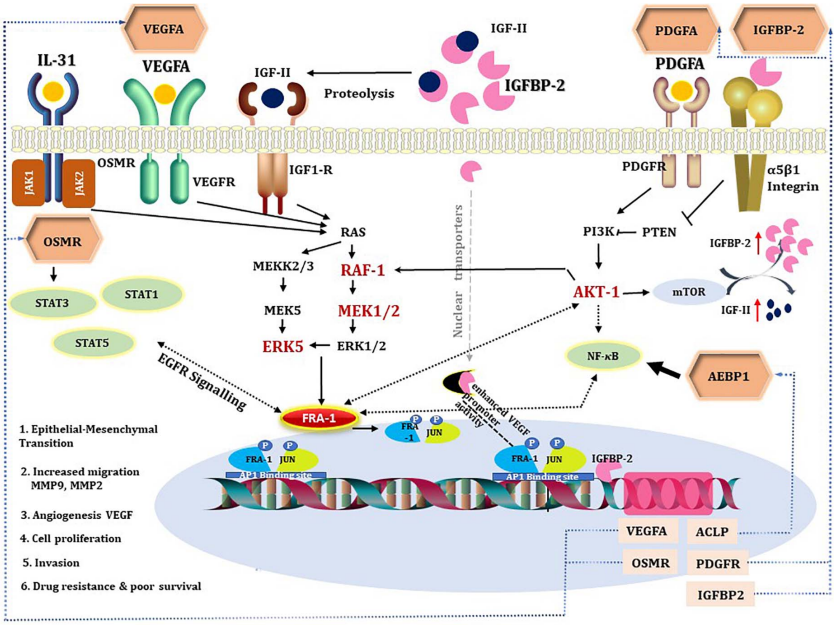

FIGURE 5 | A diagram combining prior knowledge about the role of IGFBP2 in GBM and the gene regulatory network developed in the study. The hexagons are master regulators identified in our analysis. All the intermediates of the gene regulatory network are colored red if upregulated in STS and black if not present in the network. Dotted black line indicates the knowledge is through literature and continuous black line if known through gene regulatory network. Blue dotted lines represent gene regulatory connections between master regulators and their corresponding genes transcribed by target transcription factors. VEGFA, PDGFA, IGFs, and IL-31 activate RAF/MEK/ERK signaling, which mediate cell survival through PI3K-AKT pathway (Yao et al., 2016; Simpson et al., 2017). MEK2/RAF1/ERK5 and AKT-1 are found to be upregulated in STS, suggestive of activated ERK signaling, which can contribute to drug resistance (Abrams et al., 2010; Salaroglio et al., 2019). IGFBP2 activates IGFR either by increasing bioavailability of IGFs or by direct interaction with its functional domain. Integrin acts as receptor for IGFBP2 extracellular signals (Schütt et al., 2004; Yau et al., 2015) and modulates NF-κB signaling. IGFBP2 by nuclear translocation (Azar et al., 2011) is involved in transcriptional regulation of the VEGF gene and modulates angiogenesis (Azar et al., 2011). STAT3 and NF-κB are said to be the two major downstream transcription factors of IGFBP2 that direct tumorigenic intracellular signaling (Phillips et al., 2016) *via* EGFR signaling. Oncostatin M, a receptor for cytokine IL31, is a regulator of EGFR signaling (Jahani-As et al., 2016). FRA-1 is required for AKT activation in cancers to promote AKT-dependent cell growth (Zhang et al., 2016). NF-κB can regulate AP1 (FOS and JUN) thereby VEGF expression in pancreatic tumor cell lines (Fujioka et al., 2004). All the five master regulators have binding sites for FRA-1. In the figure, we depicted the possible positive feedback loop between FRA-1 and the master regulators to orchestrate a complex tumorigenic program of invasiveness, migration, drug resistance, and angiogenesis.

(Wang etal., 2003) support our computational prediction of IGFBP2 as a therapeutic target. Hence, the gene regulatory networks proposed by our computational analysis suggest a novel molecular mechanism associated with GBM survival in which FRA-1 acts as a transcription regulator of IGFBP2. The study of Kesari and Bota (2011) confirmed our hypothesis that IGFBP2 can enhance GBM invasion *via* TF AP1 (FOS-JUN). Metalloproteinases like MMP-2/MMP-9 have been reported earlier to be regulated by FRA-1 in several cancers including GBM (Debinski and Gibo, 2005; Adiseshaiah et al., 2008; Kimura et al., 2011; Prywes and Henckels, 2013). Taking these findings together, our work proposes that the regulation of IGFBP2 gene expression *via* AP1 (FOS-JUN) can be an important mechanism of GBM invasion. An overview of the gene regulatory network developed in this work and supporting literature evidence is illustrated in **Figure 5**.

In summary, our work proposes a gene regulatory network associated with STS in GBM, which is regulated by five master regulators, namely, IGFBP2, VEGFA, PDGFA, OSMR, and AEBP1. Furthermore, these five master regulators may present biomarkers of GBM prognosis and/or as therapeutic targets for enhancing survival in GBM. This work also proposes a novel mechanism of gene dysregulation by IGFBP2 by modulating a key molecule of tumor invasiveness and progression—FRA-1 TF. All the genes encoding these five master regulators have binding sites for FRA-1 in their promoters. FRA-1 and the master regulators cooperate in a positive feedback loop to orchestrate a complex tumorigenic program leading to poor survival in GBM.

TABLE 5 | Statistics of datasets under study.

	Platform	Short-term survivors	Long-term survivors
GSE53733 (Reifenberger et al., 2014)	HU133 plus 2.0 arrays	16	23
GSE108474 (Gusev et al., 2018)	HU133 plus 2.0 arrays	97	35

The datasets with labels GSE' were collected from the GEO database.

MATERIALS AND METHODS

Data Collection

The genome-wide expression profiles based on Human Genome U133 plus 2.0 array and clinical information of patients with GBM were collected from the public repository of GEO database—GSE108474 (Gusev et al., 2018)[2] and GSE53733 (Reifenberger et al., 2014)[3]. The two datasets were pooled together leading to 113 and 58 samples corresponding to STS (survival <12 months) and LTS (survival >36 months) with GBM, respectively (**Table 5**). Duplicates were not removed. Sample information and cleaned datasets are given in GitHub.

Affymetrix Microarray Data Pre-processing

The raw data files (.CEL format) for GSE108474 and GSE53733 were collected from the GEO database—from here on called as GSE dataset. RMA algorithm is used in R (affy package) for background correction, quality check, and normalization to obtain log2-transformed expression values (Gautier et al., 2004). Batch correction of the pooled expression data was performed using empirical Bayes framework (Leek et al., 2012). This batch-corrected file is used for further analysis. Multiple Affymetrix IDs were summarized to gene IDs by choosing the maximum out of the probe intensities of multiple probes belonging to a single gene. The final expression matrix comprised 21,526 probes and 171 samples.

Differential Gene Expression Analysis

The LIMMA method was applied to identify DEGs (Ritchie et al., 2015). It is an efficient tool that is stable even for experiments with small samples. A differential gene expression analysis of 171 samples of the GSE dataset was performed with Benjamini–Hochberg adjusted *p*-value. Nine hundred fifty-seven genes were significantly (adjusted *p*-value < 0.05) differentially expressed (DEGs). One hundred fifteen of them were significantly upregulated (adjusted *p*-value < 0.05 and log2FC > 0.5) and 83 were significantly downregulated [adjusted *p*-value < 0.05 and log2FC $< (-0.5)$].

[2]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE108474
[3]https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE53733

Databases Used in the Study

Transcription factor binding sites in promoters and enhancers of DEGs were analyzed using known DNA-binding motifs described in the TRANSFAC® library, release 2019.3 (geneXplain GmbH, Wolfenbüttel, Germany)[4] (Wingender et al., 1996). The master regulator search uses the TRANSPATH® database, release 2019.3 (geneXplain GmbH, Wolfenbüttel, Germany)[5] (Krull et al., 2003). A comprehensive signal transduction network of human cells is built by the Genome Enhancer software based on reactions annotated in TRANSPATH®. The information about drugs corresponding to identified drug targets and clinical trials references were extracted from the HumanPSD™ database (Wingender et al., 2007), release 2020.2[6]. The Ensembl database build 99.38[7] (Aken et al., 2016) was used for gene ID representation and GO[8] (Ashburner et al., 2000) was used for functional classification of the studied gene set.

Functional Annotation

To explore the biological importance of gene signatures, a gene set enrichment analysis is performed. All the adjusted *p*-value significant genes were used. GSEA is an efficient method to determine whether the genes of interest show statistically significant enrichment between different biological states. GO enrichments for cellular component, biological process, and molecular functions were performed. To investigate the top enriched ontology terms, 1,000 random permutations were done and an adjusted *p*-value cutoff of 0.05 is used. The dysregulated gene network enrichment also gives a useful insight about known disease signatures (Subramanian et al., 2005). The hallmark gene set of MSigDB (Liberzon et al., 2011) defines specific biological states or processes. Enrichment analysis is performed in R using DOSE package (Yu et al., 2015). PANTHER pathway enrichment of the identified TFs was performed using the EnrichR tool (Chen et al., 2013). TRANSPATH® (Krull et al., 2003) pathway enrichment was performed using the geneXplain platform.

Genome Enhancer Pipeline

The approaches mentioned above help us in understanding the impact of the DEGs in GBM biology. To understand the reason behind this dysregulation, the genome enhancer pipeline of geneXplain is used. The genome enhancer is a multi-omics analysis service (see text footnote 1) that is built using an open-source programming environment BioUML (Kolpakov et al., 2019)[9] and incorporates an automated pipeline for the previously published "upstream analysis" (Koschmann et al., 2015; Boyarskikh et al., 2018) and the advanced approach "walking pathways" (Kel et al., 2019). Significantly upregulated genes in STS were used in this workflow.

The workflow works in 2 steps.

[4]https://genexplain.com/transfac
[5]https://genexplain.com/transpath
[6]https://genexplain.com/humanpsd
[7]http://www.ensembl.org
[8]http://geneontology.org
[9]www.biouml.org

A. Analysis of enriched transcription factor binding sites and composite modules

Binding of TFs to the specific sites in promoters and enhancers is the key to the transcriptional regulation of genes. Identifying clusters of binding sites for TFs (composite modules) in the upstream regulatory regions [−1,000 bp upstream of transcription start site (TSS)] of the genes of interest is a determining step to understand the gene regulatory mechanism (composite regulatory modules) (Kel-Margoulis et al., 2002).

We use the CMA (Waleev et al., 2006) to detect such potential enhancers, as targets of multiple TFs bound to the regulatory regions of the genes of interest. The TFs are ranked based on (a) the yes/no ratio: given a set of promoter sequences of dysregulated genes, denoted as a yes set, and promoter sequences of unchanged genes under the same experimental condition, denoted as a no set, motifs are considered important if they have a high yes/no ratio, the ratio of motif occurrences per promoter in yes and no sets, and a statistically significant enrichment of occurrences in yes sequences assessed by the binomial *p*-value. (b) A regulatory score, which is a measure of involvement of a TF in controlling the expression of genes that encode master regulators. CMA identifies the TFs that, through their cooperation, provide a synergistic effect and thus have a great influence on the gene regulation process.

B. Finding master regulators in networks

The second step involves the signal transduction database TRANSPATH® and special graph search algorithms to identify common regulators of the revealed TFs. These master regulators appear to be the key candidates for therapeutic targets as they have a master effect on the regulation of intracellular pathways that activate the pathological process of our study. Master regulators regulating the TFs revealed in step A are ranked based on (a) logFC, (b) CMA score, which signifies how strong is the potential for this gene to be regulated by TFs of interest, and (c) master regulator score, which signifies how strong is the potential of this gene product to regulate the activity of those TFs. Selected master regulators can also be visualized and with the possibility to map the logFC and *p*-value on the created regulatory network.

Validation of Observed Gene Signatures

The raw microarray data of 560 TCGA-GBM samples were downloaded from TCGA legacy. The GSE16011 raw.CEL data was downloaded from the GEO repository. Both raw datasets were processed and analyzed independently following same steps as mentioned earlier. These two datasets are used to observe and validate the expression pattern of master regulators across the two survival groups (see **Table 6**). GSE16011 comprises of data generated at a single center and is used in several studies (Prasad et al., 2020), unlike TCGA. TCGA-GBM microarray data PCA plots are given **Supplementary Figure 3**, and no significant batch effects in the context of survival groups were found.

TABLE 6 | Statistics of the two validation datasets.

Datasets	Platform	Short-term survivors	Long-term survivors
GSE16011 (Gravendeel et al., 2009)	HU133 plus 2.0 arrays	93	16
TCGA-GBM microarray (Grossman et al., 2016)	HU133	271	49

Validation of Master Regulators

The TCGA-GBM microarray data downloaded from TCGA legacy archive is processed in the same fashion as GSE. Similar cutoffs (log2FC and *p*-value) and parameters are used to identify enriched TFs and network analysis in order to understand drivers of gene regulatory networks in short survival.

Impact on Survival

Master regulators and their target TFs affect the whole regulatory network and therefore can have an independent impact on survival in GBM patients. Level 3 RNA-seq data and clinical data for 152 TCGA-GBM cohort is downloaded using the TCGAbiolinks package in R. Survival and survminer libraries in R were used to perform a univariate survival analysis. A univariate survival analysis was used to understand the impact of individual master regulator on survival in GBM with non-overlapping 50% upper and lower quantiles. Additionally, a univariate Cox regression for survival analysis was performed using the coxph function of the survival package to calculate the HR with *p*-value cutoff of 0.05 for significance.

CONCLUSION

In the work presented, we have identified candidate master regulators responsible for gene dysregulation in STS. These candidates have sufficient experimental evidence toward their role in GBM. Out of reported five master regulators, IGFBP2 is established as the most promising master regulator. Through the gene regulatory network analysis, we propose that IGFBP2 and FRA-1 are in a positive feedback loop that may lead to a pathological self-enhancing process responsible for poor survival in GBM.

DATA AVAILABILITY STATEMENT

The results of the analysis performed using the Genome Enhancer in geneXplain platform are available here: https://github.com/genexplain/Manasa_KP_et_al_IGFBP2_regulatory_networks_in_Gliobastoma.

AUTHOR CONTRIBUTIONS

AK involved in conceptualization, providing resources, supervision, and manuscript reviewing. MK had conceptualized

the work, performed the data collection and data analysis, interpreted the results, and wrote the manuscript. DW supervised the data analysis pipeline and manuscript writing. EW and TB involved in supervision of work and in reviewing the draft. All authors contributed to the article and approved the submitted version.

FUNDING

This project had received funding from the European Union's Horizon 2020 Research and Innovation Programme under the Marie Skłodowska-Curie grant agreement no. 766069 and by German Ministry of Education and Research (BMBF) e:Med project MyPathSem (031L0024).

ACKNOWLEDGMENTS

It is the authors' pleasure to acknowledge Ravi Kumar Nadella who has read and given comments on every version of this paper and their colleague Philip Stegmaier for his expertise, discussions, and assistance in improving the manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2021.670240/full#supplementary-material

REFERENCES

Abrams, S. L., Steelman, L. S., Shelton, J. G., Wong, E. W. T., Chappell, W. H., Bäsecke, J., et al. (2010). The Raf/MEK/ERK pathway can govern drug resistance, apoptosis and sensitivity to targeted therapy. *Cell Cycle* 9, 1781–1791. doi: 10.4161/cc.9.9.11483

Adiseshaiah, P., Peddakama, S., Zhang, Q., Kalvakolanu, D. V., and Reddy, S. P. (2005). Mitogen regulated induction of FRA-1 proto-oncogene is controlled by the transcription factors binding to both serum and TPA response elements. *Oncogene* 24, 4193–4205. doi: 10.1038/sj.onc.1208583

Adiseshaiah, P., Vaz, M., Machireddy, N., Kalvakolanu, D. V., and Reddy, S. P. (2008). A Fra-1-dependent, matrix metalloproteinase driven EGFR activation promotes human lung epithelial cell motility and invasion. *J. Cell. Physiol.* 216, 405–412. doi: 10.1002/jcp.21410

Aken, B. L., Ayling, S., Barrell, D., Clarke, L., Curwen, V., Fairley, S., et al. (2016). The ensembl gene annotation system. *Database* 2016:baw093. doi: 10.1093/database/baw093

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., et al. (2000). Gene ontology: tool for the unification of biology. *Nat. Genet.* 25, 25–29. doi: 10.1038/75556

Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., et al. (2019). ArrayExpress update - From bulk to single-cell expression data. *Nucleic Acids Res.* 47, D711–D715. doi: 10.1093/nar/gky964

Avci, N. G., Ebrahimzadeh-Pustchi, S., Akay, Y. M., Esquenazi, Y., Tandon, N., Zhu, J. J., et al. (2020). NF-κB inhibitor with Temozolomide results in significant apoptosis in glioblastoma via the NF-κB(p65) and actin cytoskeleton regulatory pathways. *Sci. Rep.* 10:13352. doi: 10.1038/s41598-020-70392-5

Azar, W. J., Azar, S. H. X., Higgins, S., Hu, J. F., Hoffman, A. R., Newgreen, D. F., et al. (2011). IGFBP-2 enhances VEGF gene promoter activity and consequent promotion of angiogenesis by neuroblastoma cells. *Endocrinology* 152, 3332–3342. doi: 10.1210/en.2011-1121

Azar, W. J., Zivkovic, S., Werther, G. A., and Russo, V. C. (2014). IGFBP-2 nuclear translocation is mediated by a functional NLS sequence and is essential for its pro-tumorigenic actions in cancer cells. *Oncogene* 33, 578–588. doi: 10.1038/onc.2012.630

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., et al. (2013). NCBI GEO: archive for functional genomics data sets - Update. *Nucleic Acids Res.* 41, D991–D995. doi: 10.1093/nar/gks1193

Bi, W. L., and Beroukhim, R. (2014). Beating the odds: extreme long-term survival with glioblastoma. *Neuro. Oncol.* 16, 1159–1160. doi: 10.1093/neuonc/nou166

Boyarskikh, U., Pintus, S., Mandrik, N., Stelmashenko, D., Kiselev, I., Evshin, I., et al. (2018). Computational master-regulator search reveals mTOR and PI3K pathways responsible for low sensitivity of NCI-H292 and A427 lung cancer cell lines to cytotoxic action of p53 activator Nutlin-3. *BMC Med. Genomics* 11(Suppl. 1):12. doi: 10.1186/s12920-018-0330-5

Bradshaw, A., Wickremsekera, A., Tan, S. T., Peng, L., Davis, P. F., and Itinteang, T. (2016). Cancer stem cell hierarchy in glioblastoma multiforme. *Front. Surg.* 3:21. doi: 10.3389/fsurg.2016.00021

Cai, J., Chen, Q., Cui, Y., Dong, J., Chen, M., Wu, P., et al. (2018). Immune heterogeneity and clinicopathologic characterization of IGFBP2 in 2447 glioma samples. *Oncoimmunology* 7:e1426516. doi: 10.1080/2162402X.2018.1426516

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14:128. doi: 10.1186/1471-2105-14-128

Chen, J. R., Yao, Y., Xu, H. Z., and Qin, Z. Y. (2016). Isocitrate dehydrogenase (IDH)1/2 mutations as prognostic markers in patients with glioblastomas. *Med. (United States)* 95:e2583. doi: 10.1097/MD.0000000000002583

da Hora, C. C., Pinkham, K., Carvalho, L., Zinter, M., Tabet, E., Nakano, I., et al. (2019). Sustained NF-κB-STAT3 signaling promotes resistance to Smac mimetics in glioma stem-like cells but creates a vulnerability to EZH2 inhibition. *Cell Death Discov.* 5:72. doi: 10.1038/s41420-019-0155-9

Das, P., Puri, T., Jha, P., Pathak, P., Joshi, N., Suri, V., et al. (2011). A clinicopathological and molecular analysis of glioblastoma multiforme with long-term survival. *J. Clin. Neurosci.* 18, 66–70. doi: 10.1016/j.jocn.2010.04.050

De Vega, S., Iwamoto, T., and Yamada, Y. (2009). Fibulins: multiple roles in matrix structures and tissue functions. *Cell. Mol. Life Sci.* 66, 1890–1902. doi: 10.1007/s00018-009-8632-6

Debinski, W., and Gibo, D. M. (2005). Fos-related antigen 1 modulates malignant features of glioma cells. *Mol. Cancer Res.* 3, 237–249. doi: 10.1158/1541-7786.MCR-05-0004

Debinski, W., Slagle-Webb, B., Achen, M. G., Stacker, S. A., Tulchinsky, E., Gillespie, G. Y., et al. (2001). VEGF-D is an X-linked/AP-1 regulated putative onco-angiogen in human glioblastoma multiforme. *Mol. Med.* 7, 598–608. doi: 10.1007/bf03401866

Ellis, H. P., and Kurian, K. M. (2014). Biological rationale for the use of PPARÎ³ agonists in glioblastoma. *Front. Oncol.* 4:52. doi: 10.3389/fonc.2014.00052

Franceschi, S., Mazzanti, C. M., Lessi, F., Aretini, P., Carbone, F. G., La Ferla, M., et al. (2015). Investigating molecular alterations to profile short- and long-term recurrence-free survival in patients with primary glioblastoma. *Oncol. Lett.* 10:3599–3606. doi: 10.3892/ol.2015.3738

Fujioka, S., Niu, J., Schmidt, C., Sclabas, G. M., Peng, B., Uwagawa, T., et al. (2004). NF-κB and AP-1 connection: mechanism of NF-κB-dependent regulation of AP-1 activity. *Mol. Cell. Biol.* 24, 7806–7819. doi: 10.1128/mcb.24.17.7806-7819.2004

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy–analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi: 10.1093/bioinformatics/btg405

Gravendeel, L. A. M., Kouwenhoven, M. C. M., Gevaert, O., De Rooi, J. J., Stubbs, A. P., Duijm, J. E., et al. (2009). Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res.* 69, 9065–9072. doi: 10.1158/0008-5472.CAN-09-2307

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., et al. (2016). Toward a shared vision for cancer genomic data. *N. Engl. J. Med.* 375, 1109–1112. doi: 10.1056/nejmp1607591

Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J. C., Fine, H., and Madhavan, S. (2018). Data descriptor: the REMBRANDT study, a large collection of genomic data from brain cancer patients. *Sci. Data* 5:180158. doi: 10.1038/sdata.2018.158

Han, S., Meng, L., Han, S., Wang, Y., and Wu, A. (2014). Plasma IGFBP-2 levels after postoperative combined radiotherapy and chemotherapy predict prognosis in elderly glioblastoma patients. *PLoS One* 9:e93791. doi: 10.1371/journal.pone.0093791

Holmes, K. M. (2012). *Elucidating the IGFBP2 Signaling Pathway in Glioma Development Elucidating the IGFBP2 Signaling Pathway in Glioma Development and Progression and Progression*. Available online at: https://digitalcommons.library.tmc.edu/utgsbs_dissertations (accessed September 5, 2020).

Iwadate, Y. (2016). Epithelial-mesenchymal transition in glioblastoma progression. *Oncol. Lett.* 11, 1615–1620. doi: 10.3892/ol.2016.4113

Jahani-As, A., Yin, H., Soleimani, V. D., Haque, T., Luchman, H. A., Chang, N. C., et al. (2016). Control of glioblastoma tumorigenesis by feed-forward cytokine signaling. *Nat. Neurosci.* 19, 798–806. doi: 10.1038/nn.4295

Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., et al. (2020). The reactome pathway knowledgebase. *Nucleic Acids Res.* 48, D498–D503. doi: 10.1093/nar/gkz1031

Kamal, M. M., Sathyan, P., Singh, S. K., Zinn, P. O., Marisetty, A. L., Liang, S., et al. (2012). REST regulates oncogenic properties of glioblastoma stem cells. *Stem Cells* 30, 405–414. doi: 10.1002/stem.1020

Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., and Tanabe, M. (2020). KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* 49, D545–D551. doi: 10.1093/nar/gkaa970

Kel-Margoulis, O. V., Kel, A. E., Reuter, I., Deineko, I. V., and Wingender, E. (2002). TRANSCompel§: a database on composite regulatory elements in eukaryotic genes. *Nucleic Acids Res.* 30, 332–334. doi: 10.1093/nar/30.1.332

Kel, A., Boyarskikh, U., Stegmaier, P., Leskov, L. S., Sokolov, A. V., Yevshin, I., et al. (2019). Walking pathways with positive feedback loops reveal DNA methylation biomarkers of colorectal cancer. *BMC Bioinformatics* 20:119. doi: 10.1186/s12859-019-2687-7

Kel, A. E., Gößling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., and Wingender, E. (2003). MATCH^{TM}: a tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 31, 3576–3579. doi: 10.1093/nar/gkg585

Kel, A., Voss, N., Jauregui, R., Kel-Margoulis, O., and Wingender, E. (2006). Beyond microarrays: finding key transcription factors controlling signal transduction pathways. *BMC Bioinformatics* 7:S13. doi: 10.1186/1471-2105-7-S2-S13

Kesari, S., and Bota, D. A. (2011). Fos-related antigen-1 (Fra-1) is a regulator of glioma cell malignant phenotype. *Cancer Biol. Ther.* 11, 307–310. doi: 10.4161/cbt.11.3.14718

Kimura, R., Ishikawa, C., Rokkaku, T., Janknecht, R., and Mori, N. (2011). Phosphorylated c-Jun and Fra-1 induce matrix metalloproteinase-1 and thereby regulate invasion activity of 143B osteosarcoma cells. *Biochim. Biophys. Acta Mol. Cell Res.* 1813, 1543–1553. doi: 10.1016/j.bbamcr.2011.04.008

Kolmykov, S., Yevshin, I., Kulyashov, M., Sharipov, R., Kondrakhin, Y., Makeev, V. J., et al. (2021). Gtrd: an integrated view of transcription regulation. *Nucleic Acids Res.* 49, D104–D111. doi: 10.1093/nar/gkaa1057

Kolpakov, F., Akberdin, I., Kashapov, T., Kiselev, L., Kolmykov, S., Kondrakhin, Y., et al. (2019). BioUML: an integrated environment for systems biology and collaborative analysis of biomedical data. *Nucleic Acids Res.* 47, W225–W233. doi: 10.1093/nar/gkz440

Koschmann, J., Bhar, A., Stegmaier, P., Kel, A., and Wingender, E. (2015). "Upstream Analysis": an integrated promoter-pathway analysis approach to causal interpretation of microarray data. *Microarrays* 4, 270–286. doi: 10.3390/microarrays4020270

Krex, D., Klink, B., Hartmann, C., von Deimling, A., Pietsch, T., Simon, M., et al. (2007). Long-term survival with glioblastoma multiforme. *Brain* 130, 2596–2606. doi: 10.1093/brain/awm204

Krull, M., Voss, N., Choi, C., Pistor, S., Potapov, A., and Wingender, E. (2003). TRANSPATH§: an integrated database on signal transduction and a tool for array analysis. *Nucleic Acids Res.* 31, 97–100. doi: 10.1093/nar/gkg089

Lee, Y., Scheck, A. C., Cloughesy, T. F., Lai, A., Dong, J., Farooqi, H. K., et al. (2008). Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Med. Genomics* 1:52. doi: 10.1186/1755-8794-1-52

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi: 10.1093/bioinformatics/bts034

Liao, Y., Smyth, G. K., and Shi, W. (2013). The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.* 41:e108. doi: 10.1093/nar/gkt214

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi: 10.1093/bioinformatics/btr260

Lindström, M. S. (2019). Expanding the scope of candidate prognostic marker IGFBP2 in glioblastoma. *Biosci. Rep.* 39:BSR20190770. doi: 10.1042/BSR20190770

Liu, Y., Song, C., Shen, F., Zhang, J., and Song, S. W. (2019). IGFBP2 promotes immunosuppression associated with its mesenchymal induction and FcγRIIB phosphorylation in glioblastoma. *PLoS One* 14:e0222999. doi: 10.1371/journal.pone.0222999

Majdalawieh, A. F., Massri, M., and Ro, H. S. (2020). AEBP1 is a novel oncogene: mechanisms of action and signaling pathways. *J. Oncol.* 2020:8097872. doi: 10.1155/2020/8097872

Marbach, D., Costello, J. C., Küffner, R., Vega, N. M., Prill, R. J., Camacho, D. M., et al. (2012). Wisdom of crowds for robust gene network inference. *Nat. Methods* 9, 796–804. doi: 10.1038/nmeth.2016

Martinho, O., and Reis, R. M. (2011). "Malignant gliomas: role of platelet-derived growth factor receptor A (PDGFRA)," in *Tumors of the Central Nervous System*, ed. M. Hayat (Dordrecht: Springer), 109–118. doi: 10.1007/978-94-007-0344-5_12

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., et al. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* 34, D108–D110. doi: 10.1093/nar/gkj143

McDonald, K. L., O'Sullivan, M. G., Parkinson, J. F., Shaw, J. M., Payne, C. A., Brewer, J. M., et al. (2007). IQGAP1 and IGFBP2. *J. Neuropathol. Exp. Neurol.* 66, 405–417. doi: 10.1097/nen.0b013e31804567d7

Phillips, L. M., Zhou, X., Cogdell, D. E., Chua, C. Y., Huisinga, A., Hess, K. R., et al. (2016). Glioma progression is mediated by an addiction to aberrant IGFBP2 expression and can be blocked using anti-IGFBP2 strategies. *J. Pathol.* 239, 355–364. doi: 10.1002/path.4734

Prasad, B., Tian, Y., and Li, X. (2020). Large-scale analysis reveals gene signature for survival prediction in primary glioblastoma. *Mol. Neurobiol.* 57, 5235–5246. doi: 10.1007/s12035-020-02088-w

Prywes, R., and Henckels, E. (2013). Fra-1 regulation of Matrix Metallopeptidase-1 (MMP-1) in metastatic variants of MDA-MB-231 breast cancer cells. *F1000Research* 2:229. doi: 10.12688/f1000research.2-229.v1

Reifenberger, G., Weber, R. G., Riehmer, V., Kaulich, K., Willscher, E., Wirth, H., et al. (2014). Molecular characterization of long-term survivors of glioblastoma using genome- and transcriptome-wide profiling. *Int. J. Cancer* 135, 1822–1831. doi: 10.1002/ijc.28836

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43:e47. doi: 10.1093/nar/gkv007

Salaroglio, I. C., Mungo, E., Gazzano, E., Kopecka, J., and Riganti, C. (2019). ERK is a pivotal player of chemo-immune-resistance in cancer. *Int. J. Mol. Sci.* 20:2505. doi: 10.3390/ijms20102505

Schütt, B. S., Langkamp, M., Rauschnabel, U., Ranke, M. B., and Elmlinger, M. W. (2004). Integrin-mediated action of insulin-like factor binding protein-2 in tumor cells. *J. Mol. Endocrinol.* 32, 859–868. doi: 10.1677/jme.0.0320859

Scott, J. N., Rewcastle, N. B., Brasher, P. M. A., Fulton, D., MacKinnon, J. A., Hamilton, M., et al. (1999). Which glioblastoma multiforme patient will become a long-term survivor? A population-based study. *Ann. Neurol* 46, 183–188. doi: 10.1002/1531-8249

Shinawi, T., Hill, V. K., Krex, D., Schackert, G., Gentle, D., Morris, M. R., et al. (2013). DNA methylation profiles of long- and short-term glioblastoma survivors. *Epigenetics* 8, 149–156. doi: 10.4161/epi.23398

Simpson, A., Petnga, W., Macaulay, V. M., Weyer-Czernilofsky, U., and Bogenrieder, T. (2017). Insulin-like growth factor (IGF) pathway targeting in cancer: role of the IGF axis and opportunities for future combination studies. *Target. Oncol.* 12, 571–597. doi: 10.1007/s11523-017-0514-5

Sonoda, Y., Kumabe, T., Watanabe, M., Nakazato, Y., Inoue, T., Kanamori, M., et al. (2009). Long-term survivors of glioblastoma: clinical features and molecular analysis. *Acta Neurochir. (Wien).* 151, 1349–1358. doi: 10.1007/s00701-009-0387-1

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Thomas, P. D., Thomas, P. D., Campbell, M. J., Kejariwal, A., Al, E., Thomas, P. D., et al. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.* 13, 2129–2141. doi: 10.1101/gr.772403

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., et al. (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. 17, 98–110. doi: 10.1016/j.ccr.2009.12.020

Vial, E., and Marshall, C. J. (2003). Elevated ERK-MAP kinase activity protects the FOS family member FRA-1 against proteasomal degradation in colon carcinoma cells. *J. Cell Sci.* 116, 4957–4963. doi: 10.1242/jcs.00812

Waleev, T., Shtokalo, D., Konovalova, T., Voss, N., Cheremushkin, E., Stegmaier, P., et al. (2006). Composite module analyst: identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res.* 34:W541. doi: 10.1093/nar/gkl342

Wang, H., Wang, H., Shen, W., Huang, H., Hu, L., Ramdas, L., et al. (2003). *Insulin-like Growth Factor Binding Protein 2 Enhances Glioblastoma Invasion by Activating Invasion-enhancing Genes 1*. Available online at: www.mdanderson.org/genomics (accessed January 13, 2021).

Wen, P. Y., and Kesari, S. (2008). Malignant gliomas in adults. *N. Engl. J. Med.* 359, 492–507. doi: 10.1056/NEJMra0708126

Wingender, E., Dietze, P., Karas, H., and Knüppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res.* 24, 238–241. doi: 10.1093/nar/24.1.238

Wingender, E., Hogan, J., Schacherer, F., Potapov, A. P., and Kel-Margoulis, O. (2007). Integrating pathway data for systems pathology. *In Silico Biol.* 7(2 Suppl), S17–S25.

Xu, C., Wu, X., and Zhu, J. (2013). VEGF promotes proliferation of human glioblastoma multiforme stem-like cells through VEGF receptor 2. *Sci. World J.* 2013:417413. doi: 10.1155/2013/417413

Yamini, B. (2018). NF-κB, Mesenchymal differentiation and glioblastoma. *Cells* 7:125. doi: 10.3390/cells7090125

Yao, X., Sun, S., Zhou, X., Guo, W., and Zhang, L. (2016). IGF-binding protein 2 is a candidate target of therapeutic potential in cancer. *Tumor Biol.* 37, 1451–1459. doi: 10.1007/s13277-015-4561-1

Yau, S. W., Azar, W. J., Sabin, M. A., Werther, G. A., and Russo, V. C. (2015). IGFBP-2 - taking the lead in growth, metabolism and cancer. *J. Cell Commun. Signal.* 9, 125–142. doi: 10.1007/s12079-015-0261-2

Yu, G., Wang, L.-G., Yan, G.-R., and He, Q.-Y. (2015). DOSE: an R/Bioconductor package for disease ontology semantic and enrichment analysis. *Bioinformatics* 31, 608–609. doi: 10.1093/bioinformatics/btu684

Zhang, X., Wu, J., Luo, S., Lechler, T., and Zhang, J. Y. (2016). FRA1 promotes squamous cell carcinoma growth and metastasis through distinct AKT and c-Jun dependent mechanisms. *Oncotarget* 7:34371–34383. doi: 10.18632/oncotarget.9110

Zhang, X., Zhang, W., Cao, W. D., Cheng, G., and Zhang, Y. Q. (2012). Glioblastoma multiforme: molecular characterization and current treatment strategy (Review). *Exp. Ther. Med.* 3, 9–14.

Conflict of Interest: MK, DW, and TB are from the Department of Medical Bioinformatics, University Medical Center Göttingen. MK, AK, and EW are employees of geneXplain GmbH.

The remaining author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

I am among those who think that science has great beauty.

Marie Curie

4

Machine Learning based Survival Group Prediction in Glioblastoma

This work is published in preprints in February, 2022

Kalya, M.; Kel, A.; Leha, A.; Altynbekova, K.; Wingender, E.; beissbarth, T. Machine Learning based Survival Group Prediction in Glioblastoma . Preprints 2022, 202202005 1 (doi: 10.20944/preprints202202.005 1.v1) (link to the preprints) [*]

Availability of software, data and materials

The datasets analyzed in the current study, supplementary files plots and link to webtool are available in the GitHub project here - Machine Learning based Survival Group Prediction in Glioblastoma [†]

[*] https://www.preprints.org/manuscript/202202.0051/v1

[†] https://github.com/genexplain/Manasa_KP_et_al_MLmodels_predictionofGBMsurvivorgroups

Declaration of my contributions

I conceptualized and executed this work and Dr. Alexander Kel participated in finalising pipeline for data analysis in the direction of successful publication. Dr. Andreas Leha has participated in finalising the Machine Learning methods. Kamilya Altynbekova has extensively participated in making and maintainace of webtool aspects. Prof. Edgar Wingender has participated extensively in improving the manuscript and Prof. Tim Beißbarth has supervised the work, read and corrected the manuscript, and approved the final version.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

Machine Learning based Survival Group Prediction in Glioblastoma

Manasa Kalya [1,2], Alexander Kel [2,4 *], Andreas Leha[3], Kamilya Altynbekova[2], Edgar Wingender[2], Tim Beißbarth[1]

1. Department of Medical Bioinformatics, University Medical Center Göttingen, 37099 Göttingen, Germany
2. geneXplain GmbH, 38302 Wolfenbüttel, Germany
3. Institute for Medical Statistics, University Medical Center Göttingen, 37099 Göttingen, Germany
4. Institute of Chemical Biology and Fundamental Medicine SBRAS, 630090, Novosibirsk, Russia

* Correspondence: Alexander E. Kel: alexander.kel@genexplain.com

Keywords: Glioblastoma, survival prediction, Machine Learning, biomarkers, HumanPSD™, Long-term survivors.

Abstract:

Glioblastoma (GBM) is a very aggressive malignant brain tumor with the vast majority of patients surviving less than 12 months (Short-term survivors [STS]). Only around 2% of patients survive more than 36 months (Long-term survivors [LTS]). Studying these extreme survival groups might help in better understanding GBM biology. This work aims at exploring application of machine learning methods in predicting survival groups(STS, LTS). We used age and gene expression profiles belonging to 249 samples from publicly available datasets. 10 Machine learning methods have been implemented and compared for their performances. Hyperparameter tuned random forest model performed best with accuracy of 80% (AUC of 74% and F1_score of 85%). The performance of this model is validated on external test data of 16 samples. The model predicted the true survival group for 15 samples achieving an accuracy of 93.75%. This classification model is deployed as a web tool GlioSurvML. The top 1500 features which retained classification efficiency (Accuracy of 80%, AUC of 74%) were studied for enriched pathways and disease-causal biomarker associations using the HumanPSD™ database. We identified 199 genes as possible biomarkers of GBM and/or similar diseases (like Glioma, astrocytoma, and others). 57 of these genes are shown to be differentially expressed across survival groups and/or have impact on survival. This work demonstrates the application of machine learning methods in predicting survival groups of GBM.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

1. Introduction

The majority of patients with glioblastoma (GBM) have a short-term survival rate of fewer than 12 months (short-term survivors [STS]), however there is a minority of individuals who have a long-term survival rate of more than three years (36 months), referred to as long-term survivors (LTS)(Hwang et al., 2019a). Clinical, radiological, and histological characteristics have not been found to be predictors of long-term survival or response to therapy in studies (Davis, 2016). (Hwang et al., 2019) Machine Learning (ML) techniques are increasingly being applied in GBM research, as evidenced by a rise in the number of publications in the recent decade (Valdebenito and Medina, 2019). With enormous volumes of high-dimensional data, machine learning aids in recognizing patterns, forecasting events, and interpreting the interactions of complex biochemical networks (Valdebenito and Medina, 2019).

A biomarker is a biological marker that indicates a biological condition and can signal illness-associated molecular alterations at the molecular level which is valuable in understanding the disease state or diagnosis. ML based classification and feature selection methods have aided such a biomarker discovery (Mamoshina et al., 2018; Torres and Judson-Torres, 2019; Fortino et al., 2020; Xie et al., 2021). Some of the major examples of ML use in GBM research are the Stemness Subtype(I/II) Predictor (Wang et al., 2021), NF1 activation status predictor , GBM subtype-specific classifiers (Ensenyat-Mendez et al., 2021), and temozolomide treatment response predictor(Geldof et al., 2020). (Senders et al., 2020)Joeky et al.,2020 has developed an online survival calculator for patients with glioblastoma based on demographic, socioeconomic, clinical, and radiographic variables to predict overall survival.

Transcriptomics approaches have been demonstrated to be highly promising as they offer prognostic techniques for gaining a better knowledge of the condition. Using TCGA RNA-seq data from 129 samples, a study has used an Autoencoder (AE)-based approach for the prediction of GBM patient survival (short-term or long-term survivors) with an accuracy 89%.(Kirtania et al., 2021) In this study, we evaluated 10 ML models to build a classifier which can classify GBM patients into short-term and long-term survivo groups using transcriptomic profiles and clinical information(age) of 249 patients, pooled from 5 publicly available datasets. Random forest model has performed best with an accuracy of 80% and is deployed as a webtool - GlioSurvML. Following model identification, the top 1500 features are used for further analysis to identify important biological pathways and biomarkers.

2. Materials and Methods

2.1. Data Collection

The genome-wide expression profiles based on the Human Genome U133 Plus 2.0 array and **clinical** information of patients with GBM were collected from the public repository of the GEO database. Age information was available for 75.5% of the samples, whereas information on Gender, Karnofsky score, MGMT status, or IDH status were not available for most of them (<30%) and hence only information of age is considered along with the transcriptome to build the survival predictor.

All the datasets were pooled together leading to 176 and 73 samples corresponding to short-term survivors (STS; survival < 12 months) and long-term survivors (LTS; survival > 36 months), respectively (**Table 1).** Duplicates were not removed. Raw data, sample information, and cleaned datasets are given in **Supplementary file 1.**

Table 1. Statistics of datasets studied in this work.

	Platform	Short-term survivors	Long-term survivors
GSE53733 (Reifenberger et al., 2014)	HU133 plus 2.0 arrays	16	23
GSE108474 (Gusev et al., 2018)	HU133 plus 2.0 arrays	97	35
GSE13041 (Lee et al., 2008)	HU133 plus 2.0 arrays	20	02
GSE7696 (Murat et al., 2008)	HU133 plus 2.0 arrays	29	09
GSE43378 (Kawaguchi et al., 2013)	HU133 plus 2.0 arrays	14	04

2.2. Affymetrix microarray data pre-processing

The raw data files (. CEL format) for the above-mentioned datasets were collected from the GEO database- from here on called as GSE dataset. RMA algorithm is used in R (affy package) for background correction, quality check, and normalization to obtain log2 transformed expression values (Gautier et al., 2004). Batch correction of the pooled expression data was performed using empirical Bayes framework is performed (Leek et al., 2012). PCA plot for the batch corrected data is given in **Supplementary file 2.** This batch corrected file is used for further analysis. Multiple Affymetrix ids were summarized to genes ids by choosing the maximum out of probe intensities of multiple probes belonging to a single gene. The final expression matrix comprised 21526 probes and 249 samples is given in **Table S1-C**.

2.3 Development of a Prediction Model Using a Machine Learning Algorithm

To develop a machine learning model, we have used several functionalities of model building in python sklearn (Pedregosa FABIANPEDREGOSA et al., 2011). The dataset used to build the model contains transcriptomics profiles of 176 STS and 73 LTS and the age of the corresponding patient. Using a variance filter the top 10,000 highly variant genes are identified and were considered for model building. Labels were encoded using label encoder. Figure 1 shows the work flow of model development.The samples were first split into 80% training and 20% test data. All the downstream operations to build the predictive model were performed only on training data and is later tested on test data. The training data is scaled and quantile transformed. The scaling and quantiles were saved so that they can be applied to test data.

To deal with the problem of class imbalance during model training (training - STS:139, LTS=60), we have used the synthetic minority oversampling technique SMOTE of the imblearn package (Lemaˆıtre Lemaˆıtre et al., 2017). This oversampling strategy first randomly selects an instance from the minority class and finds its k nearest minority class neighbors. Synthetic data would then be made between the random data and the randomly selected k-nearest neighbor. With SMOTE oversampling, the number of samples in the minority class was increased to 139. On this resampled training data, we applied 10 ML models. However, only the random forest model performed better in terms of classifying the minority classes. For hyperparameter tuning of model parameters we used GridSearchCV. Models

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2022 doi:10.20944/preprints202202.0051.v1

were tuned for their hyperparameters (Table 2) for optimal performances. Hyperparameter tuning results for all ML models are given in **Table S3-A.**

Table 2. Hyperparameter tuning in ML models

Method	Parameters
Random forest	Criterion, max_depth, n_estimators
Logistic regression	penalty, Solver & C
Linear Support Vector Classification (Linear SVC)	C, kernel, gamma,
Support Vector Classification (SVC)	Kernel, C, gamma
Nu-Support Vector Classification (NuSVC)	Nu, Kernel, decision_function_shape
Naïve Bayes	var_smoothing
Classification and Regression Trees (CART)	Criterion, max_features
k-nearest neighbors (KNN)	N_neighbors,algorithm & weights
Balanced random forest	max_features, n_estimators, max_depth, criterion
Balanced Bagging	n_estimators

Hyperparameter tuned models were applied on the (20%) test data to evaluate model performances and choose the best performing classifier. The best performing model was evaluated on an external independent microarray data to evaluate the application of this classifier as a reliable tool for predicting Glioblastoma survival groups. The top best features based which retains higher classification efficiency were extracted and evaluated for biological relevance by using Gene set enrichment, Differential expression, Survival significance and their association with Glioblastoma or similar diseases.

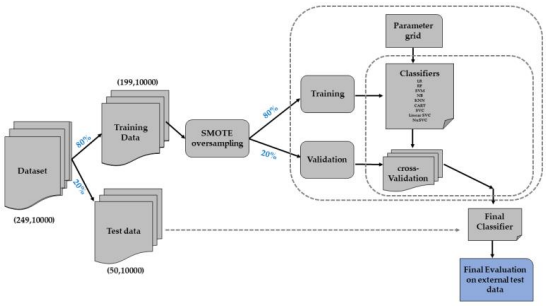

Figure 1. Workflow explaining the steps of building ML models.

2.4.Gene Enrichment Analysis

To explore the biological importance of these 1500features, gene list enrichment tool enrichR (Chen et al., 2013) is used. Enrichment for Molecular Signature Database (MSigDB) (Liberzon et al., 2011) is used.

2.5 Differential gene expression (DEG) analysis

LIMMA (Linear Models for Microarray Data) method was applied to identify differentially expressed genes (Ritchie et al., 2015) . Differential gene expression analysis for short-term and long-

term survivors is performed in GSE108474 and TCGA GBM microarray data. Clinical information and cleaned datasets of GSE108474 and TCGA GBM microarray data are given in **Supplementary 4.**

2.6. Impact on survival

Survival and Survminer libraries in R are used to perform univariate survival analysis. Univariate Cox regression for survival analysis is performed using the coxph function of the Survival package to calculate the Hazard ratio (HR) with p-value cutoff of 0.05 for significance (Therneau, 2021). KMplots are used to depict impact of genes on survival with non-overlapping 50% upper and lower quantiles. **supplementary 4**

2.7 Identification of biomarkers

Causal molecular mechanisms present a unifying principle for disease classification, analysis of clinical disorder associations, as well as prediction of disease genes, diagnostic markers, and therapeutic targets. A novel approach published (Stegmaier et al., 2010) built of 1000 causal gene-disease networks is now updated and available in the HumanPSD™ database (Wingender et al., 2007). The important features identified using the ML model can serve as biomarkers of survival/prognosis in GBM. HumanPSD™ database 2021.2 is mined to fetch information on the association of these features with GBM or similar diseases

3. Results

3.1. Development of ML model:

The genome-wide expression profiles from 5 independent experiments using Human Genome U133 Plus 2.0 arrays with corresponding clinical information of Glioblastoma patients were collected, normalized and integrated to obtain a data matrix of 176 and 73 samples corresponding to short-term survivors (STS; survival < 12 months) and long-term survivors (LTS; survival > 36 months), respectively. Top 10k highly variant genes were used for building ML models for classification. See more details in methods section (**Supplementary file 1 and 2)**

In the current work, we have used machine learning methods to predict the survival class of GBM patients using gene expression profiles.

Ten ML models such as random forest, Naïve Bayes, Support Vector Classification, Linear SVC, NuSVC, Logistic Regression, Classification and Regression Trees (CART), k-nearest neighbors (KNN), and specialized packages of imbalanced learning like Balanced Random forest and Balanced Bagging are evaluated in this study. The dataset was split into 80% training and 20% test data. To address the problem of class imbalance, SMOTE oversampling is applied during the training of the model to balance the classes. GridSearchCV upon StratifiedShuffleSplit on the oversampled training data is used for hyperparameter tuning of the models (Table S3-A and Table S3-B). The performance of all the hyperparameter tuned models on the test data is given in **Table 1**.

We found that hyperparameter-tuned random forest model (**Figure 2)** performed best out of all other models mentioned earlier, with f1_score of 86.48%, Accuracy of 80%, and AUC of 74% on test data. This corresponds to 86% of true labels in majority class and 62% true labels in minority class (**Figure 3A)**

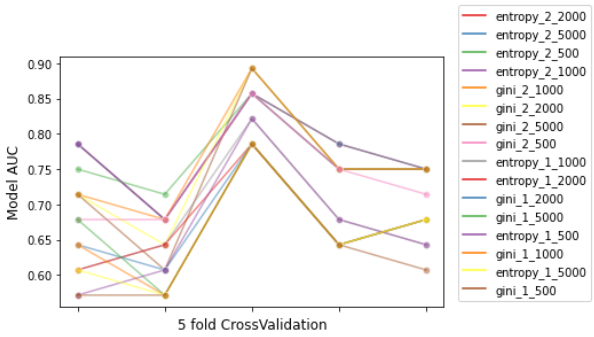

Figure 2. Hyperparameter Tuning in RF. The following hyperparameters were tuned: Tuning parameters of criterion(gini/entropy), maximum depth (1/2) and number of estimators (500/1000/2000/5000) for random forest model upon 5-fold cross validation using GridSearchCV.

The hyperparameter tuned BalancedRandomForest model performed with f1_score of 82.3%, Accuracy of 76%, AUC of 76.29% on test data. The model positively identified 77% of minority labels and 78% of majority labels (**Figure 3B)**. The linear models like LR, SVC, NuSVC, LinearSVC had lower AUC values as they identified less than 35% of the minority class (LTS) and hence were not considered in our further analysis.

Table 3. Performance of 10 ML models under study on 20% test data upon hyperparameter tuning

Hyperparameter tuned ML model	F1_Score	Accuracy	AUC
Logistic Regression	0.81	0.720	0.636
Random forest	0.864	0.800	0.740
NuSVC	0.864	0.780	0.626
SVC	0.864	0.787	0.626
Balanced random forest	0.823	0.760	0.762
Balanced Bagging	0.853	0.780	0.701
Linear SVC	0.746	0.660	0.645
Naïve Bayes	0.805	0.720	0.661
KNN	0.407	0.360	0.417
CART- Decision Trees	0.788	0.700	0.647

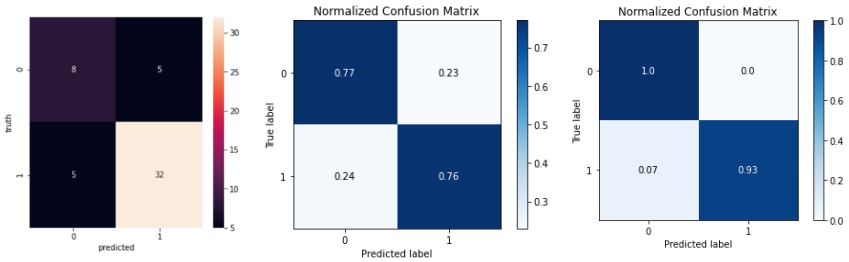

Figure3. Normalized Confusion Matrix for ML models.
Normalized Confusion matrix for the classification of survival groupsis shown here.For the classes, 0(LTS) and 1(STS), the X-axis in the plot is for the predicted class and the Y-axis is for the true class. The true class elements of a row are spread across columns and the elements of the matrix are normalized row wise, i.e., sum of fractions along a row sum to 1. The only true predictions are along the diagonal, i.e., each of the i–ith element of the matrix and all other off-diagonal elements along a row are wrong predictions. The more the correctness of a class, the darker the blue hue it has in a cell of the plot of the confusion matrix. A) Normalized Confusion Matrix of Random forest model on internal (20%) test data B) Normalized Confusion Matrix for BalancedRandom forest model without oversampling C) Normalized Confusion Matrix for Random forest model on external test data

To build a robust machine learning model which can identify the survival class of the GBM patients, we tested the random forest model on an external microarray dataset (**Supplementary file 7)**. The LTS are rare events and hard to find adequate samples for testing. The external dataset containing 16 samples (1-LTS and 15-STS) was from a single experiment. Random forest model performed with an accuracy of 93.75% (AUC of 96.66%) (**Figure 3C).**

Age was found to be one of the top important (Top 7) features of the random forest model developed. The random forest model built on gene-expression and age had better sensitivity (93.75%) than the random forest model built on gene expression alone (81.25%) **(Supplementary file 7).**

3.2. Deployment of ML model:

The random forest model developed here for survival class prediction is deployed as a webtool-GlioSurvML. All information associated is given in github repository. Webtool has 2 models of RF one with including age and one without age. The webtool prints the output as a PDF report as well as an excel-table. (**Supplementary file 8)**

3.3. Feature Importances

Ranking of features/genes according to their importance in the random forest classification model discussed above is given in **Table S3-C**. The performance of the model using top 100/500/1000/1500/2000 features (**Table S3-D**) is investigated. We observed that the top 1500 features (**Table S3-E**) were sufficient enough to maintain the 80% accuracy of prediction. These genes are looked for their relevance in the disease using gene enrichment analysis, differential expression analysis, univariate survival analysis to investigate prognostic value and by utilizing existing knowledge on biomarkers of the glioblastoma.

We found that TNF-alpha Signaling via NF-kB, mTORsignalling, G2-M checkpoints, Epithelial to Mesenchymal transition are some of the top overlapping gene sets according to MsigDB **Table S3-F.**

3.4 Biomarker Identification

Exploiting the previously reported method on unifying disease mechanisms based on causal gene-disease associations as described in HumanPSD™ database (**Supplementary file 5**), we find that, out of top 1500 genes, 63 known gene expression biomarkers of Glioblasytoma and 136 gene expression biomarkers from similar diseases to Glioblastoma and 35 markers were reported both in Glioblastoma and in one of the similar diseases according to HumanPSD™ database (199 unique biomarkers in total). **Figure 4.** Based on this analysis, we propose 171(136+35) gene expression based biomarkers to Glioblastoma. According to the database, these genes were mapped to 8 diseases like Osteosarcoma, Melonoma, Ovarian neoplasm, Nasopharangeal neoplasm including Glioma , astrocytoma, brain

neoplasms. Top 10 (based on feature ranking in random forest model) of these new proposed biomarkers of Glioblastoma prognoisis are given in **Table 4**.

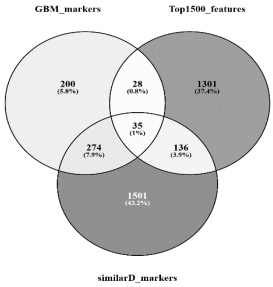

Figure 4: Venn Diagram of HumanPSD™ biomarkers and important features.
HumanPSD™ database reports 537 mRNAexpression based Glioblastoma markers, 1946 mRNA expression based biomarkers of diseases similar to GBM. Out of the top 1500 important features required for classifying the survival group of GBM, 63 Glioblastoma and 171 similar disease biomarkers were found overlapping. 35 genes were found associated with both GBM and related disease.

These biomarkers are checked for differential gene expression between STS and LTS and univariate impact on survival. The analysis is performed in GSE108474 dataset which is U133 plus 2 affymetrix platform and TCGA-GBM of 560 microarray (U133 Affy array) datasets (**Supplementary file 4).**

Table 4. Top 10 features proposed as biomarkers of prognosis in Glioblastoma in our study

Features	Feature_Rank	Molecule	Disease	Disease_Association	PMID
CBX3	25	mRNA	Osteosarcoma	increased expression of CBX3 mRNA correlates with increased neoplasm metastasis associated with osteosarcoma	22870217
GHR	29	mRNA	Melanoma	increased expression of GHR mRNA correlates with neoplasm metastasis associated with melanoma	24134847
HNRNPA2B1	38	mRNA	Brain Neoplasms	increased expression of HNRNPA2B1 mRNA correlates with oligodendroglioma tumors associated with brain neoplasms	11485829
NES	41	mRNA	Astrocytoma	increased expression of NES mRNA may correlate with disease progression associated with astrocytoma	17611714
SKP2	44	mRNA	Ovarian Neoplasms	decreased expression of SKP2 mRNA may correlate with increased response to salinomycin associated with ovarian neoplasms	23807222
RARRES2	48	mRNA	Glioma	increased expression of RARRES2 mRNA correlates with glioma	21949124
ERBB2	58	mRNA	Ovarian Neoplasms	increased expression of ERBB2 mRNA may correlate with malignant form of ovarian neoplasms	8094034
ELAVL1	63	mRNA	Ovarian Neoplasms	decreased expression of ELAVL1 mRNA may prevent increased positive regulation of gene expression associated with ovarian neoplasms	23394580
TGIF2	68	mRNA	Ovarian Neoplasms	increased expression of TGIF2 mRNA correlates with ovarian neoplasms	11006116

FZD1	80	mRNA	Ovarian Neoplasms	increased expression of FZD1 mRNA correlates with glandular and epithelial neoplasms associated with ovarian neoplasms	19148501

The information of differential gene expression (Log2FC, adj.pvalue) and survival significance (Hazard Ratio and FDR <0.05) for these 199 biomarkers in GSE108474 are given in **Supplementary File 6**. Out of these, 17 genes were significantly differentially expressed, 28 had survival significance and 12 biomarkers were both differentially expressed and had significant impact on survival.

4. Discussion

In this study, we evaluated application of 10 ML models to build a classifier to differentiate patients between STS and LTS groups based on their transcriptomic profiles and clinical information(age) from 249 patients data which is pooled from publicly available datasets. To the best of our knowledge this is the first application of its kind. Of the models evaluated, a random forest model performed best with accuracy of 80% (F1_score=86.4% AUC =74%). Furthermore, this model is evaluated on external microarray data and found to have high accuracy of 93.75% (AUC of 96.66%). The identification of age as an important feature is in line with the observation that age is an important clinical predictor for survival.We have noted that the top 1500 features alone can preserve the classification efficiency of the model and these are only used for further analysis.

The enrichment analysis revealed enrichment of TNF-Alpha via NF-kB, mTOR signalling, G2-M checkpoints, Epithelial to Mesenchymal transition signaling pathways. All of these pathways are identified as therapeutic targets in GBM (ref) and play a role in response to Temozolomide (ref), which is a first line of treatment in GBM.

Using HumanPSD™ we have identified 8 disorders which are mapped to be similar to Glioblastoma. Of these three are related to central nervous system tumors and others include ovarian, osteosarcoma, melanoma, nasopharyngeal tumors and general neoplasms. This identified overlap of GBM with gliomas and melanoma is interesting as studies have shown increased risk of gliomas in malignant melanoma patients (Scarbrough et al., 2014) and increased representation of melanoma in GBM patients (Yang et al., 2021). The gliomas and melanoma are shown to be responsive to Temozolomide which is indicative of a common potential pathophysiological pathway (Desai and Grossman, 2008).

From the HumanPSD™ we have identified 199 mRNA biomarkers that have previously been linked to Glioblastoma and/or related.

Some of the important biomarkers include retinoic acid receptor responder 2(RERRES2), Distinct Subgroup of The Ras Family Member 3 (DIRAS3), DEP Domain Containing MTOR Interacting Protein (DEPTOR), Insulin like Growth Binding Protein 5(IGFBP5) and C-Type Lectin Domain Family 2 Member B (CLEC2B). RERRES2 is a critical gene of retinoic acid signaling which is reported to be highly upregulated in STS in GBM (Barbus et al., 2011). DIRAS3 drives autophagy by Ras/AKT/mTOR pathway in GBM and is reported to be significantly downregulated in long-term survivors of GBM (Zhong et al., 2019). DEPTOR is a natural inhibitor of MTORc1 and mTORc2 which plays an important role in autophagy. Inhibitors of mTOR signaling are widely discussed as an adjuvant therapy to regulate

autophagy in GBM (Xia et al., 2020). IGFBP5 promotes cell invasion by regulating Epithelial to Mesenchymal Transition and inhibits cell proliferation by suppressing the phosphorylation of AKT in GBM (Dong et al., 2020). Its expression was upregulated in high grades of glioma and is correlated with worse prognosis (Dong et al., 2020). CLEC2B - A rise in expression of CLEC2B was linked to a rise in the progression-free Hazard ratio (Serão et al., 2011)

Identifying the signaling pathways and biomarkers that are related to Glioblastoma, mapping to the diseases which are related to CNS or those with shared biology gives strength to our machine learning model and reinforces the idea that machine learning models can be used for understanding the biology of GBM. Our analysis has shown inclusion of clinical information i.e. age has increased the sensitivity of survival group prediction which shows the importance of adding clinical information to the machine learning models. Other clinically important variables are not added to the model due to high levels of missingness in the datasets which needs to be addressed while collecting the future data. One important limitation of the current study is that the method is applicable only for microarray platforms and extension of this model for application in RNA-seq data requires further work.

5. Conclusion

The current study presents a Machine Learning model for use in research to classify patients into Glioblastoma survival groups, deploys application as a webtool, discusses important features for relevance in the disease, proposes new plausible markers of survival in Glioblastoma.

Availability of software, data and materials: All the datasets analyzed in the current study are available from previous publications. All datasets, models and supplementary materials are available here:

https://github.com/genexplain/Manasa_KP_et_al_MLmodels_predictionofGBMsurvivorgroups

Conflicts of Interest: The authors Manasa Kalya, Andreas Leha and Tim Beißbarth are from Department of Medical Bioinformatics, University Medical Center Göttingen, Kamilya Altynbekova, Alexander Kel and Edgar Wingender are employees of geneXplain GmbH.

Author Contributions: AK is involved in conceptualization, providing resources, supervision and manuscript reviewing. MKP has conceptualized the work, performed data collection, data analysis, interpreting results and writing the manuscript. AL has extensively participated in developing the data analysis pipeline, KA has participated in making the webtool application. EW and TB are involved in supervision of work and in reviewing the draft.

Funding: This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 766069.

Acknowledgements: It is my pleasure to acknowledge Ravi Kumar Nadella who has read and given comments on every version of this paper and for his expertise, discussions and assistance in improving the manuscript.

References

Barbus, S., Tews, B., Karra, D., Hahn, M., Radlwimmer, B., Delhomme, N., et al. (2011). Differential retinoic acid signaling in tumors of long- and short-term glioblastoma survivors. *J. Natl. Cancer Inst.* 103, 598–601. doi:10.1093/JNCI/DJR036.

Chen, E. Y., Tan, C. M., Kou, Y., Duan, Q., Wang, Z., Meirelles, G. V., et al. (2013). Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinformatics* 14. doi:10.1186/1471-2105-14-128.

Davis, M. E. (2016). Glioblastoma: Overview of disease and treatment. *Clin. J. Oncol. Nurs.* 20, 1–8. doi:10.1188/16.CJON.S1.2-8.

Desai, A. S., and Grossman, S. A. (2008). Association of melanoma with glioblastoma multiforme. *https://doi.org/10.1200/jco.2008.26.15_suppl.2082* 26, 2082–2082. doi:10.1200/JCO.2008.26.15_SUPPL.2082.

Dong, C., Zhang, J., Fang, S., and Liu, F. (2020). IGFBP5 increases cell invasion and inhibits cell proliferation by EMT and Akt signaling pathway in Glioblastoma multiforme cells. *Cell Div.* 15, 1–9. doi:10.1186/S13008-020-00061-6/FIGURES/5.

Ensenyat-Mendez, M., Íñiguez-Muñoz, S., Sesé, B., and Marzese, D. M. (2021). iGlioSub: an integrative transcriptomic and epigenomic classifier for glioblastoma molecular subtypes. *BioData Min.* 14, 1–16. doi:10.1186/S13040-021-00273-8/FIGURES/5.

Fortino, V., Wisgrill, L., Werner, P., Suomela, S., Linder, N., Jalonen, E., et al. (2020). Machine-learning–driven biomarker discovery for the discrimination between allergic and irritant contact dermatitis. *Proc. Natl. Acad. Sci. U. S. A.* 117, 33474–33485. doi:10.1073/PNAS.2009192117.

Gautier, L., Cope, L., Bolstad, B. M., and Irizarry, R. A. (2004). affy--analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics* 20, 307–315. doi:10.1093/bioinformatics/btg405.

Geldof, T., van Damme, N., Huys, I., and van Dyck, W. (2020). Patient-level effectiveness prediction modeling for glioblastoma using classification trees. *Front. Pharmacol.* 10, 1665. doi:10.3389/FPHAR.2019.01665/BIBTEX.

Gusev, Y., Bhuvaneshwar, K., Song, L., Zenklusen, J. C., Fine, H., and Madhavan, S. (2018). The REMBRANDT study, a large collection of genomic data from brain cancer patients. *Sci. data* 5. doi:10.1038/SDATA.2018.158.

Hwang, T., Mathios, D., McDonald, K. L., Daris, I., Park, S. H., Burger, P. C., et al. (2019a). Integrative analysis of DNA methylation suggests down-regulation of oncogenic pathways and reduced somatic mutation rates in survival outliers of glioblastoma. *Acta Neuropathol. Commun.* 7, 5. doi:10.1186/S40478-019-0744-0.

Hwang, T., Mathios, D., McDonald, K. L., Daris, I., Park, S. H., Burger, P. C., et al. (2019b). Integrative analysis of DNA methylation suggests down-regulation of oncogenic pathways and reduced somatic mutation rates in survival outliers of glioblastoma. *Acta Neuropathol. Commun.* 7, 5. doi:10.1186/S40478-019-0744-0.

Kawaguchi, A., Yajima, N., Tsuchiya, N., Homma, J., Sano, M., Natsumeda, M., et al. (2013). Gene expression signature-based prognostic risk score in patients with glioblastoma. *Cancer Sci.* 104, 1205–1210. doi:10.1111/CAS.12214.

Kirtania, R., Banerjee, S., Laha, S., Shankar, B. U., Chatterjee, R., and Mitra, S. (2021). Deepsgp:Deep learning for gene selection and survival group prediction in glioblastoma. *Electron.* 10, 1463. doi:10.3390/ELECTRONICS10121463/S1.

Lee, Y., Scheck, A. C., Cloughesy, T. F., Lai, A., Dong, J., Farooqi, H. K., et al. (2008). Gene expression analysis of glioblastomas identifies the major molecular basis for the prognostic benefit of younger age. *BMC Med. Genomics* 1. doi:10.1186/1755-8794-1-52.

Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E., and Storey, J. D. (2012). The SVA package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* 28, 882–883. doi:10.1093/bioinformatics/bts034.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics* 27, 1739–1740. doi:10.1093/bioinformatics/btr260.

Mamoshina, P., Volosnikova, M., Ozerov, I. V., Putin, E., Skibina, E., Cortese, F., et al. (2018). Machine learning on human muscle transcriptomic data for biomarker discovery and tissue-specific drug target identification. *Front. Genet.* 9, 242. doi:10.3389/FGENE.2018.00242/BIBTEX.

Murat, A., Migliavacca, E., Gorlia, T., Lambiv, W. L., Shay, T., Hamou, M. F., et al. (2008). Stem cell-related "self-renewal" signature and high epidermal growth factor receptor expression associated with resistance to concomitant chemoradiotherapy in glioblastoma. *J. Clin. Oncol.* 26, 3015–3024. doi:10.1200/JCO.2007.15.7164.

Reifenberger, G., Weber, R. G., Riehmer, V., Kaulich, K., Willscher, E., Wirth, H., et al. (2014). Molecular characterization of long-term survivors of glioblastoma using genome- and transcriptome-wide profiling. *Int. J. Cancer* 135, 1822–1831. doi:10.1002/ijc.28836.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., et al. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* 43, e47–e47. doi:10.1093/nar/gkv007.

Scarbrough, P. M., Akushevich, I., Wrensch, M., and Il'yasova, D. (2014). Exploring the association between melanoma and glioma risks. *Ann. Epidemiol.* 24, 469. doi:10.1016/J.ANNEPIDEM.2014.02.010.

Senders, J. T., Staples, P., Mehrtash, A., Cote, D. J., Taphoorn, M. J. B., Reardon, D. A., et al. (2020). An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Neurosurgery* 86, E184–E192. doi:10.1093/NEUROS/NYZ403.

Serão, N. V., Delfino, K. R., Southey, B. R., Beever, J. E., and Rodriguez-Zas, S. L. (2011). Cell cycle and aging, morphogenesis, and response to stimuli genes are individualized biomarkers of glioblastoma progression and survival. *BMC Med. Genomics* 4, 1–21. doi:10.1186/1755-8794-4-49/FIGURES/6.

Stegmaier, P., Krull, M., Voss, N., Kel, A. E., and Wingender, E. (2010). Molecular mechanistic associations of human diseases. *BMC Syst. Biol.* 4, 124. doi:10.1186/1752-0509-4-124/FIGURES/9.

Therneau, T. (2021). A package for survival analysis in R.

Torres, R., and Judson-Torres, R. L. (2019). Research Techniques Made Simple: Feature Selection for Biomarker Discovery. *J. Invest. Dermatol.* 139, 2068-2074.e1. doi:10.1016/J.JID.2019.07.682.

Valdebenito, J., and Medina, F. (2019). Machine learning approaches to study glioblastoma: A review of the last decade of applications. *Cancer Rep.* 2. doi:10.1002/CNR2.1226.

Wang, Z., Wang, Y., Yang, T., Xing, H., Wang, Y., Gao, L., et al. (2021). Machine learning revealed stemness features and a novel stemness-based classification with appealing implications in discriminating the prognosis, immunotherapy and temozolomide responses of 906 glioblastoma patients. *Brief. Bioinform.* 22, 1–20. doi:10.1093/BIB/BBAB032.

Wingender, E., Hogan, J., Schacherer, F., Potapov, A. P., and Kel-Margoulis, O. (2007). Integrating

pathway data for systems pathology. in *In Silico Biology*.

Xia, Q., Xu, M., Zhang, P., Liu, L., Meng, X., and Dong, L. (2020). Therapeutic Potential of Autophagy in Glioblastoma Treatment With Phosphoinositide 3-Kinase/Protein Kinase B/Mammalian Target of Rapamycin Signaling Pathway Inhibitors. *Front. Oncol.* 10, 1886. doi:10.3389/FONC.2020.572904/BIBTEX.

Xie, Y., Meng, W. Y., Li, R. Z., Wang, Y. W., Qian, X., Chan, C., et al. (2021). Early lung cancer diagnostic biomarker discovery by machine learning methods. *Transl. Oncol.* 14, 100907. doi:10.1016/J.TRANON.2020.100907.

Yang, K., Stein, T. D., Huber, B. R., Sartor, E. A., Rachlin, J. R., and Mahalingam, M. (2021). Glioblastoma and malignant melanoma: Serendipitous or anticipated association? *Neuropathology* 41, 65–71. doi:10.1111/NEUP.12702.

Zhong, C., Shu, M., Ye, J., Wang, X., Chen, X., Liu, Z., et al. (2019). Oncogenic Ras is downregulated by ARHI and induces autophagy by Ras/AKT/mTOR pathway in glioblastoma. *BMC Cancer* 19, 1–14. doi:10.1186/S12885-019-5643-Z/FIGURES/7.

5 Conclusion

Glioblastomas (GBM) are grade IV gliomas. The prognosis is extremely poor and a majority of patients have a median survival time of fewer than 12 months and are referred to as short-term survivors (STS). Long-term survivors (LTS) are a rare group of patients who survive more than three years. Understanding the predictors and biological underpinnings of the LTS is an active area of research and investigations into clinical, radiological, histological, and genetic aspects have failed to yield an agreement on predictors of LTS. Using a variety of computational approaches, the present research thesis tries to uncover probable drivers of mechanisms driving prognosis in Glioblastoma.

In Chapter 1, I have used a meta-analytic approach to combine 14 independent GBM studies containing information of 2309 glioblastoma patients to study the predictors of survival (clinical and molecular) using a time-to-event analysis. Age and MGMT promoter methylation status were found to have a significant impact on survival. Several authors agree to the fact that age is an important prognostic factor, however, there is no age-group based information on the degree of risk a patient experiences because of his age factor. This work attempts to answer this question. Patients belonging to the age group of B (51 - 60 yrs) experience 20% more risk & age group C (61 - 70 yrs) experience 40% more risk of death respectively compared to a younger age group (40 - 50 yrs), group A. The older age group – group D (71 - 90 yrs) experiences a 2.4 times higher risk of death compared to group A.

Using time to event analysis, gene-expression profiles of 560 GBM patients of the TCGA database were analyzed to investigate their contribution to overall survival Chapter 2. This

analysis has revealed 720 genes that have prognostic value. Gene regulatory networks built on these genes revealed 14 TFs and 43 important drivers of upstream analysis. Some of the TFs enriched are know to be involved in the GBM pathogenesis such as 1) JUN a protooncogene that plays a critical role in cell proliferation and malignant transformation , 2) Glucocorticoid receptor(GR), which is shown to promote stem cells-like phenotype and resistance to chemotherapy , and 3) STAT3, persistent activation of which induces cell proliferation, anti-apoptosis, glioma stem cell maintenance, tumor invasion, angiogenesis, and immune evasion. The STAT3 which is significantly differentially expressed in short-term survivors (Log2FC = 0.403427421, adj.p-value = 0.00129) and also had a significant impact on survival with HR = 1.4 (p-value = 0.0015 with FDR = 0.009). Some of the master regulators proposed in the study have drugs that can target them, thus making these master regulators potential molecules for therapeutic intervention. This work is published in Biochemistry (Moscow), Supplement Series B: Biomedical Chemistry volume in November 2021. [Kalya et al., 2021a]

Comparative analyses of short-term and long-term GBM survivors revealed that gene expression patterns implemented across survival groups differ considerably. In light of these findings, Chapter 3 discusses the upstream analysis technique to learn about the gene regulatory networks that are responsible for the short survival. To the best of our knowledge, this is the first of its kind approach which attempts to explain the master regulators of poor prognosis and the molecular mechanism behind it. The gene regulatory network associated with STS in GBM is regulated by five master regulators, Insulin like growth factor binding protein-2 (IGFBP2), Vascular-Endothelial Growth Factor A (VEGFA), Platelet-Derived Growth Factor (PDGFA), Oncostatin M receptor (OSMR), and Adipocyte Enhancer Binding protein(AEBP1), and These five master regulators may present biomarkers of GBM prognosis and/or as therapeutic targets for enhancing survival in GBM. This work also proposes a novel mechanism of gene dysregulation by IGFBP2 by modulating a key molecule of tumor invasiveness and progression – the FRA-1 transcription factor. All the genes encoding these five master regulators have binding sites for FRA-1 in their promoters. FRA-1 and the master regulators cooperate in a positive feedback loop to orchestrate a complex tumorigenic program leading to poor survival in GBM. This work is published in Frontiers in Genetics, June 2021. [Kalya et al., 2021b]

In Chapter 4, I [Kalya et al., 2022] have explored different machine learning classifers that can predict the survival of GBM patients. To build these models I have used clinical information (Age) and transcriptomic profiles of 249 patients integrated from several microarray experiments. Of the 10 ML models that are tested, RandomForest model outperformed others with an accuracy of 80% on the test data. The model performed well on validation using external test data. A webtool (GlioSurvML) has been built using this model and made available for research applications. The features which are important for the classifier are discussed in detail. The biological relevance of these important features are explained based

on known biomarker information in the HumanPSD^{TM} database. 199 biomarkers of prognosis in Glioblastoma are proposed in this work of which 12 were found to have a significant impact on survival in GBM and were differentially expressed.This work is published in preprints, February 2022. [Kalya et al., 2022]

In summary, the entire work tried to identify probable biomarkers for GBM by using time to event analysis, upstream analysis and ML approach on gene expression databases. Around 242 gene-expression based biomarkers were identified of which PDGFA, AEBP1, and VEGFA were found to be important in all the approaches. They were also found to be master regulators driving gene-dysregulation with differential expression across survivor groups. 171 out of 242 biomarkers are not previously reported in GBM but they are found in other diseases like Ovarian carcinoma, Glioma and Melanoma. This underscores the importance of shared biology or similar pathogenesis in cancer development.I believe that the work contributes to understanding drivers as well as mechanisms driving prognosis in Glioblastoma.

A
Research Outputs

PEER-REVIEWED JOURNAL PUBLICATIONS

Published Publications

Kalya M, Kel A, Wlochowitz D, Wingender E, Beißbarth T. IGFBP2 Is a Potential Master Regulator Driving the Dysregulated Gene Network Responsible for Short Survival in Glioblastoma Multiforme. Front Genet. 2021 Jun 15;12:670240. DOI: 10.3389/fgene.2021.670240. PMID: 34211498; PMCID: PMC8239365
(link to the article)

Kalya, M., Beißbarth, T. & Kel, A.E. Master Regulators Associated with Poor Prognosis in Glioblastoma Multiforme. Biochem. Moscow Suppl. Ser. B 15, 263–273 (2021).
(link to the article)

Kalya, M.; Altynbekova.K, Alexander Kel. Master-regulators of host response to SARS-CoV-2 as promising targets for drug repurposing, Dec 2020 (link to the article)

Preprint Publications

Kalya, M.; Kel, A.; Leha, A.; Altynbekova, K.; Wingender, E.; beissbarth, T. Machine Learning based Survival Group Prediction in Glioblastoma . Preprints 2022, 2022020051 (doi: 10.20944/preprints202202.0051.v1)
"Machine Learning based Survival Group Prediction in Glioblastoma"

Publications In Process

Manasa Kalya;Alexander Kel;Tim Beißbarth *"Predictors of survival outcome in Glioblastoma: meta-analysis of individual patient data"*

Web Application developed in the study

GlioSurvML : Machine Learning based survival group prediction in Glioblastoma
(link to github)

Poster and Oral Presentations

"Modelling Therapeutic resistance in Glioblastoma using multi-Omics computational models" , *1st course on computational systems biology of Cancer* Institut Curie, Paris, France, Sep 2018

"Machine Learning methods for mechanism-based study on Glioblastoma Multiforme", *Brain Tumor Meeting*, Berlin, May 2019

Certifications

"Machine Learning: Data to Decisions" from Massachusetts Institute of Technology, USA, 2019

"Transcriptomics summer school", A workshop on Next-Generation Sequencing Data-Analysis, VIB, Belgium, 2019

Teaching Experience

Conducted hands on training workshop on **Basic R and RNA seq Data Analysis** in the University
In Collaboration with my colleague Darius Wlochowitz and mentoring of Prof.Tim Beißbarth, I supervised a 6-week internship of a Master's student on **Integrating Machine Learning and Upstream Analysis** approaches

References

[Abedalthagafi et al., 2018] Abedalthagafi, M., Barakeh, D., & Foshay, K. M. (2018). Immunogenetics of glioblastoma: the future of personalized patient management. *npj Precision Oncology 2018 2:1*, 2(1), 1–8.

[Aldape et al., 2015] Aldape, K., Zadeh, G., Mansouri, S., Reifenberger, G., & von Deimling, A. (2015). Glioblastoma: pathology, molecular mechanisms and markers. *Acta neuropathologica*, 129(6), 829–848.

[Armocida et al., 2019] Armocida, D., Pesce, A., Di Giammarco, F., Frati, A., Santoro, A., & Salvati, M. (2019). Long Term Survival in Patients Suffering from Glio-blastoma Multiforme: A Single-Center Observational Cohort Study. *Diagnostics 2019, Vol. 9, Page 209*, 9(4), 209.

[Baysan et al., 2012] Baysan, M., Bozdag, S., Cam, M. C., Kotliarova, S., Ahn, S., Walling, J., Killian, J. K., Stevenson, H., Meltzer, P., & Fine, H. A. (2012). G-Cimp Status Prediction Of Glioblastoma Samples Using mRNA Expression Data. *PLOS ONE*, 7(11), e47839.

[Bell et al., 2018] Bell, E. H., Zhang, P., Fisher, B. J., Macdonald, D. R., McElroy, J. P., Lesser, G. J., Fleming, J., Chakraborty, A. R., Liu, Z., Becker, A. P., Fabian, D., Aldape, K. D., Ashby, L. S., Werner-Wasik, M., Walker, E. M., Bahary, J. P., Kwok, Y., Yu, H. M., Laack, N. N., Schultz, C. J., Gray, H. J., Robins, H. I., Mehta, M. P., & Chakravarti, A. (2018). Association of MGMT Promoter Methylation Status with Survival Outcomes in Patients with High-Risk Glioma Treated with Radiotherapy and Temozolomide: An Analysis from the NRG Oncology/RTOG 0424 Trial. *JAMA Oncology*, 4(10), 1405–1409.

[Bi & Beroukhim, 2014] Bi, W. L. & Beroukhim, R. (2014). Beating the odds: Extreme long-term survival with glioblastoma. *Neuro-Oncology*, 16(9), 1159–1160.

[Boyarskikh et al., 2018] Boyarskikh, U., Pintus, S., Mandrik, N., Stelmashenko, D., Kiselev, I., Evshin, I., Sharipov, R., Stegmaier, P., Kolpakov, F., Filipenko, M., & Kel, A.

(2018). Computational master-regulator search reveals mTOR and PI3K pathways responsible for low sensitivity of NCI-H292 and A427 lung cancer cell lines to cytotoxic action of p53 activator Nutlin-3. *BMC Medical Genomics*, 11(S1), 12.

[Brennan et al., 2013] Brennan, C. W., Verhaak, R. G., McKenna, A., Campos, B., Noushmehr, H., Salama, S. R., Zheng, S., Chakravarty, D., Sanborn, J. Z., Berman, S. H., Beroukhim, R., Bernard, B., Wu, C. J., Genovese, G., Shmulevich, I., Barnholtz-Sloan, J., Zou, L., Vegesna, R., Shukla, S. A., Ciriello, G., Yung, W. K., Zhang, W., Sougnez, C., Mikkelsen, T., Aldape, K., Bigner, D. D., Van Meir, E. G., Prados, M., Sloan, A., Black, K. L., Eschbacher, J., Finocchiaro, G., Friedman, W., Andrews, D. W., Guha, A., Iacocca, M., O'Neill, B. P., Foltz, G., Myers, J., Weisenberger, D. J., Penny, R., Kucherlapati, R., Perou, C. M., Hayes, D. N., Gibbs, R., Marra, M., Mills, G. B., Lander, E. S., Spellman, P., Wilson, R., Sander, C., Weinstein, J., Meyerson, M., Gabriel, S., Laird, P. W., Haussler, D., Getz, G., Chin, L., Benz, C., Barrett, W., Ostrom, Q., Wolinsky, Y., Bose, B., Boulos, P. T., Boulos, M., Brown, J., Czerinski, C., Eppley, M., Kempista, T., Kitko, T., Koyfman, Y., Rabeno, B., Rastogi, P., Sugarman, M., Swanson, P., Yalamanchii, K., Otey, I. P., Liu, Y. S., Xiao, Y., Auman, J. T., Chen, P. C., Hadjipanayis, A., Lee, E., Lee, S., Park, P. J., Seidman, J., Yang, L., Kalkanis, S., Poisson, L. M., Raghunathan, A., Scarpace, L., Bressler, R., Eakin, A., Iype, L., Kreisberg, R. B., Leinonen, K., Reynolds, S., Rovira, H., Thorsson, V., Annala, M. J., Paulauskis, J., Curley, E., Hatfield, M., Mallery, D., Morris, S., Shelton, T., Shelton, C., Sherman, M., Yena, P., Cuppini, L., DiMeco, F., Eoli, M., Maderna, E., Pollo, B., Saini, M., Balu, S., Hoadley, K. A., Li, L., Miller, C. R., Shi, Y., Topal, M. D., Wu, J., Dunn, G., Giannini, C., Aksoy, B. A., Antipin, Y., Borsu, L., Cerami, E., Gao, J., Gross, B., Jacobsen, A., Ladanyi, M., Lash, A., Liang, Y., Reva, B., Schultz, N., Shen, R., Socci, N. D., Viale, A., Ferguson, M. L., Chen, Q. R., Demchok, J. A., Dillon, L. A., Mills Shaw, K. R., Sheth, M., Tarnuzzer, R., Wang, Z., Yang, L., Davidsen, T., Guyer, M. S., Ozenberger, B. A., Sofia, H. J., Bergsten, J., Eckman, J., Harr, J., Smith, C., Tucker, K., Winemiller, C., Zach, L. A., Ljubimova, J. Y., Eley, G., Ayala, B., Jensen, M. A., Kahn, A., Pihl, T. D., Pot, D. A., Wan, Y., Hansen, N., Hothi, P., Lin, B., Shah, N., Yoon, J. G., Lau, C., Berens, M., Ardlie, K., Carter, S. L., Cherniack, A. D., Noble, M., Cho, J., Cibulskis, K., DiCara, D., Frazer, S., Gabriel, S. B., Gehlenborg, N., Gentry, J., Heiman, D., Kim, J., Jing, R., Lawrence, M., Lin, P., Mallard, W., Onofrio, R. C., Saksena, G., Schumacher, S., Stojanov, P., Tabak, B., Voet, D., Zhang, H., Dees, N. N., Ding, L., Fulton, L. L., Fulton, R. S., Kanchi, K. L., Mardis, E. R., Wilson, R. K., Baylin, S. B., Harshyne, L., Cohen, M. L., Devine, K., Sloan, A. E., Van Den Berg, S. R., Berger, M. S., Carlin, D., Craft, B., Ellrott, K., Goldman, M., Goldstein, T., Grifford, M., Ma, S., Ng, S., Stuart, J., Swatloski, T., Waltman,

P., Zhu, J., Foss, R., Frentzen, B., McTiernan, R., Yachnis, A., Mao, Y., Akbani, R., Bogler, O., Fuller, G. N., Liu, W., Liu, Y., Lu, Y., Protopopov, A., Ren, X., Sun, Y., Yung, W. K., Zhang, J., Chen, K., Weinstein, J. N., Bootwalla, M. S., Lai, P. H., Triche, T. J., Van Den Berg, D. J., Gutmann, D. H., Lehman, N. L., Brat, D., Olson, J. J., Mastrogianakis, G. M., Devi, N. S., Zhang, Z., Lipp, E., & McLendon, R. (2013). The somatic genomic landscape of glioblastoma. *Cell*, 155(2), 462.

[Cao et al., 2019] Cao, M., Cai, J., Yuan, Y., Shi, Y., Wu, H., Liu, Q., Yao, Y., Chen, L., Dang, W., Zhang, X., Xiao, J., Yang, K., He, Z., Yao, X., Cui, Y., Zhang, X., & Bian, X. (2019). A four-gene signature-derived risk score for glioblastoma: prospects for prognostic and response predictive analyses. *Cancer Biology and Medicine*, 16(3), 595–605.

[Cooper, 2000] Cooper, G. M. (2000). Regulation of Transcription in Eukaryotes.

[Davis, 2016] Davis, M. E. (2016). Glioblastoma: Overview of disease and treatment. *Clinical Journal of Oncology Nursing*, 20(5), 1–8.

[Ellis & Kurian, 2014] Ellis, H. P. & Kurian, K. M. (2014). Biological Rationale for the Use of PPARÎ³ Agonists in Glioblastoma. *Frontiers in Oncology*, 4, 52.

[Ellor et al., 2014] Ellor, S. V., Pagano-Young, T. A., & Avgeropoulos, N. G. (2014). Glioblastoma: background, standard treatment paradigms, and supportive care considerations. *The Journal of law, medicine ethics : a journal of the American Society of Law, Medicine Ethics*, 42(2), 171–182.

[Endersby & Baker, 2008] Endersby, R. & Baker, S. J. (2008). PTEN signaling in brain: neuropathology and tumorigenesis. *Oncogene 2008 27:41*, 27(41), 5416–5430.

[Ensenyat-Mendez et al., 2021] Ensenyat-Mendez, M., Íñiguez-Muñoz, S., Sesé, B., & Marzese, D. M. (2021). iGlioSub: an integrative transcriptomic and epigenomic classifier for glioblastoma molecular subtypes. *BioData Mining*, 14(1), 1–16.

[Ferguson et al., 2021] Ferguson, S. D., Hodges, T. R., Majd, N. K., Alfaro-Munoz, K., Al-Holou, W. N., Suki, D., de Groot, J. F., Fuller, G. N., Xue, L., Li, M., Jacobs, C., Rao, G., Colen, R. R., Xiu, J., Verhaak, R., Spetzler, D., Khasraw, M., Sawaya, R., Long, J. P., & Heimberger, A. B. (2021). A validated integrated clinical and molecular glioblastoma long-term survival-predictive nomogram. *Neuro-Oncology Advances*, 3(1).

[Gan et al., 2009] Gan, H. K., Kaye, A. H., & Luwor, R. B. (2009). The EGFRvIII variant in glioblastoma multiforme. *Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia*, 16(6), 748–754.

[Gately et al., 2018] Gately, L., McLachlan, S. A., Philip, J., Ruben, J., & Dowling, A. (2018). Long-term survivors of glioblastoma: a closer look. *Journal of neuro-oncology*, 136(1), 155–162.

[Gerber et al., 2014] Gerber, N. K., Goenka, A., Turcan, S., Reyngold, M., Makarov, V., Kannan, K., Beal, K., Omuro, A., Yamada, Y., Gutin, P., Brennan, C. W., Huse, J. T., & Chan, T. A. (2014). Transcriptional diversity of long-term glioblastoma survivors. *Neuro-oncology*, 16(9), 1186–1195.

[Gilard et al., 2021] Gilard, V., Tebani, A., Dabaj, I., Laquerrière, A., Fontanilles, M., Derrey, S., Marret, S., & Bekri, S. (2021). Diagnosis and Management of Glioblastoma: A Comprehensive Perspective. *Journal of Personalized Medicine 2021, Vol. 11, Page 258*, 11(4), 258.

[Gustafsson et al., 2018] Gustafsson, J. R., Katsioudi, G., Degn, M., Ejlerskov, P., Issazadeh-Navikas, S., & Kornum, B. R. (2018). DNMT1 regulates expression of MHC class i in post-mitotic neurons. *Molecular Brain*, 11(1), 1–16.

[Hanif et al., 2017] Hanif, F., Muzaffar, K., Perveen, K., Malhi, S. M., & Simjee, S. U. (2017). Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. *Asian Pacific Journal of Cancer Prevention : APJCP*, 18(1), 3.

[Hatanpaa et al., 2010] Hatanpaa, K. J., Burma, S., Zhao, D., & Habib, A. A. (2010). Epidermal growth factor receptor in glioma: signal transduction, neuropathology, imaging, and radioresistance. *Neoplasia (New York, N.Y.)*, 12(9), 675–684.

[Henson, 2006] Henson, J. W. (2006). Treatment of Glioblastoma Multiforme: A New Standard. *Archives of Neurology*, 63(3), 337–341.

[Hu et al., 2006] Hu, Z., Fan, C., Oh, D. S., Marron, J. S., He, X., Qaqish, B. F., Livasy, C., Carey, L. A., Reynolds, E., Dressler, L., Nobel, A., Parker, J., Ewend, M. G., Sawyer, L. R., Wu, J., Liu, Y., Nanda, R., Tretiakova, M., Orrico, A. R., Dreher, D., Palazzo, J. P., Perreard, L., Nelson, E., Mone, M., Hansen, H., Mullins, M., Quackenbush, J. F., Ellis, M. J., Olopade, O. I., Bernard, P. S., & Perou, C. M. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7, 96.

[Hwang et al., 2019] Hwang, T., Mathios, D., McDonald, K. L., Daris, I., Park, S. H., Burger, P. C., Kim, S., Dho, Y. S., Carolyn, H., Bettegowda, C., Shin, J. H., Lim, M., & Park, C. K. (2019). Integrative analysis of DNA methylation suggests down-regulation of oncogenic pathways and reduced somatic mutation rates in survival outliers of glioblastoma. *Acta neuropathologica communications*, 7(1), 5.

[Iacob & Dinca, 2009] Iacob, G. & Dinca, E. B. (2009). Current data and strategy in glioblastoma multiforme. *Journal of Medicine and Life*, 2(4), 386.

[Irizarry et al., 2003] Irizarry, R. A., Bolstad, B. M., Collin, F., Cope, L. M., Hobbs, B., & Speed, T. P. (2003). Summaries of Affymetrix GeneChip probe level data. *Nucleic acids research*, 31(4), e15.

[Jiang et al., 2021] Jiang, H., Yu, K., Cui, Y., Ren, X., Li, M., Zhang, G., Yang, C., Zhao, X., Zhu, Q., & Lin, S. (2021). Differential Predictors and Clinical Implications Associated With Long-Term Survivors in IDH Wildtype and Mutant Glioblastoma. *Frontiers in Oncology*, 11.

[Johannessen et al., 2018] Johannessen, L. E., Brandal, P., Myklebust, T. Å., Heim, S., Micci, F., & Panagopoulos, I. (2018). MGMT Gene Promoter Methylation Status - Assessment of Two Pyrosequencing Kits and Three Methylation-specific PCR Methods for their Predictive Capacity in Glioblastomas. *Cancer genomics proteomics*, 15(6), 437–446.

[Kalya et al., 2021a] Kalya, M., Beißbarth, T., & Kel, A. E. (2021a). Master Regulators Associated with Poor Prognosis in Glioblastoma Multiforme. *Biochemistry (Moscow) Supplement Series B: Biomedical Chemistry*, 15(4), 263–273.

[Kalya et al., 2022] Kalya, M., Kel, A., Leha, A., Altynbekova, K., Wingender, E., & Beissbarth, T. (2022). Machine Learning based Survival Group Prediction in Glioblastoma.

[Kalya et al., 2021b] Kalya, M., Kel, A., Wlochowitz, D., Wingender, E., & Beißbarth, T. (2021b). IGFBP2 Is a Potential Master Regulator Driving the Dysregulated Gene Network Responsible for Short Survival in Glioblastoma Multiforme. *Frontiers in Genetics*, 12, 652.

[Kloosterhof et al., 2013] Kloosterhof, N. K., De Rooi, J. J., Kros, M., Eilers, P. H., Smitt, P. A., Van den Bent, M. J., & French, P. J. (2013). Molecular subtypes of glioma identified by genome-wide methylation profiling. *Genes, chromosomes cancer*, 52(7), 665–674.

[Koch et al., 2018] Koch, A., Joosten, S. C., Feng, Z., De Ruijter, T. C., Draht, M. X., Melotte, V., Smits, K. M., Veeck, J., Herman, J. G., Neste, L. V., Criekinge, W. V., De Meyer, T., & Engeland, M. V. (2018). Analysis of DNA methylation in cancer: location revisited. *Nature reviews. Clinical oncology*, 15(7), 459–466.

[Koschmann et al., 2015] Koschmann, J., Bhar, A., Stegmaier, P., Kel, A., & Wingender, E. (2015). "Upstream Analysis": An Integrated Promoter-Pathway Analysis Approach to Causal Interpretation of Microarray Data. *Microarrays*, 4(2), 270–286.

[Krex et al., 2007] Krex, D., Klink, B., Hartmann, C., von Deimling, A., Pietsch, T., Simon, M., Sabel, M., Steinbach, J. P., Heese, O., Reifenberger, G., Weller, M., & Schackert, G. (2007). Long-term survival with glioblastoma multiforme. *Brain*, 130(10), 2596–2606.

[Lofton-Day & Lesche, 2003] Lofton-Day, C. & Lesche, R. (2003). DNA methylation markers in patients with gastrointestinal cancers. Current understanding, potential applications for disease management and development of diagnostic tools. *Digestive diseases (Basel, Switzerland)*, 21(4), 299–308.

[Ma et al., 2020] Ma, H., Zhao, C., Zhao, Z., Hu, L., Ye, F., Wang, H., Fang, Z., Wu, Y., & Chen, X. (2020). Specific glioblastoma multiforme prognostic-subtype distinctions based on DNA methylation patterns. *Cancer gene therapy*, 27(9), 702–714.

[Mazaris et al., 2014] Mazaris, P., Hong, X., Altshuler, D., Schultz, L., Poisson, L. M., Jain, R., Mikkelsen, T., Rosenblum, M., & Kalkanis, S. (2014). Key determinants of short-term and long-term glioblastoma survival: A 14-year retrospective study of patients from the Hermelin Brain Tumor Center at Henry Ford Hospital. *Clinical Neurology and Neurosurgery*, 120, 103–112.

[MC & M, 2015] MC, C. & M, S. (2015). Combined analysis of TERT, EGFR, and IDH status defines distinct prognostic glioblastoma classes. *Neurology*, 84(19), 2007.

[Mrugala, 2013] Mrugala, M. M. (2013). Advances and Challenges in the Treatment of Glioblastoma: A Clinician's Perspective. *Discovery Medicine*, 15(83), 221–230.

[Muhammad et al., 2018] Muhammad, J. S., Khan, M. R., & Ghias, K. (2018). DNA methylation as an epigenetic regulator of gallbladder cancer: An overview. *International journal of surgery (London, England)*, 53, 178–183.

[Noushmehr et al., 2010] Noushmehr, H., Weisenberger, D. J., Diefes, K., Phillips, H. S., Pujara, K., Berman, B. P., Pan, F., Pelloski, C. E., Sulman, E. P., Bhat, K. P., Verhaak,

R. G., Hoadley, K. A., Hayes, D. N., Perou, C. M., Schmidt, H. K., Ding, L., Wilson, R. K., Van Den Berg, D., Shen, H., Bengtsson, H., Neuvial, P., Cope, L. M., Buckley, J., Herman, J. G., Baylin, S. B., Laird, P. W., & Aldape, K. (2010). Identification of a CpG Island Methylator Phenotype that Defines a Distinct Subgroup of Glioma. *Cancer cell*, 17(5), 510.

[Nutt et al., 2003] Nutt, C. L., Mani, D. R., Betensky, R. A., Tamayo, P., Cairncross, J. G., Ladd, C., Pohl, U., Hartmann, C., Mclaughlin, M. E., Batchelor, T. T., Black, P. M., Von Deimling, A., Pomeroy, S. L., Golub, T. R., & Louis, D. N. (2003). Gene Expression-based Classification of Malignant Gliomas Correlates Better with Survival than Histological Classification 1. *CANCER RESEARCH*, 63, 1602–1607.

[Park et al., 2019] Park, A. K., Kim, P., Ballester, L. Y., Esquenazi, Y., & Zhao, Z. (2019). Subtype-specific signaling pathways and genomic aberrations associated with prognosis of glioblastoma. *Neuro-oncology*, 21(1), 59–70.

[Paul et al., 2017] Paul, Y., Mondal, B., Patil, V., & Somasundaram, K. (2017). DNA methylation signatures for 2016 WHO classification subtypes of diffuse gliomas. *Clinical Epigenetics*, 9(1), 1–18.

[Phillips et al., 2006] Phillips, H. S., Kharbanda, S., Chen, R., Forrest, W. F., Soriano, R. H., Wu, T. D., Misra, A., Nigro, J. M., Colman, H., Soroceanu, L., Williams, P. M., Modrusan, Z., Feuerstein, B. G., & Aldape, K. (2006). Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. *Cancer cell*, 9(3), 157–173.

[Poon et al., 2020] Poon, M. T., Sudlow, C. L., Figueroa, J. D., & Brennan, P. M. (2020). Longer-term (≥ 2 years) survival in patients with glioblastoma in population-based studies pre- and post-2005: a systematic review and meta-analysis. *Scientific reports*, 10(1).

[Prasad et al., 2020] Prasad, B., Tian, Y., & Li, X. (2020). Large-Scale Analysis Reveals Gene Signature for Survival Prediction in Primary Glioblastoma. *Molecular Neurobiology*, 57(12), 5235–5246.

[Schena et al., 1995] Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235), 467–470.

[Schmitt et al., 2021] Schmitt, M. J., Company, C., Dramaretska, Y., Barozzi, I., Göhrig, A., Kertalli, S., Großmann, M., Naumann, H., Sanchez-Bailon, M. P., Hulsman, D.,

Glass, R., Squatrito, M., Serresi, M., & Gargiulo, G. (2021). Phenotypic Mapping of Pathologic Cross-Talk between Glioblastoma and Innate Immune Cells by Synthetic Genetic Tracing. *Cancer Discovery*, 11(3), 754–777.

[Scott et al., 2011] Scott, J., Tsai, Y. Y., Chinnaiyan, P., & Yu, H. H. M. (2011). Effectiveness of Radiotherapy for Elderly Patients With Glioblastoma. *International Journal of Radiation Oncology, Biology, Physics*, 81(1), 206–210.

[Senders et al., 2020] Senders, J. T., Staples, P., Mehrtash, A., Cote, D. J., Taphoorn, M. J., Reardon, D. A., Gormley, W. B., Smith, T. R., Broekman, M. L., & Arnaout, O. (2020). An Online Calculator for the Prediction of Survival in Glioblastoma Patients Using Classical Statistics and Machine Learning. *Neurosurgery*, 86(2), E184–E192.

[Songtao et al., 2012] Songtao, Q., Lei, Y., Si, G., Yanqing, D., Huixia, H., Xuelin, Z., Lanxiao, W., & Fei, Y. (2012). IDH mutations predict longer survival and response to temozolomide in secondary glioblastoma. *Cancer science*, 103(2), 269–273.

[Spiegl-Kreinecker et al., 2015] Spiegl-Kreinecker, S., Lötsch, D., Ghanim, B., Pirker, C., Mohr, T., Laaber, M., Weis, S., Olschowski, A., Webersinke, G., Pichler, J., & Berger, W. (2015). Prognostic quality of activating TERT promoter mutations in glioblastoma: interaction with the rs2853669 polymorphism and patient age at diagnosis. *Neuro-oncology*, 17(9), 1231–1240.

[Stegmaier et al., 2011] Stegmaier, P., Voss, N., Meier, T., Kel, A., Wingender, E., & Borlak, J. (2011). Advanced Computational Biology Methods Identify Molecular Switches for Malignancy in an EGF Mouse Model of Liver Cancer. *PLOS ONE*, 6(3), e17738.

[Tejero et al., 2019] Tejero, R., Huang, Y., Katsyv, I., Kluge, M., Lin, J. Y., Tome-Garcia, J., Daviaud, N., Wang, Y., Zhang, B., Tsankova, N. M., Friedel, C. C., Zou, H., & Friedel, R. H. (2019). Gene signatures of quiescent glioblastoma cells reveal mesenchymal shift and interactions with niche microenvironment. *EBioMedicine*, 42, 252–269.

[Teo et al., 2019] Teo, W. Y., Sekar, K., Seshachalam, P., Shen, J., Chow, W. Y., Lau, C. C., Yang, H. K., Park, J., Kang, S. G., Li, X., Nam, D. H., & Hui, K. M. (2019). Relevance of a TCGA-derived Glioblastoma Subtype Gene-Classifier among Patient Populations. *Scientific Reports*, 9(1).

[Thakkar et al., 2014] Thakkar, J. P., Dolecek, T. A., Horbinski, C., Ostrom, Q. T., Lightner, D. D., Barnholtz-Sloan, J. S., & Villano, J. L. (2014). Epidemiologic and Molecular Prognostic Review of Glioblastoma. *Cancer epidemiology, biomarkers prevention :*

a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology, 23(10), 1985.

[Valdebenito & Medina, 2019] Valdebenito, J. & Medina, F. (2019). Machine learning approaches to study glioblastoma: A review of the last decade of applications. *Cancer Reports*, 2(6).

[Verano-Braga et al., 2018] Verano-Braga, T., Gorshkov, V., Munthe, S., Sørensen, M. D., Kristensen, B. W., Kjeldsen, F., Verano-Braga, T., Gorshkov, V., Munthe, S., Sørensen, M. D., Kristensen, B. W., & Kjeldsen, F. (2018). SuperQuant-assisted comparative proteome analysis of glioblastoma subpopulations allows for identification of potential novel therapeutic targets and cell markers. *Oncotarget*, 9(10), 9400–9414.

[Verhaak et al., 2010] Verhaak, R. G., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James, C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G., Perou, C. M., & Hayes, D. N. (2010). Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, 17(1), 98–110.

[Wang et al., 2021] Wang, Z., Wang, Y., Yang, T., Xing, H., Wang, Y., Gao, L., Guo, X., Xing, B., Wang, Y., & Ma, W. (2021). Machine learning revealed stemness features and a novel stemness-based classification with appealing implications in discriminating the prognosis, immunotherapy and temozolomide responses of 906 glioblastoma patients. *Briefings in Bioinformatics*, 22(5), 1–20.

[Wen & Kesari, 2008] Wen, P. Y. & Kesari, S. (2008). Malignant gliomas in adults.

[Yin et al., 2018] Yin, A. A., Lu, N., Etcheverry, A., Aubry, M., Barnholtz-Sloan, J., Zhang, L. H., Mosser, J., Zhang, W., Zhang, X., Liu, Y. H., & He, Y. L. (2018). A novel prognostic six-CpG signature in glioblastomas. *CNS Neuroscience Therapeutics*, 24(3), 167.

[Yin et al., 2019] Yin, W., Tang, G., Zhou, Q., Cao, Y., Li, H., Fu, X., Wu, Z., & Jiang, X. (2019). Expression profile analysis identifies a novel five-gene signature to improve prognosis prediction of glioblastoma. *Frontiers in Genetics*, 10(MAY), 419.

[Yuan et al., 2016] Yuan, Y., Qi, C., Maling, G., Xiang, W., Yanhui, L., Ruofei, L., Yunhe, M., Jiewen, L., & Qing, M. (2016). TERT mutation in glioma: Frequency, prognosis

and risk. *Journal of clinical neuroscience : official journal of the Neurosurgical Society of Australasia*, 26, 57–62.

[Zeng et al., 2017] Zeng, W., Ren, X., Cui, Y., Jiang, H., Zhang, X., & Lin, S. (2017). 1q/19p co-polysomy predicts longer survival in patients with astrocytic gliomas. *Oncotarget*, 8(40), 67104.

[Zuo et al., 2019] Zuo, S., Zhang, X., & Wang, L. (2019). A RNA sequencing-based six-gene signature for survival prediction in patients with glioblastoma. *Scientific Reports 2019 9:1*, 9(1), 1–10.

Declaration

I hereby confirm that the work presented in this thesis entitled *"Machine learning and statistical analysis to identify determinants of survival in primary glioblastoma using genome wide expression studies"* is my own. The information derived from other sources has been quoted rightly in the thesis.

Göttingen, 03/02/2022

Manasa Kalya Purushothama