Integration of RNA-Seq and proteomics data identifies glioblastoma multiforme surfaceome signature

Saiful Effendi Syafruddin[1], Wan Fahmi Wan Mohamad Nazarie[2], Nurshahirah Ashikin Moidu[1], Bee Hong Soon[3], M. Aiman Mohtar [1,*]

[1] UKM Medical Molecular Biology Institute, UKM Medical Centre, Universiti Kebangsaan Malaysia, Bandar Tun Razak, 56000 Cheras, Kuala Lumpur, Malaysia

[2] Faculty of Science and Natural Resources, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia

[3] Neurosurgery Division, Department of Surgery, Faculty of Medicine, Universiti Kebangsaan Malaysia, Bandar Tun Razak, 56000 Cheras, Kuala Lumpur, Malaysia

* Correspondence: M. Aiman Mohtar, Email: m.aimanmohtar@ppukm.ukm.edu.my; Tel: +60391459062; Fax: +6039171 7185

ABSTRACT

Background

Glioblastoma multiforme (GBM) is a highly lethal, stage IV brain tumor with a prevalence of approximately 2 per 10000 people globally. The cell surface proteins or surfaceome play significant roles in modulating cancer phenotypes and acting as information gateway in many oncogenic signaling pathways. Hence, surfaceome are attractive targets for cancer therapy due to their direct accessibility to drugs. Nonetheless, a comprehensive GBM surfaceome landscape has not been fully defined. Thus, the aim of this study is to define GBM-specific surfaceome genes and identify key cell surface genes that could potentially be developed as novel GBM biomarkers, or therapeutically targeted.

Methods

We integrated the RNA-Seq data from TCGA GBM (n=166) and GTEx normal brain cortex (n=408) databases to identify the significantly dysregulated surfaceome in GBM. This was followed by integrative analysis that combines transcriptomics, proteomics and protein-protein interaction network data to prioritize the high-confidence GBM surfaceome signature.

Results

Among the 2,381 significantly dysregulated genes in GBM, 395 genes were classified as surfaceome. Via the integrative analysis, we identified 6 high-confidence GBM molecular signature, HLA-DRA, CD44, SLC1A5, EGFR, ITGB2, PTPRJ, which were significantly upregulated in GBM. The expression of these genes were validated in an independent transcriptomics database, which confirmed their upregulated expression in GBM. Importantly,

high expression of CD44, PTPRJ and HLA-DRA is significantly associated with poor disease-free survival. Lastly, using the Drugbank database, we identified several clinically-approved drugs targeting the GBM molecular signature suggesting potential drug repurposing.

Conclusions

In summary, we identified and highlighted the key GBM surface-enriched repertoires that could be biologically relevant in supporting GBM pathogenesis. These genes could be further interrogated experimentally in future studies that could lead to efficient GBM diagnostic/prognostic markers or a therapeutic regimen to treat GBM.

KEYWORDS

Differentially expressed genes, protein-protein interaction, cell surface proteins, network analysis, TCGA, GTEx

RUNNING TITLE

Syafruddin et al: Integrative analysis identifies high-confidence glioblastoma multiforme surfaceome signature for drug repurposing.

INTRODUCTION

Glioblastoma multiforme (GBM) is the most common and lethal tumor of the central nervous system in adults [1]. Despite decades of efforts to tackle this disease, the median survival rate of GBM patients is still not improving [2]. GBM patients have an average life expectancy of 15 months post-diagnosis and the 5-years survival rate is less than 3% [3]. The standard-of-care GBM treatment generally consists of maximal safe surgical resection followed by radiotherapy and concomitant chemotherapy. However rapid post-treatment relapse and high intra-tumoral heterogeneity that could either arise naturally during disease progression or treatments-induced have made this disease intractable and more challenging to treat [4, 5]. Therefore, there is a pressing need for better and efficient diagnostic and therapeutic strategies for this disease.

Temozolomide, an orally-administered DNA-alkylating drug, is the current and commonly used chemotherapy agent to treat GBM in clinic [6]. This combination treatment of temozolomide and radiotherapy is referred to as the Stupp regimen and it is widely used as the standard-of-care for the treatment of GBM. The landmark study showed that the combination of radiotherapy and concomitant chemotherapy with temozolomide improve the patient's prognosis compared to radiotherapy alone (median survival of 14.6 months vs 12.1 months, respectively) [6]. Alternative GBM treatment options such as the VEGF-targeting monoclonal antibody Bevacizumab, other DNA alkylating agents such as lomustine and carmustine implants, alternating electric field therapy and the checkpoint blockade inhibitor have thus far yielded low efficacy in treating GBM [2, 7, 8]. TCGA comprehensive GBM molecular characterizations have identified significant genetic alterations in several important oncogenic signalling pathways such as the RTK/Ras/PI3K (88%), p53 (87%) and pRB signaling pathways (78%) in GBM patients [9]. Several clinical trials are currently ongoing that aim to target these altered GBM oncogenic signaling pathways components using small molecule inhibitors and/or monoclonal antibodies. However, the results thus far were far from satisfactory [10]. This seems to suggest that instead of using a single agent targeting a specific component or pathway, novel treatments should consider the administration of several inhibitors targeting multiple different pathways.

The cell surface proteins or surfaceome serve as information gateway that integrates and transduces extracellular cues into intracellular signaling cascades or *vice versa*. Surfaceome also play important role in cell adhesion and migration which are among the

critical processes during tumorigenesis. Indeed, aberrant surfaceome expression and activity are frequently observed in many cancer types and therefore are good candidate for cancer diagnostic or biomarkers as well as therapeutic targets. Recent evidence has demonstrated that 56% of cell surface proteins are differentially expressed in GBM which are also present in cerebrospinal fluid or plasma, suggesting their potential use as biomarkers [11]. Of note, surfaceome expression is more dynamic than intracellular proteins and they could be sometimes cell type-specific [12, 13]. Mass spectrometry analysis showed that the average surfaceome size in brain cancer cell lines is higher than other cancer types [12]. Thus, surfaceome genes in GBM may hold the key to understand GBM pathogenesis and drug responsiveness, in which targeting these genes may unravel potential '*druggable*' stage in GBM pathways.

A comprehensive overview of GBM surfaceome landscape has not been fully defined. Therefore, the aim of this study was to characterize the GBM surfaceome genes expression profile by unifying the two large RNA-Seq datasets from the TCGA (GBM) and GTEx (normal brain). We integrated and performed differential gene expression analysis on these two datasets because of the low number of normal brain tissue samples in the TCGA database. Previously annotated surfaceome gene set was employed to filter and identify the significant differentially expressed surfaceome genes in GBM. To further prioritize the high-confidence GBM cell surface signature, we integrated our transcriptomics analysis with GBM tissues and cell surface proteomics, and PPI hub gene analysis. Collectively, we identified a list of upregulated surfaceome genes in GBM that include *CD44, PTPRJ* and *HLA-DRA* in which their biological relevance in supporting GBM pathogenesis could be comprehensively investigated in the future studies for the development of novel GBM diagnostic/prognostic or therapeutic strategies.

MATERIALS AND METHODS

TCGA and GTEx data acquisition, normalization and quality control: The analysis combined of the TCGA-GBM and GTEx normal brain RNA-Seq read count data. The GBM RNA-Seq gene raw read counts from TCGA were downloaded from Genomics Data Commons Data Portal (https://portal.gdc.cancer.gov). GTEx data were used for the normal brain tissues. The GTEx data used for the analyses described in this manuscript were obtained from the GTEx Portal on 29/03/19. We downloaded RNA-Seq gene raw read counts (from the cortex, frontal cortex, anterior cingulate cortex) from GTEx portal (https://gtexportal.org/home/datasets).

This allows us to perform the differentially expressed genes analysis on the 166 samples of GBM tumor from TCGA and 408 samples of normal brain tissues data from GTEx. The RNA-Seq raw read counts pre-processing steps involve are data filtering and data normalization. The normalization process of both data set are then performed by using mean as gene-level normalization using $\log_2$-counts per million where raw data are adjusted to account for factors that will prevent a direct comparison of expression measures and to safeguard the expression distributions are similar for each sample across the whole experiment. Data that unlikely to be informative or simply erroneous data will be removed by using variance filter (less than 15) and low abundance (less than 4).

Cell surface gene set classification and analysis: The identified DEGs of glioblastoma were classified into cell surface genes set as discussed in the main text (See Results 2.4). The classification of the gene sets was performed based on the mapping set of DEGs with this resource. Other genes, which did not map to this resource were removed from the final dataset.

Differential gene expression: Differential expressed genes (DEGs) analysis was performed using NetworkAnalyst [14], a web-based application tool for visualizing molecular and entity interactions. This platform utilizes the statistical method on data comparison from R package, limma to identify genes whose expression is different. Genes that have adjusted p-value <0.05 and log2 fold change |2| were considered as statistically significant DEGs.

Functional annotation and pathway analysis: The enrichment analysis of the identified glioblastoma associated genes was performed using DAVID (https://david.ncifcrf.gov/), a web-based tool for analyzing functional gene analysis. The tool comprises database from various public resources for biological analysis. The enrichment analysis such as GO and KEGG pathways were performed with top results as per gene counts.

Identification of hub genes through PPI network analysis: A biological database for known and predicted protein–protein interactions called IMEx interactome database (https://www.imexconsortium.org) was used to construct the protein–protein interaction (PPI) of the DEGs. The network of interacting proteins was extracted and visualized using NetworkAnalyst. The top 87-gene modules of highly interacting gene clusters among the DEG were found with default parameters. For the classified gene sets, the PPI network was

constructed and the network topological parameters i.e. degree and betweenness centrality were calculated.

Co-expression network of CD44: Co-expression analysis was perfomed using Graphia Professional (https://kajeka.com/graphia-professional/), previously known as BioLayout *Express*^{3D} [15] using raw read counts and then saved as an ‘‘.expression’’ file. This contains a unique identifier for each row of data. Following import into Graphia Professional, a pairwise Pearson correlation matrix was calculated thereby performing an gene vs. gene comparison of the expression profile of each genes. All Pearson correlations where r>0.7 were saved to a ‘‘.pearson’’ file. Based on a user defined threshold of r>0.75, an undirected network graph of the data was generated. In this context, nodes represent individual genes and the edges between them represent Pearson correlation coefficients above the selected threshold (r>0.75). *CD44* was selected along with its neighbor in the network, representing *CD44* co-expression partners. The class set of *CD44* co-expressed genes were visualized to compare the expression values in this class set with genes in normal samples.

RESULTS

Patients' characteristics of TCGA and GTEx: We utilized the publicly available TCGA and GTEx RNA-Seq database as our primary sources of GBM tumour and normal brain tissue transcriptomic data, respectively. We downloaded the datasets containing RNA-Seq gene expression profiles and clinical information of 166 patients from TCGA-GBM and 408 normal brain tissues from GTEx database. The combined data were stratified based on gender, age and treatment as shown in Table 1. Out of a total of 166 GBM cases, 104 cases (62.7%) were male and 56 cases (33.7%) were female. GBM is more prevalent in patients aged ≥ 60 years old which accounts for 42.8% of total cases in the TCGA GBM cohort. Fifty-two patients (31.3%) have undergone treatments whereas 62.1% cases did not have any treatment data. Unfortunately, the clinical data for the GTEx normal brain samples are not publicly available.

Identification of differentially expressed genes in glioblastoma: The analysis pipeline employed in this study is depicted in Fig 1. Briefly, the RNA-Seq raw read counts from the two large compendiums, TCGA and GTEx were utilized to identify the differentially expressed genes between GBM and normal brain tissues. Since most GBM cases are generally found in the supratentorial region of the brain such as the cerebral hemisphere [16], we only extracted

the RNA-Seq profiles of this region namely the cortex, frontal cortex, anterior cingulate cortex as per GTEx description. We performed t-distributed stochastic neighbor embedding (t-SNE) analysis to reflect the directionality of transcripts expression among GBM tumor and normal brain tissues read count values. The t-SNE plot showed that all RNA-Seq profiles of all GTEx cortex region clustered together while the GBM RNA-Seq profiles form a separate cluster, thus confirming distinct expression patterns between these groups (Fig. 2a). In total, there are 18,021 genes that have the RNA expression data from these combined TCGA and GTEx dataset but only 13,548 genes passed the quality control check. By applying the cut-off criteria $\log_2$ fold change |2| and adjusted p-value <0.05, we identified 2381 genes as significant differentially expressed genes (DEGs) in GBM, of which 648 genes were upregulated and 1733 genes were downregulated (Fig. 2b). The detailed information of the differential gene expression analysis is listed in Supplementary Table S1.

Functional enrichment analysis and classification of DEGs: The significant DEGs were then subjected to functional enrichment analysis using Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) tools to define their properties and putative biological relevance in GBM. Interestingly, the GO cellular component analysis of both upregulated and downregulated DEGs showed enrichment of cell surface and membrane-associated proteins (Supplementary Figure S1a and S1b). The KEGG pathway enrichment indicated that the upregulated DEGs are involved in pathways related to infectious diseases, pathway in cancer and cell adhesion (Supplementary Figure S1c). Downregulated genes mainly involve in neuroactive ligand-receptor interaction and major cellular signaling pathways (Supplementary Figure S1d).

Identification of GBM cell-surface antigen candidates: The DEGs were then further filtered and classified into the surfaceome gene set as previously defined by Bausch-Fluck et. al [13], Cunha et. al [17] and Lee et. al [18]. These studies utilized different criteria and stringency in curating the surfaceome gene list. From the overall DEGs in GBM, we identified 395 common cell surface genes within these three surfaceome definitions, including 124 upregulated and 271 downregulated genes (Supplementary Figure S2a and Supplementary Table S2). We further classified the surfaceome according to their main subclasses, which are receptors, transporters, enzymes, miscellaneous and unclassified, as previously reported by Almén et. al [19]. Among the defined surfaceome subclasses, 42.8% of the significant differentially expressed surfaceome in GBM belong to the receptor subclass (Supplementary Figure S2b).

KEGG analysis of the GBM-enriched cell surface proteins identified pathways related to immune defense and infectious disease pathways while GBM-deficient cell surface genes are enriched in pathways related to in neuroactive ligand-receptor interaction and major cellular signaling pathways (Supplementary Figure S3a and S3b). These findings are almost similar with the enrichment analysis of overall DEGs in GBM (Supplementary Figure S1c and S1d) suggesting that surfaceome has significant roles in dictating GBM cellular activities.

Identification of GBM cell-surface signature by integration of proteomics and transcriptomics data analysis: Thus far, we have (i) classified the overall DEGs in GBM using transcriptomics data and (ii) highlighted the differentially expressed cell surface genes in GBM. Even though this transcriptomics analysis is very informative for biomarker discovery, we aimed to add another layer of analysis to select for a more high-confidence cell surface signature for GBM. To attain this, we integrated our transcriptomics analysis data with the publicly available proteomics data. This integration will validate the cell surface genes prediction and eliminate possible discrepancy between the expression levels of mRNAs and proteins due to post-transcriptional and post-translational modifications. Thus, we gathered the publicly available quantitative mass spectrometry analysis data for both GBM tissues and cell lines. We postulated that GBM tissues and cell lines might have different cell surface repertoires and therefore it is important to stratify between these two sources. Additionally, GBM cell lines cell surface signature, as identified in this present study, could be validated experimentally in future functional studies.

Mass spectrometry analysis of five GBM cell lines revealed the upregulation of EGFR, CD44, PTPRJ, SLC1A5, F2R, and TSPAN6 proteins in these samples [12], whereby the expression level of these proteins were in concordance with our transcriptomics data analysis (Fig. 3). For tissue proteomics, we found several studies that performed comparative GBM vs. normal brain tissues proteome profiling [11, 20–23]. However, some of these studies either identified only limited number of proteins or the data are not downloadable. Only one study by Polisetty et al. that has identified large number of proteins in their proteome profiling study that included 1834 high-confidence membrane proteins with more than 2-fold change [11]. We therefore used this dataset where we performed integrative analysis with our analyzed transcriptomics data and identified 10 overlapped genes, *MRC2, FCGR3A, HLA-DRA, CD44, CD74, MSR1, CD163, EGFR, ITGB2, PTPRZ1* (Fig. 3). The mRNA expression levels correlated with the protein expression levels except the *PTPRZ1* where the mRNA levels showed upregulation while

proteomics data showed downregulation (Supplementary Table S3 and S4). In total, there are 14 genes from the combined tissues and cell lines proteomics that overlapped with our transcriptomics data (Fig. 3). It is important to note that proteins identification in mass spectrometry can be limiting due to protein isolation methods, proteins solubility, and other intrinsic variations that affect the proteins abundance as well as the sensitivity and detection capability of the MS instrumentation. Thus, these limitations may underestimate the results between transcriptomics prediction and proteomics discoveries.

Surfaceome protein-protein interaction network cluster analysis and prioritization of high-confidence GBM cell surface markers: We set out to further analyze the GBM-enriched cell surface markers using protein-protein interaction (PPI) network analysis. This is to better understand the interplay between the cell surface genes within the identified DEGs as well as with other genes. More importantly, this would enable us to further select the genes that are highly interconnected from the integrated proteomics and transcriptomics analysis. Network analysis of the identified differentially expressed cell surface protein genes was performed using NetworkAnalyst [14] to determine the relationship between genes according to the network topological parameters such as degree and betweenness. These parameters reflect the role and property of proteins within the network. The nodes and edges in the PPI network represent the proteins and their interactions, respectively. The GBM-enriched cell surface proteins network contains 1,321 nodes and 1,767 edges interactions based on a number of validated features including functional experiments, co-expression analysis, text mining, neighborhood, gene fusion and databases (Figure 4a). We identified 87-gene modules of clusters and the top cluster genes with more than 30 interactions include *VCAM1, EGFR, TGFBR1, CD44, NGFR, ITGB2, DCC, PTPRJ, ANBCA1, HLA-DRA, CCR5* and *CSF1R* (Fig. 4a and Supplementary Table S5). Vascular Cell Adhesion Molecule 1 (VCAM1) has the highest interacting cluster as it was found to have 426 degree with 422,712.18 betweenness score. We subsequently mapped the 14 genes identified from the integrated transcriptomics and proteomics data analysis (Fig. 3) with the top genes that have at least 20 interactions from the PPI network analysis. We found 6 genes that were in common between these two datasets which represent the high-confidence GBM predictive surfaceome markers (Fig. 4b).

Validation of high-confidence GBM cell surface markers and correlation of expression of individual DEGs in disease-free survival: Next, we validated the expression profiles of the identified 6 high-confidence cell surface markers using an independent database, Gene

Expression Profiling Interactive Analysis (GEPIA) [24]. GEPIA also combines the TCGA and GTEx gene expression data that were processed from raw reads count and unified using its own pipeline. The identified GBM cell surface signature genes were confirmed to be significantly upregulated in GBM (Supplementary Figure S4a – S4f). To investigate whether these signature genes could modulate/influence the GBM patient prognosis, Kaplan–Meier survival analysis was performed using the aforementioned GEPIA online tool. We examined the disease-free survival profile of the GBM patients who have either high or low expression level of these 6 GBM signature genes. Since GBM patients have low overall survival rate (average <2 years' survival post-diagnosis), we supposed it would be more reasonable to look at the disease-free survival end point rather than the overall survival. Overall survival end point is more suited towards longer follow-up period (typically 5 years) for the data to be meaningful [25]. The disease-free survival analysis showed that high expression of only *CD44*, *PTPRJ* and *HLA-DRA* were significantly correlated ($p<0.05$) with poor disease-free survival (Supplementary Figure S5a – S5f).

Co-expression network of CD44: CD44 is a transmembrane receptor and has multifaceted functions in both normal and disease physiology. OMICS studies have identified CD44 to be overexpressed in many types of cancer including glioblastoma [26, 27]. Based on our analysis, CD44 seems particularly important as it can be both identified in transcriptomics and proteomics-based approaches, among the top hub gene and whose high expression correlate with poor disease-free survival. We performed co-expression network analysis to further interrogate its association with other genes using our transcriptomics. The nodes represent in the network analysis represent genes, while the edges represent Pearson correlation above $r>0.75$. The neighboring genes connected to CD44 was extracted and shown in Fig 5a. There are 27 genes in this complex connected to CD44. Among the highly correlated genes are ELK3, CLIC4, GALNT2, TNC, and VIM. All genes in this CD44 co-expression cluster are highly expressed in GBM compared to normal brain samples (Fig 5b), further corroborating the biological relevance of CD44 in supporting GBM pathogenesis.

Identifying existing drugs targeting GBM signature and CD44 network: We next determined whether there are any clinically approved drugs or binding molecules for the identified high confidence GBM cell surface markers (Supplementary Figure S4) and components of the constructed CD44 co-expression network (Fig. 5a). To achieve this, we utilized the genes-drugs database curated by Cheng and colleagues [28] and cross-checked with the Drugbank

website (https://www.drugbank.ca/). Based on our analysis, we found several approved drugs or binding targets that are available for some of these proteins which are summarized in Table 2. Hyaluronic acid, which is the known binding target for CD44 receptor, has been approved in treating disease such as osteoarthritis [29]. On top of this, hyaluronic acid is widely used in ophthalmology and dermatology treatments and procedures [30].

Due to its high binding affinity to CD44 receptor, which are commonly observed to be overexpressed in many cancer types [31] including in GBM (reported in this present study), hyaluronic acid has been further developed for cancer-specific drugs or gene delivery nanomaterials [32, 33]. Thus, the excellent features of hyaluronic acid in mediating enhanced drugs or genes delivery to cancer cells via the overexpressed CD44 receptor could potentially be applied and developed for novel GBM therapeutic strategies. Moreover, within the components of CD44 co-expression network, only complement C1R (C1R) and calreticulin (CALR) are targetable by drugs. For C1R, several monoclonal antibodies such as cetuximab and bevacizumab have been clinically-approved to target this protein whereby antihemophilic factor, tenecteplase and melatonin are approved to target CALR. Overall, the identification of these molecules and drugs (Table 2) could disrupt GBM interactome and open new avenues for better GBM treatment options. This could be either a single drug or multi-drugs treatment approaches as well as in potential combination with chemotherapy, temozolomide.

DISCUSSION

The surfaceome comprise cellular frontiers that permit/inhibit signal transduction as well as playing important roles in modulating cells proliferation, migration and invasion, and cells-cells interaction. The surfaceome can organize themselves at a nanoscale resolution [34]. This spatiotemporal nanoscale organization could define the cell identity and phenotypes, and capacity to communicate with microenvironments such as the extracellular matrix, growth factors, hormones and drugs. Due to their accessibility on the cell membrane, surfaceome are ideal candidates for biomarkers and often targeted for drugs development. In fact, over 50% of drugs curated in the DrugBank target the surfaceome. In addition to their ubiquitous expression on the plasma membrane, the extracellular stalks of these cell surface proteins can be cleaved and released into the bloodstream, making them as suitable targets for blood-based diagnostics. Surfaceome can also be draped with glycans during post-translational modifications, which

will mediate their interaction with other proteins that reside on either the same or neighboring cells as well as with the microenvironments [34]. Dysregulated surfaceome expression and functions have been shown to promote tumor formation and progression [35]. Therefore, scientists have begun profiling and cataloging surfaceome in various types of cancers [36–39]. Cell surface proteins can be elevated in cancer cells in which they can response to increased level of growth factors, rendering cancer cells to sustain their infinite proliferative capabilities [40] and interact with the microenvironment that could either direct or indirectly modulate the tumor growth and metastatic capabilities [41].

The GBM transcriptomics dataset have been previously utilized to uncover genes that support GBM pathogenesis as well as genes that have potential prognostic values [42–44]. For example, Nicolasjilwan et al. analyzed the TCGA database to predict the survival of GBM patients based on clinical features, MRI images genomics alterations [42]. However, most TCGA GBM differential genes expression analyses either relied on low number of normal brain tissue samples, in which the TCGA GBM cohort contained only 5 normal brain tissues RNA-Seq data, or the data were combined with the GBM TCGA microarray data. This might create imbalance that would lead to inaccuracy or bias in the downstream analysis. Hence, in order to increase the robustness of this study in identifying the significantly upregulated GBM surfaceome repertoire, we included the normal brain tissues GTEx RNA-Seq database TCGA in our analysis. On a similar scale, the GTEx studies have performed genes expression profiling in more than 11,000 samples across multiple human tissues from nearly 1,000 healthy donors. We compared the TCGA GBM and normal cortex GTEx RNA-seq data and identified 2,381 significant differentially expressed genes in GBM, in which 648 were upregulated and 1,733 downregulated genes. In agreement with previous GBM proteomics profiling study [12], the GO cellular compartment analysis showed that most of the dysregulated genes in GBM encode for the cell surface proteins, suggesting the importance of cell surface proteins in GBM pathogenesis.

Out of the 2,381 significant DEGs in GBM, 395 genes encode for cell surface proteins, in which 124 and 271 genes were found to be significantly upregulated and downregulated, respectively. Interestingly, receptor subclass was the predominant dysregulated genes in GBM, suggesting the crucial roles of cell surface receptors in supporting GBM pathogenesis. This was indeed in line with a plethora of studies that have reported the implications of cell surface receptors dysregulation in the pathogenesis of many cancer types [45]. For this reason, the

development of cancer treatment strategies have been revolved around targeting the cell surface receptors such as the receptor tyrosine kinases (RTKs) [46] and G protein-coupled receptors (GPCRs) [47]. Therefore, targeting the cell surface proteins particularly the receptor subclass could potentially be further explored as novel GBM therapeutic options.

Robust cancer biomarkers are those that could be reproducibly identified by multi-omics platforms or reported in several different studies. To this end, we integrated the analyzed transcriptomics data with publicly available GBM proteomics data to prioritize for high-confidence cell surface proteins. Also, due to post-transcriptional and post-translational modifications, the mRNAs expression level are sometimes not correlated with their respective protein expression levels [48]. After mapping the prioritized genes from the transcriptomics-proteomics integrative analysis with the PPI network analysis data, we identified 6 genes; *HLA-DRA, CD44, SLC1A5, EGFR, ITGB2, PTPRJ*, whereby we considered these genes as the high-confidence GBM predictive surface markers. Disease-free survival analysis demonstrated that *CD44, PTPRJ,* and *HLA-DRA5* expression level correlated with GBM patient prognostic outcomes. For example, GBM patients who had lower *CD44* expression had better disease-free survival rate than patients whose *CD44* expression was upregulated. This indicates that these 3 genes, *CD44, PTPRJ,* and *HLA-DRA5,* could potentially be developed as GBM prognostic markers in the clinic.

Using a graph-based analytics [15], we constructed the CD44 co-expression network and checked for the approved drugs or binding targets for CD44, components of the CD44 co-expression network as well as the other high confidence GBM predictive surface markers. Only CD44, C1R, CALR, EGFR and ITGB2 are targetable by drugs or binding targets curated in the Drugbank database. *CD44* encodes a transmembrane glycoprotein that serves as the receptor for hyaluronic acid and several other ligands such as osteopontin, fibronectin and collagen [27]. CD44 has been implicated in supporting the tumorigenesis of many cancer types [27, 49]. Proteogenomic profiling of GBM tissues showed high expression of CD44 [23] and recent GBM systematic analysis further supported CD44 as a GBM cell surface antigen [26]. Interestingly, this transmembrane glycoprotein can be cleaved and secreted into the vasculatures, signifying its potential to be developed as diagnostic marker [50]. It has also been reported that the activation of CD44 by its ligand promotes cancer stem cell-like phenotypes in GBM and increases therapeutic resistance [51].

The current approved therapies to treat GBM are far from satisfactory and have remained unchanged for more than a decade [52]. This includes the alkylating agent temozolomide, which is the first line of drug used in treating GBM. Therefore, there is a need for novel or alternative treatment strategies for GBM. Due to the upregulated expression of CD44 in GBM, drugs targeting CD44 are currently undergoing clinical trials and the results are thus far promising in that CD44 inhibition impedes GBM cells growth [53]. In addition to this, our drug mapping analysis revealed hyaluronic acid as an actionable CD44 binding molecule. It is therefore appealing to investigate the activity of this existing drug in eradicating GBM. Our co-expression network analysis demonstrated that genes connected to CD44 were also highly expressed in GBM compared to normal brain tissues, suggesting that CD44 signaling axis is important in GBM tumorigenesis. Within the CD44 network, there are two proteins, namely the C1R and CALR that are druggable with several molecules and these can be potentially used to disrupt CD44 interactome and reduce GBM fitness. In addition, there are a number of approved drugs to target another high-confidence GBM signature, EGFR [54]. Studying a combination of these available drugs targeting EGFR or CD44 axis along with temozolomide could synergistically improve overall GBM patients's survival.

CONCLUSIONS

In summary, we identified GBM surfaceome by combining RNA-seq data. Through an integrative multi-OMICS strategy, we highlighted 6 GBM surface-enriched genes that could be important in driving GBM development. Some of these genes can be targeted by known drugs or molecules in the Drugbank suggesting potential drug repurposing. Additionally, further studies of these genes could lead to potential GBM diagnostic/prognostic markers or a therapeutic regimen to treat GBM.

ACKNOWLEDGEMENTS

The authors thank David Shorthouse (MRC Cancer Unit, University of Cambridge) and Low Teck Yew (UKM Medical Molecular Biology Institute) for discussion, critical insight and proofreading the manuscript.

FUNDING

MAM is supported by Fundamental Research Grant Scheme by the Ministry of Education, Malaysia (FRGS/1/2018/STG04/UKM/03/1) and Collaborative Research Programme - International Centre for Genetic Engineering and Biotechnology Grant (CRP/MYS19-04_EC).

AUTHORS CONTRIBUTION

Conceptualization, MAM and SES; methodology, WFWMN; software, WFWN; formal analysis, WFWMN, MAM and SES; investigation, WFWMN, MAM and SES; resources, WFWMN and NAM; data curation, NAM and SBH.; writing—original draft preparation, SES, MAM and WFWMN; supervision, MAM; writing—review & editing, MAM and SES; funding acquisition, MAM.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AVAILABALITY OF DATA MATERIALS

The data are included within the manuscript and in the supplementary files. The TCGA GBM data can be obtained from Genomics Data Commons Data Portal (https://portal.gdc.cancer.gov). The normal brain tissues RNA-seq data were obtained from the GTEx Portal (https://gtexportal.org/home/datasets).

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA Cancer J Clin. 2016;66:7–30.

2. Kamiya-Matsuoka C, Gilbert MR. Treating recurrent glioblastoma: an update. CNS Oncol. 2015;4:91–104.

3. Ohgaki H. Epidemiology of brain tumors. Methods Mol Biol Clifton NJ. 2009;472:323–42.

4. Qazi MA, Vora P, Venugopal C, Sidhu SS, Moffat J, Swanton C, et al. Intratumoral heterogeneity: pathways to treatment resistance and relapse in human glioblastoma. Ann Oncol. 2017;28:1448–56.

5. Shergalis A, Bankhead A, Luesakul U, Muangsin N, Neamati N. Current Challenges and Opportunities in Treating Glioblastoma. Pharmacol Rev. 2018;70:412–45.

6. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJB, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. N Engl J Med. 2005;352:987–96.

7. Ito H, Nakashima H, Chiocca EA. Molecular responses to immune checkpoint blockade in glioblastoma. Nat Med. 2019;25:359.

8. Nam JY, de Groot JF. Treatment of Glioblastoma. J Oncol Pract. 2017;13:629–38.

9. Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature. 2008;455:1061–8.

10. Pearson JRD, Regad T. Targeting cellular pathways in glioblastoma multiforme. Signal Transduct Target Ther. 2017;2:17040.

11. Polisetty RV, Gautam P, Sharma R, Harsha HC, Nair SC, Gupta MK, et al. LC-MS/MS Analysis of Differentially Expressed Glioblastoma Membrane Proteome Reveals Altered Calcium Signaling and Other Protein Groups of Regulatory Functions. Mol Cell Proteomics. 2012;11:M111.013565.

12. Bausch-Fluck D, Hofmann A, Bock T, Frei AP, Cerciello F, Jacobs A, et al. A Mass Spectrometric-Derived Cell Surface Protein Atlas. PLOS ONE. 2015;10:e0121314.

13. Bausch-Fluck D, Goldmann U, Müller S, Oostrum M van, Müller M, Schubert OT, et al. The in silico human surfaceome. Proc Natl Acad Sci. 2018;115:E10988–97.

14. Xia J, Gill EE, Hancock REW. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc. 2015;10:823–44.

15. Theocharidis A, van Dongen S, Enright AJ, Freeman TC. Network visualization and analysis of gene expression data using BioLayout Express(3D). Nat Protoc. 2009;4:1535–50.

16. Nakada M, Kita D, Watanabe T, Hayashi Y, Teng L, Pyko IV, et al. Aberrant Signaling Pathways in Glioma. Cancers. 2011;3:3242–78.

17. Cunha JPC da, Galante P a. F, Souza JE de, Souza RF de, Carvalho PM, Ohara DT, et al. Bioinformatics construction of the human cell surfaceome. Proc Natl Acad Sci. 2009;106:16752–7.

18. Lee JK, Bangayan NJ, Chai T, Smith BA, Pariva TE, Yun S, et al. Systemic surfaceome profiling identifies target antigens for immune-based therapy in subtypes of advanced prostate cancer. Proc Natl Acad Sci. 2018;115:E4473–82.

19. Almén MS, Nordström KJV, Fredriksson R, Schiöth HB. Mapping the human membrane proteome: a majority of the human membrane proteins can be classified according to function and evolutionary origin. BMC Biol. 2009;7:50.

20. Banerjee HN, Mahaffey K, Riddick E, Banerjee A, Bhowmik N, Patra M. Search for a diagnostic/prognostic biomarker for the brain cancer glioblastoma multiforme by 2D-DIGE-MS technique. Mol Cell Biochem. 2012;367:59–63.

21. Collet B, Guitton N, Saïkali S, Avril T, Pineau C, Hamlat A, et al. Differential analysis of glioblastoma multiforme proteome by a 2D-DIGE approach. Proteome Sci. 2011;9:16.

22. Heroux MS, Chesnik MA, Halligan BD, Al-Gizawiy M, Connelly JM, Mueller WM, et al. Comprehensive characterization of glioblastoma tumor tissues for biomarker identification using mass spectrometry-based label-free quantitative proteomics. Physiol Genomics. 2014;46:467–81.

23. Song Y-C, Lu G-X, Zhang H-W, Zhong X-M, Cong X-L, Xue S-B, et al. Proteogenomic characterization and integrative analysis of glioblastoma multiforme. Oncotarget. 2017;8:97304–12.

24. Tang Z, Li C, Kang B, Gao G, Li C, Zhang Z. GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. Nucleic Acids Res. 2017;45:W98–102.

25. Sargent DJ, Wieand HS, Haller DG, Gray R, Benedetti JK, Buyse M, et al. Disease-Free Survival Versus Overall Survival As a Primary End Point for Adjuvant Colon Cancer Studies: Individual Patient Data From 20,898 Patients on 18 Randomized Trials. J Clin Oncol. 2005;23:8664–70.

26. Ghosh D, Funk CC, Caballero J, Shah N, Rouleau K, Earls JC, et al. A Cell-Surface Membrane Protein Signature for Glioblastoma. Cell Syst. 2017;4:516-529.e7.

27. Chen C, Zhao S, Karnad A, Freeman JW. The biology and role of CD44 in cancer progression: therapeutic implications. J Hematol OncolJ Hematol Oncol. 2018;11:64.

28. Cheng F, Kovács IA, Barabási A-L. Network-based prediction of drug combinations. Nat Commun. 2019;10:1–11.

29. Bowman S, Awad ME, Hamrick MW, Hunter M, Fulzele S. Recent advances in hyaluronic acid based therapy for osteoarthritis. Clin Transl Med. 2018;7. doi:10.1186/s40169-017-0180-3.

30. Huynh A, Priefer R. Hyaluronic acid applications in ophthalmology, rheumatology, and dermatology. Carbohydr Res. 2020;489:107950.

31. Chen C, Zhao S, Karnad A, Freeman JW. The biology and role of CD44 in cancer progression: therapeutic implications. J Hematol OncolJ Hematol Oncol. 2018;11. doi:10.1186/s13045-018-0605-5.

32. Kim JH, Moon MJ, Kim DY, Heo SH, Jeong YY. Hyaluronic Acid-Based Nanomaterials for Cancer Therapy. Polymers. 2018;10. doi:10.3390/polym10101133.

33. Kim K, Choi H, Choi ES, Park M-H, Ryu J-H. Hyaluronic Acid-Coated Nanomedicine for Targeted Cancer Therapy. Pharmaceutics. 2019;11.

34. Bausch-Fluck D, Milani ES, Wollscheid B. Surfaceome nanoscale organization and extracellular interaction networks. Curr Opin Chem Biol. 2019;48:26–33.

35. Teh JLF, Chen S. Glutamatergic signaling in cellular transformation. Pigment Cell Melanoma Res. 2012;25:331–42.

36. Mirkowska P, Hofmann A, Sedek L, Slamova L, Mejstrikova E, Szczepanski T, et al. Leukemia surfaceome analysis reveals new disease-associated features. Blood. 2013;121:e149–59.

37. Fenner A. Surfaceome profiling for NEPC target antigens. Nat Rev Urol. 2018;15:396–7.

38. Ziegler A, Cerciello F, Bigosch C, Bausch-Fluck D, Felley-Bosco E, Ossola R, et al. Proteomic surfaceome analysis of mesothelioma. Lung Cancer. 2012;75:189–96.

39. Pais H, Ruggero K, Zhang J, Al-Assar O, Bery N, Bhuller R, et al. Surfaceome interrogation using an RNA-seq approach highlights leukemia initiating cell biomarkers in an LMO2 T cell transgenic model. Sci Rep. 2019;9:1–16.

40. Hanahan D, Weinberg RA. Hallmarks of Cancer: The Next Generation. Cell. 2011;144:646–74.

41. Leth-Larsen R, Lund RR, Ditzel HJ. Plasma membrane proteomics and its application in clinical cancer biomarker discovery. Mol Cell Proteomics MCP. 2010;9:1369–82.

42. Nicolasjilwan M, Hu Y, Yan C, Meerzaman D, Holder CA, Gutman D, et al. Addition of MR imaging features and genetic biomarkers strengthens glioblastoma survival prediction in TCGA patients. J Neuroradiol J Neuroradiol. 2015;42:212–21.

43. Han J, Puri RK. Analysis of the cancer genome atlas (TCGA) database identifies an inverse relationship between interleukin-13 receptor α1 and α2 gene expression and poor prognosis and drug resistance in subjects with glioblastoma multiforme. J Neurooncol. 2018;136:463–74.

44. Jia D, Li S, Li D, Xue H, Yang D, Liu Y. Mining TCGA database for genes of prognostic value in glioblastoma microenvironment. Aging. 2018;10:592–605.

45. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, et al. Oncogenic Signaling Pathways in The Cancer Genome Atlas. Cell. 2018;173:321-337.e10.

46. Regad T. Targeting RTK Signaling Pathways in Cancer. Cancers. 2015;7:1758–84.

47. Lundstrom K. An Overview on GPCRs and Drug Discovery: Structure-Based Drug Design and Structural Biology on GPCRs. In: Leifert WR, editor. G Protein-Coupled Receptors in Drug Discovery. Totowa, NJ: Humana Press; 2009. p. 51–66. doi:10.1007/978-1-60327-317-6_4.

48. Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. Nat Rev Genet. 2012;13:227–32.

49. Senbanjo LT, Chellaiah MA. CD44: A Multifunctional Cell Surface Adhesion Receptor Is a Regulator of Progression and Metastasis of Cancer Cells. Front Cell Dev Biol. 2017;5. doi:10.3389/fcell.2017.00018.

50. Lim S, Kim D, Ju S, Shin S, Cho I, Park S-H, et al. Glioblastoma-secreted soluble CD44 activates tau pathology in the brain. Exp Mol Med. 2018;50:1–11.

51. Pietras A, Katz AM, Ekström EJ, Wee B, Halliday JJ, Pitter KL, et al. Osteopontin-CD44 signaling in the glioma perivascular niche enhances cancer stem cell phenotypes and promotes aggressive tumor growth. Cell Stem Cell. 2014;14:357–69.

52. Kazda T, Dziacky A, Burkon P, Pospisil P, Slavik M, Rehak Z, et al. Radiotherapy of Glioblastoma 15 Years after the Landmark Stupp's Trial: More Controversies than Standards? Radiol Oncol. 2018;52:121–8.

53. Mooney KL, Choy W, Sidhu S, Pelargos P, Bui TT, Voth B, et al. The role of CD44 in glioblastoma multiforme. J Clin Neurosci Off J Neurosurg Soc Australas. 2016;34:1–5.

54. Singh D, Attri BK, Gill RK, Bariwal J. Review on EGFR Inhibitors: Critical Updates. Mini Rev Med Chem. 2016;16:1134–66.

FIGURE AND TABLE LEGENDS

Table 1 TCGA GBM patients' clinical data

Table 2 Available approved drugs or biding molecules targeting the identified GBM molecular signature and CD44 co-expression network

Fig. 1 Analysis pipeline to obtain the GBM predictive surfaceome markers applied from the initial TCGA GBM and GTEx data integration

Fig. 2 Identification of global differentially expressed genes in GBM. (a) t-SNE plots showing the GBM and GTEX data cluster. (b) Volcano plot of the differentially expressed genes in GBM versus normal brain tissues. Genes that are significantly dysregulated in GBM versus GTEx (log2 fold change |2|) were highlighted in red (downregulated) and green (upregulated)

Fig. 3 Integration of TCGA GBM transcriptomics, GBM tissues proteomics and cell lines proteomics data

Fig. 4 Prioritization of six high-confidence GBM surface marker genes. (a) Protein-protein interaction network analysis of the significantly upregulated GBM surfaceome genes. (b) Venn diagram showing the genes that are overlapped between the PPI network and transcriptomics-proteomics data integration analysis

Fig. 5 CD44 gene co-expressed network analysis. (a) CD44 gene co-expressed network with Pearson correlation value, $r > 0.75$. Nodes represent genes and edges are colored on a sliding scale according to the strength of the correlation (red, $r = 1.0$ and blue, $r = 0.75$). (b) Histograms of CD44 co-expression cluster from (a) showing the average expression of genes on GBM tumor (red bar) and normal (yellow bar)

SUPPLEMENTARY MATERIALS

Supplementary Tables

1. Supplementary Table S1. Overall differentially expressed genes in TCGA GBM tissues vs. GTEx normal brain tissues

2. Supplementary Table S2. Significantly dysregulated cell surface genes in TCGA GBM tissues vs. GTEx normal brain tissues

3. Supplementary Table S3. GBM cell lines proteomics data from Bausch-Fluck et al. 2015

4. Supplementary Table S4. GBM tissue samples proteomics data from Polisetty et. al 2012

5. Supplementary Table S5. Protein-protein interaction network analysis of surfaceome.

Supplementary Figures

1. Fig S1. Gene ontology and deregulated pathways in GBM. (A-B) Gene ontology cellular component of the significantly (A) upregulated and (B) downregulated genes in GBM. (c-d) KEGG pathway analysis of the (C) upregulated and (D) downregulated genes in GBM

2. Fig S2. Significant differentially expressed cell surface genes in GBM. (a) GBM surfaceome classification using previously annotated cell surface genes dataset identifies 395 DEGs that belongs to surfaceome. (b) Cell surface genes stratification from (a) based on its subclass

3. Fig S3. KEGG pathway analysis of differentially expressed surfaceome in GBM. (A) Upregulated surfaceome and (B) Downregulated surfaceome

4. Fig S4. Significant upregulation of the prioritized GBM surfaceome signature in GBM patients. (A-F) Boxplot showing the RNA-Seq data (transcript per million) of (A) CD44 (B) PTPRJ (C) SLC1A5 (D) EGFR (E) HLA-DRA and (F) ITGB2 in GBM and GTEx normal brain tissue samples

5. Fig S5. The prioritized GBM surfaceome signature as potential GBM prognostic biomarker. (a-f) Disease-free survival analysis of GBM patients having high and low expression of (a) CD44 (b) PTPRJ (c) SLC1A5 (d) EGFR (e) HLA-DRA and (f) ITGB2