Gene 824 (2022) 146395

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Research paper

Non-negative matrix factorization and differential expression analyses identify hub genes linked to progression and prognosis of glioblastoma multiforme

Sevinç Akçay [a], Emine Güven [b], Muhammad Afzal [c,*], Imran Kazmi [d]

[a] Department of Molecular Biology of Genetics, Kırşehir Ahi Evran University, Kırşehir, Turkey
[b] Department of Biomedical Engineering, Düzce University, Düzce, Turkey
[c] Department of Pharmacology, College of Pharmacy, Jouf University, Sakaka, AlJouf 72341, Saudi Arabia
[d] Department of Biochemistry, Faculty of Science, King Abdulaziz University, Jeddah, Saudi Arabia

ARTICLE INFO

Edited by: John Doe

Keywords:
Glioblastoma multiforme (GBM)
non-negative matrix factorization (NMF)
Metagenes
Glioblastoma stem cells (GSCs)
Differentially expressed genes (DEGs)

ABSTRACT

One of the most prevailing primary brain tumors in adult human male is glioblastoma multiforme (GBM), which is categorized by rapid cellular growth. Even though the combination therapy comprises surgery, chemotherapy, and adjuvant therapies, the survival rate, on average, is 14.6 months. Glioma stem cells (GSCs) have key roles in tumorigenesis, progression, and defiance against chemotherapy and radiotherapy. In our study, firstly, the gene expression dataset GSE124145 was retrieved; the non-negative matrix factorization (NMF) method was applied on GBM dataset, and differentially expressed genes analysis (DEGs) was performed. After which, overlapping genes between metagenes and DEGs were detected to examine the Gene Ontology (GO) categories in the biological process (BP) in the stemness of GBM. The common hub genes were used to construct protein–protein interaction (PPI) network and further GO, while Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway was utilized to pinpoint the real hub genes. The analysis of hub genes particular for the same GO categories demonstrated that specific hub genes triggered distinct features of the same biological processes. After utilizing GSE124145 and The Cancer Genome Atlas (TCGA) dataset for survival analysis, we screened five real hub genes: GUCA1A, RFC2, GNG11, MMP19, and NRG1, which are strongly associated with the progression and prognosis of GBM. The DEGs analysis revealed that all real hub genes were overexpressed in GBM and TCGA datasets, which further validates our results. The constructed study of PPI, GO, and KEGG pathway on common hub genes was performed. Finally, the KEGG pathways performed on the top 15 candidate hub genes (including six real hub genes) of the PPI network in the GBM gene expression dataset study found mitogen-activated protein kinase (Mapk) signaling pathway to be the most significant pathway. The rest of the hub genes reviewed throughout the analysis might be favorable targets for diagnosing and treating GBM and lower-grade gliomas.

1. Introduction

One of the most prevailing and highly malignant forms of brain tumors is Glioblastoma multiforme (GBM) or grade-IV glioma (Perry et al., 2009). The diagnosis of the GBM patients is very challenging, and the patient survival rate is 12–15 months even with combinational therapies (Davis, 2016). The current therapies are surgery, chemotherapy, and radiotherapy (Stupp et al., 2014). The low efficiency of all therapeutic methods necessitates identification of new therapeutic targets for GBM in recent years.

GBM is an extremely heterogeneous tumor at the pathological and cellular level (Dirks, 2008; Lai et al., 2011). Gene expression and cell

Abbreviations: GBM, Glioblastoma Multiforme; hGBM, Human GBM tissue; GSC_X01, Glioma Stem Cell of X01 cells; GSC_X03, Glioma Stem Cell of X03 cells; GBM_U251, Cell Line U-251 derived from a human malignant GBM; NMF, Non-negative Matrix Factorization; DEGs, Differentially Expressed Genes; DAVID, The Database for Annotation, Visualization and Integrated Discovery; KEGG, Kyoto Encyclopedia of Genes and Genomes; GEO, Gene Expression Omnibus; GO, Gene Ontology; MAPK, Mitogen-actived Protein Kinase; PPI, Protein-protein Interaction; BP, Biological Process; MF, Molecular Function; CC, Cellular Component; SNP, Single Nucleotide Polymorphism; NCK, Non-catalytic Region of Tyrosine Kinase Adaptor Protein.

* Corresponding author.
E-mail address: afzalgufran@ju.edu.sa (M. Afzal).

https://doi.org/10.1016/j.gene.2022.146395
Received 23 December 2021; Received in revised form 10 February 2022; Accepted 4 March 2022
Available online 11 March 2022

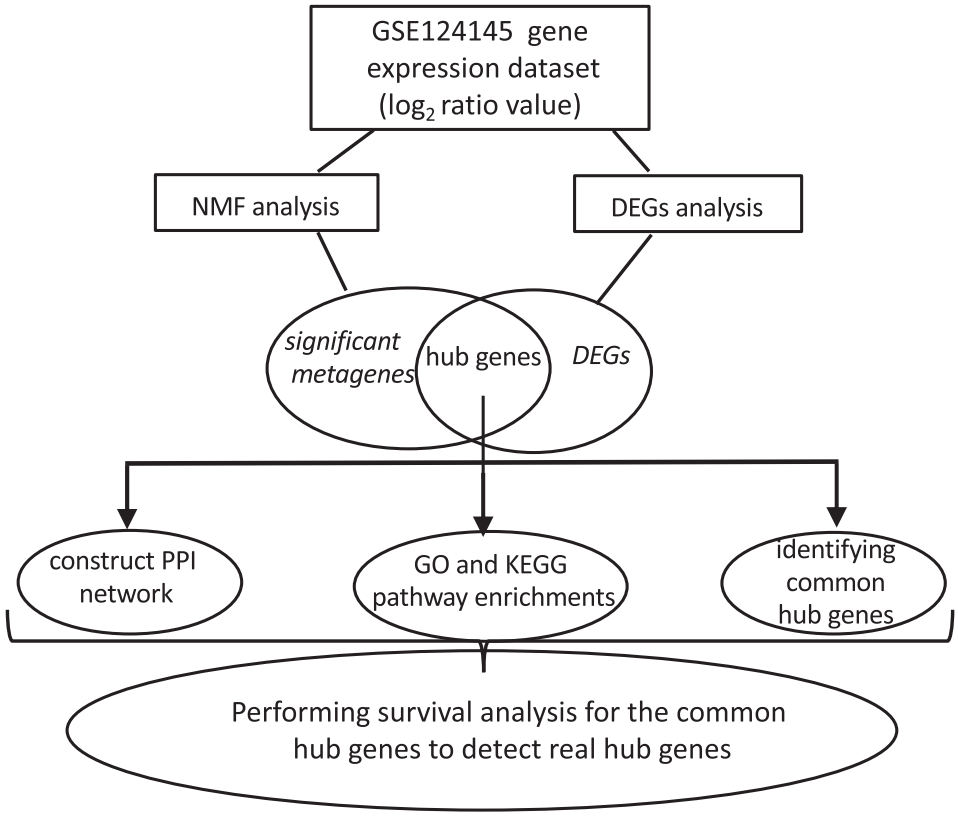

Fig. 1. Structure and workflow of the analysis steps. NMF, non-negative matrix factorization; DEGs, differentially expressed genes; PPI, protein–protein interaction.

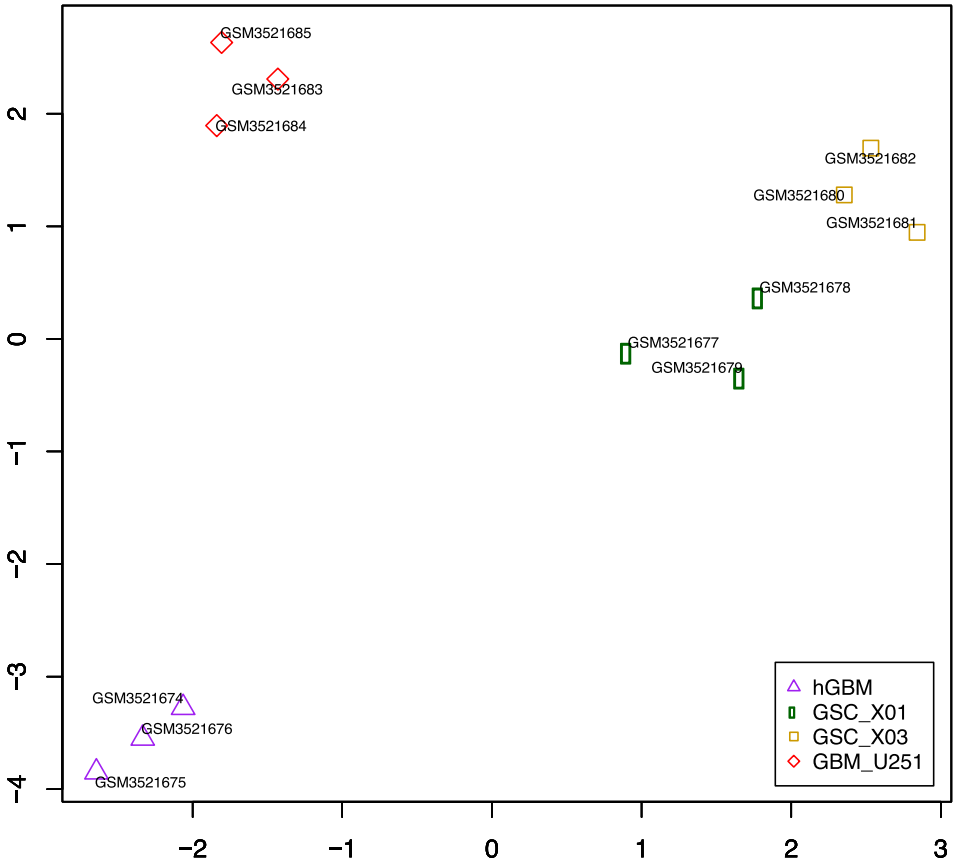

Fig. 2. UMAP of the GSE124145 gene expression values to study structure in high-dimensional datasets to present QC.

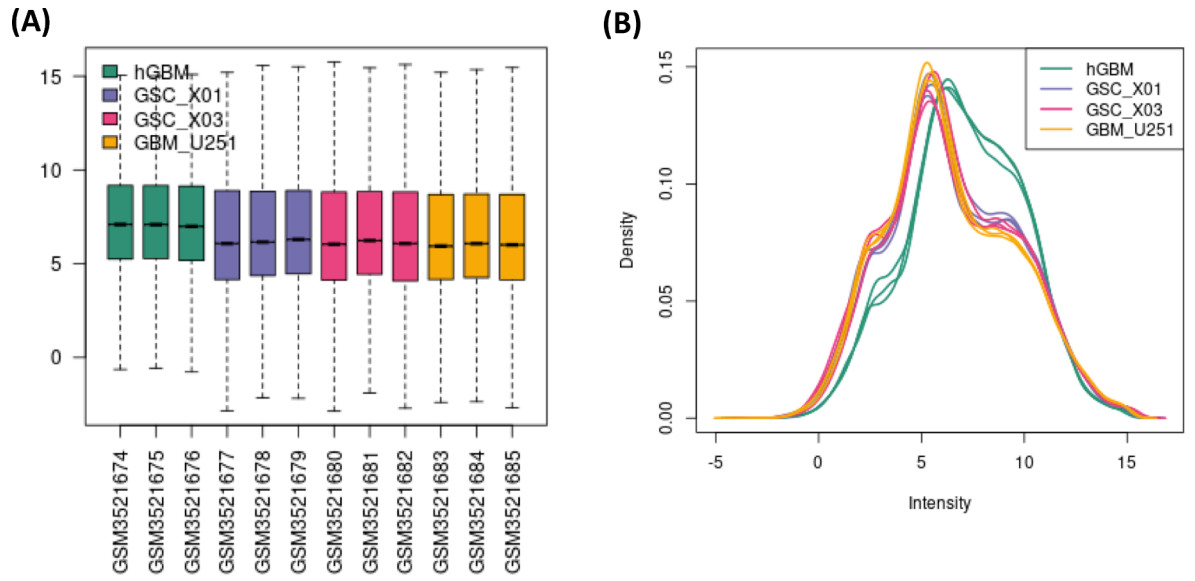

Fig. 3. (A) The boxplot and (B) the density plot of GSE124145 gene expression values of $\log_2$ base. Colors present the four clinical traits. hGBM: human glioblastoma tissue; GSC_X01 and GSC_X03: glioma stem cells; GBM_U251: glioma cell line 251.

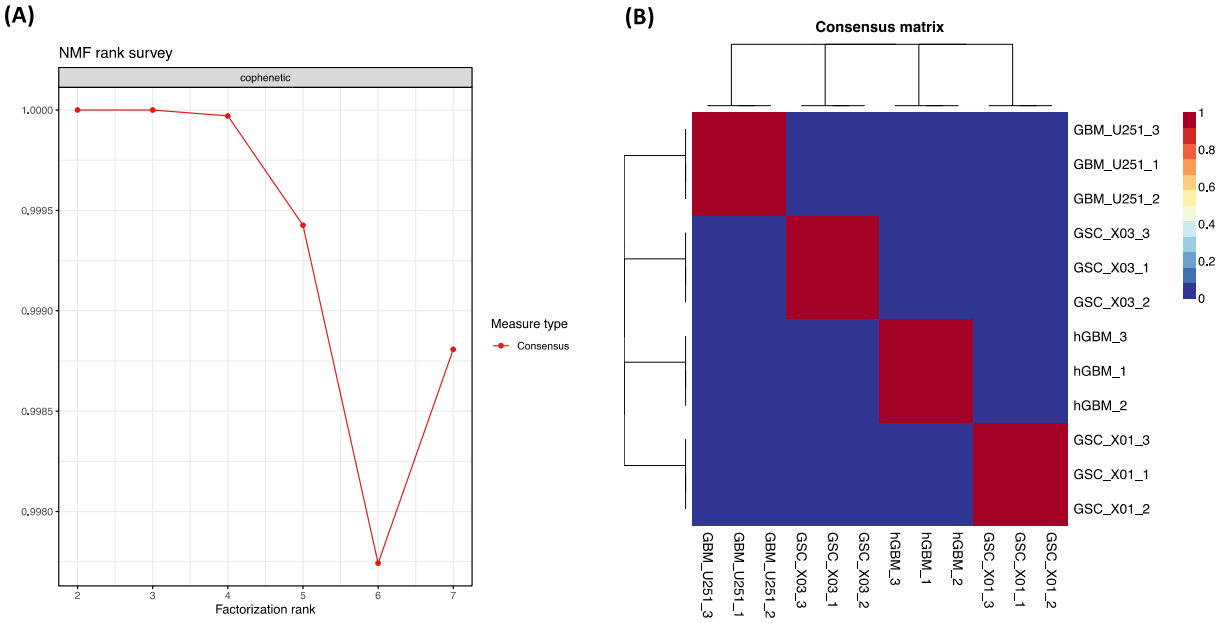

Fig. 4. (A) The cophenetic matrix plot identified factorization rank as four. (B) The consensus matrix shows the consensus rank r = 4. Pure block diagonal patterns indicate the robustness of models with 4 metagenes classes. NMF: non-negative matrix factorization; hGBM: human glioblastoma tissue; GSC_X01 and GSC_X03: glioma stem cells; GBM_U251: glioma cell line U251.

proliferation levels are also highly differing in GBM (Wu et al., 2021). Glioma stem cells (GSCs) take a central position regarding tumor formation of lower-grade gliomas and glioblastoma multiforme. GSCs have important characteristics including self-renewal ability, tumor initiation, progression ability, and resistance to GBM therapies. Several important roles of GSCs in GBM make GSCs new therapeutic targets (Lan et al., 2017; Li et al., 2019). Hence, there is an urgent need to discover new biomarkers for GBM and lower-grade astrocytomas.

Existing research has indicated that malignancy cannot be initiated by only one gene, trait, or affect. It must be a linkage of distinct genes and biological, functional, and cellular pathways organized together. Non-negative matrix factorization (NMF) is a methodology used to analyze and group the genes with associated expression motives into the same co-expression and the genes with different expression motives into different clusters. Multiple studies have indicated that NMF can be used to study genes, consensus of clustering in different cells, tissues, and cell lines (Cheng et al., 2019; Collisson et al., 2011). Furthermore, differentially expressed genes (DEGs) analysis method has been utilized in gene expression data gatherings of comparisons for multi-group data (Tang et al., 2015).

DNA microarray innovation is an influential instrument to identify the gene expression profile of thousands of genes instantaneously. The large datasets generated by microarray technology need to be analyzed and interpreted to discover the biological functions of genes and the biological mechanism of diseases. The high dimensionality of large datasets generated by microarray data needs to be reduced for visualization and clustering (Hatfield et al., 2003). Clustering analysis is very helpful in understanding unknown gene-gene relationships (Vidman

Table 1
GO enrichment analysis of categories in genes screened by 'brunet' on the GSE124145 gene expression dataset showing the top significant GO annotations of BP, CC, and MF for each metagenes (p-value < 0.05).

Metagenes	Biological Process	Number of genes	P-value	FDR
Metagenes 1	GO:0001894 ~ tissue homeostasis	5	7.61E-04	3.79E-32
(57 genes)	GO:0048873 ~ homeostasis of number of cells within a tissue	3	0.00204658	3.21E-31
	GO:0048871 ~ multicellular organismal homeostasis	5	0.00423938	8.77E-27
	GO:0060249 ~ anatomical structure homeostasis	5	0.00613093	3.61E-26
	GO:0042592 ~ homeostatic process	9	0.01859239	3.54E-25
Metagenes 2	GO:0048705 ~ skeletal system morphogenesis	5	0.00192453	5.93E-23
(80 genes)	GO:0009887 ~ organ morphogenesis	9	0.00297594	1.97E-22
	GO:0048706 ~ embryonic skeletal system development	4	0.00366075	1.97E-22
	GO:0048732 ~ gland development	6	0.00397585	1.41E-21
	GO:0007389 ~ pattern specification process	6	0.00461076	3.06E-21
Metagenes 3	GO:0006811 ~ ion transport	29	4.05E-05	4.78E-10
(229 genes)	GO:0006928 ~ movement of cell or subcellular component	33	4.44E-05	5.14E-10
	GO:0030029 ~ actin filament-based process	18	6.50E-05	6.15E-10
	GO:0030001 ~ metal ion transport	20	6.94E-05	8.69E-09
	GO:0071804 ~ cellular potassium ion transport	9	1.77E-04	9.65E-09
Metagenes 4	GO:0071495 ~ cellular response to endogenous stimulus	17	1.12E-05	1.06E-08
(84 genes)	GO:0010646 ~ regulation of cell communication	28	1.31E-05	1.06E-07
	GO:0023051 ~ regulation of signaling	28	1.79E-05	1.31E-07
	GO:0009966 ~ regulation of signal transduction	26	2.15E-05	2.53E-07
	GO:0009719 ~ response to endogenous stimulus	19	2.18E-05	3.05E-07
Metagenes	**Cellular Component**	**Number of genes**	**P-value**	**FDR**
Metagenes 1	GO:0031235~intrinsic component of the cytoplasmic side of the plasma membrane	2	2.13E-04	1.79E-10
(57 genes)	GO:0044421~extracellular region part	13	0.0035748	3.21E-9
Metagenes 2	GO:0019898~extrinsic component of membrane	4	2.78E-03	4.26E-9
(80 genes)	GO:0005834~heterotrimeric G-protein complex	2	0.00461076	2.06E-7
Metagenes 3	GO:0005887~integral component of plasma membrane	33	2.77E-05	1.68E-09
(229 genes)	GO:0031226~intrinsic component of plasma membrane	34	2.86E-05	1.59E-09
	GO:0009986~cell surface	22	3.40E-04	1.05E-09
	GO:0098978~glutamatergic synapse	12	4.98E-04	3.67E-07
	GO:0045202~synapse	26	6.68E-04	12.31E-07
Metagenes 4	GO:0098590~plasma membrane region	11	1.14E-05	0.00181278
(84 genes)	GO:0098552~side of membrane	7	8.97E-05	0.00536928
	GO:0009986~cell surface	9	1.15E-04	0.00616542
	GO:0009897~external side of plasma membrane	5	6.4E-03	0.01228229
	GO:0044432~endoplasmic reticulum part	10	1.2820512	0.02532359
Metagenes	**Molecular Function**	**Number of genes**	**P-value**	**FDR**
Metagenes 1	GO:0005125~cytokine activity	3	6.52E-04	0.01258421
(57 genes)				
Metagenes 2	GO:0035612~AP-2 adaptor complex binding	2	3.5E-7	0.00184604
(80 genes)	GO:0003712~transcription cofactor activity	5	8.77E-6	0.00382706
	GO:0000989~transcription factor activity, transcription factor binding	5	7.76E-5	0.00509563
	GO:0000988~transcription factor activity, protein binding	5	7.19E-4	0.00524945
	GO:0003700~transcription factor activity, sequence-specific DNA binding	7	1.28E-3	0.00624061
Metagenes 3	GO:0005216~ion channel activity	13	6.63E-7	3.04E-04
(229 genes)	GO:0022838~substrate-specific channel activity	13	5.36E-06	4.07E-04
	GO:0005261~cation channel activity	11	3.63E-05	4.97E-04
	GO:0051015~actin filament binding	9	4.19E-05	6.10E-04
	GO:0032394~MHC class Ib receptor activity	3	1.53E-04	7.13E-04
Metagenes 4	GO:0005086~ARF guanyl-nucleotide exchange factor activity	3	3.84E-04	0.00446343
(84 genes)	GO:0005085~guanyl-nucleotide exchange factor activity	6	7.69E-04	0.00774728
	GO:0050839~cell adhesion molecule binding	7	8.97E-04	0.01016688
	GO:0003779~actin binding	6	7.69E-03	0.02121416
	GO:0052813~phosphatidylinositol bisphosphate kinase activity	3	3.84615385	0.02793769

et al., 2019). The purpose of clustering is to categorize the genes with associated expression patterns into the same cluster and the genes with different expression patterns into distinct clusters. Gene expression data from the microarray are used to cluster genes with different methods, including self-organizing maps (SOM) (Tamayo et al., 1999), hierarchical clustering (Eisen et al., 1998), principal component analysis (PCA), and k-means (MacQueen, 1967). In the hierarchical clustering method, first genes with identical expression profiles are organized; they then form clustering trees (Jamail and Moussa, 2020). The limitations of hierarchical clustering are the high possibility of forming an inflexible clustering tree and sensitivity to similarity metrics (Brunet et al., 2004). The genes are divided into the predetermined number (k) of clusters in the k-means supervised model (Tavazoie et al., 1999). In the SOM clustering method, the dimensionality of the data is reduced and clusters gene with similar expression patterns. Nevertheless, these statistical methods have several restraints. Firstly, because these clustering methods mostly target dominant structures, alternative structures might be unnoticed in the dataset. Secondly, the grouping of genes was made based on similarities in their expression profiles. These two limitations affect the correct interpretation of huge datasets.

Recently, different bi-clustering techniques have been developed to overcome the limitations of the traditional clustering methods mentioned above. Independent component analysis (ICA), principal component analysis (PCA), and NMF are popular bi-clustering techniques that can allow simultaneous grouping of genes regulated under different conditions (Tsai and Chiu, 2010; Turner et al., 2005). The

Table 2
GO enrichment analysis of categories in genes screened by 'brunet' on the GSE124145 gene expression dataset showing the top five significant KEGG pathways for all the metagenes (p-value < 0.05).

Metagenes	KEGG pathways	Number of genes	P-value	FDR
Metagenes 1 (57 genes)	hsa04724: Glutamatergic synapse	6	3.06E-05	0.00337782
	hsa04611:Platelet activation	6	2.61E-04	0.00483749
	hsa04723:Retrograde endocannabinoid signaling	6	3.66E-04	0.01007856
	hsa04666:Fc gamma R-mediated phagocytosis	5	9.55E-04	0.01076968
	hsa04961:Endocrine and other factor-regulated calcium reabsorption	4	2.04E-05	0.01155311
Metagenes 2 (80 genes)	hsa04720:Long-term potentiation	4	3.23E-05	0.00216310
	hsa05031: Amphetamine addiction	4	4.76E-04	0.00233627
	hsa04650:Natural killer cell mediated cytotoxicity	5	6.55E-04	0.00256852
	hsa04728: Dopaminergic synapse	5	7.96E-04	0.00298064
	hsa04971:Gastric acid secretion	4	2.04E-03	0.00300012
Metagenes 3 (229 genes)	hsa04713:Circadian entrainment	7	2.11E-04	0.00153095
	hsa04022:cGMP-PKG signaling pathway	8	6.77E-04	0.00194994
	hsa04610:Complement and coagulation cascades	6	9.21E-04	0.00256897
	hsa04270:Vascular smooth muscle contraction	7	0.00118562	0.00675934
	hsa04261:Adrenergic signaling in cardiomyocytes	7	0.00211191	0.00705379
Metagenes 4 (84 genes)	hsa05214:Glioma	4	0.00222855	0.00227312
	hsa04144:Endocytosis	5	0.0155247	0.00500242
	hsa04068:FoxO signaling pathway	4	0.01659162	0.00613771
	hsa05200:Pathways in cancer	6	0.01961737	0.00844578
	hsa05210:Colorectal cancer	3	0.02599431	0.00930283

GO, Gene ontology; BP, biological function; MF, molecular function; CC, cellular component; KEGG, Kyoto Encyclopedia of Genes and Genomes.

Table 3
The number of down and up-regulated DEGs by paired features.

Features Compared	Down-Regulated DEGs	Up-Regulated DEGs
hGBM-GSC_X01	12	410
hGBM-GSC_X03	16	429
hGBM-GBM_U251	24	544
GSC_X01-GSC_X03	41	40
GSC_X01-GBM_U251	122	123
GSC_X03-GBM_U251	112	112

application of PCA depends on the linearity assumption (Swain et al., 2021). Another important disadvantage of PCA is that the positive and negative coefficients of ICA and PCA vectors might include the positive and negative values. Negative values may complicate the interpretation of a gene with a negative expression. Hence, it would be better to restrict the factors and coefficients to a non-negative setting. NMF is a bi-clustering technique that uses two non-negative matrices and is a great alternative to PCA and ICA. NMF is first used in image recognition (Lee and Seung, 1999). In recent years, NMF has become a very effective method in biomedical sciences, including metabolomics, proteomics, gene expression analysis, and sequencing analysis (Frigyesi and Höglund, 2008; Gaujoux et al., 2020; Jiang et al., 2019; Zhang et al., 2012). NMF reduces the dimension of large gene expression datasets from thousands of genes to metagenes. NMF method became popular in recent years because it is less sensitive to select genes and identify several different gene expression patterns. NMF is used in several computational biology applications, including biomedical informatics, molecular pattern discovery, class comparison, cross-platform characterization, and analyzing functional heterogeneity of genes (Devarajan, 2008). Furthermore, to gain deep knowledge of gene expression data, the DEGs analysis approach has been used over the last decade (Rau et al., 2019).

A recent study (Sakamoto et al., 2019) discovered differences in gene expression among GSCs, GBM tissue, and U251 cell line that was derived from a malignant glioblastoma tumor by explant technique (Berens et al., 1994) using PCA with factor loading, intracellular pathway analysis, and immunopathway analysis. Sakamoto et al., 2019 further demonstrated that MYCN, DPP4 and MIF are the important contributors of GSCs and deposited the microarray data in the Gene Expression Omnibus (GEO) database with the GEO accession number GSE123145.

Through NMF, the genes' behaviors are detected; this makes it easier to detect genes specific to one tumor type that are not detected by other clustering methods like PCA. One previous study showed that NMF is a helpful tool to have biological information from the microarray dataset and to understand tumor behaviors (Frigyesi and Höglund, 2008). In addition, another study showed that while NMF was a more effective tool for identifying deeper information of genes, PCA could not detect the important information (Boccarelli et al., 2018). Thus, in the present study, our aim is to analyze the GSE124145 dataset by NMF in order to get biologically relevant information and gain more insights into the pathobiology of GBM.

In our study, we first downloaded the GSE124145 dataset from the publicly available GEO database. To gain an additional understanding of the progression and prognosis of GBM, we utilized the NMF algorithm to detect real hub genes accompanied by clinical traits of the DEGs and we performed DEGs analysis. In addition, we detected overlapping genes between metagenes and DEGs, and we examined the GO categories in the biological processes (BP) in the stemness of GBM. We used the common hub genes to construct the PPI network and further GO, and the KEGG pathway was utilized to pinpoint the real hub genes. Finally, we performed survival analysis by utilizing GSE124145 and The Cancer Genome Atlas (TCGA) dataset to discover real hub genes.

2. Materials & methods

2.1. Processing of the microarray data

Microarray data for human glioblastoma and glioma stem cells were retrieved from NIH Gene Expression Omnibus (GEO) (Sakamoto et al., 2019) by typing in the search box the word "glioma" on the GEO database. The GSE124145 gene expression dataset includes total RNAs from the human glioblastoma multiforme tissues (hGBM), the human glioma stem cell lines X01 (GSC_X01), human glioma stem cell lines X03 (GSC_X03), and glioma cell line U251 (U251) from direct tumor resection of a 54-year-old female patient. Microarray data contains 12 samples such that GSM35221674 hGBM rep1, GSM35221675 hGBM rep2, GSM35221676 hGBM rep3; GSM352211677 GSCs X01 rep1, GSM352211678 GSCs X01 rep2, GSM352211679 GSCs X01 rep3; GSM352211680 GSCs X03 rep1, GSM352211681 GSCs X03 rep2, GSM352211682 GSCs X03 rep3; GSM352211683 U251 rep1, GSM352211684 U251 rep2, GSM352211685 U251 rep3.

Genomic data from cells, cell lines, and tissues of GBM gene expression were collected. All probe sets were converted to gene

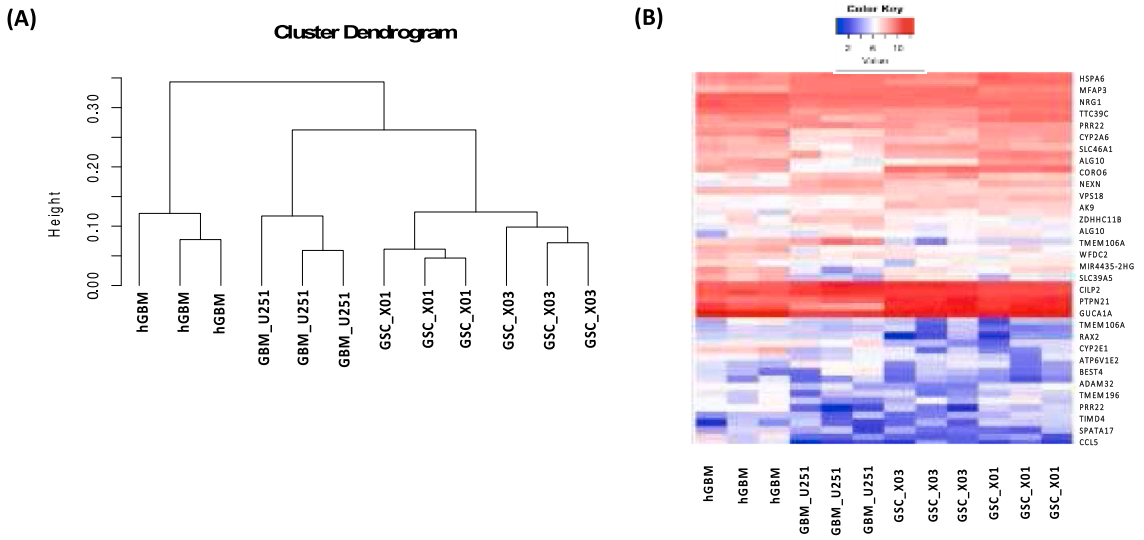

Fig. 5. (A) A cluster dendrogram of the four groups (clusters) of features by the candidate hub genes expression levels. (B) A heatmap of expression levels of 52 hub genes by four features.

symbols to the probe annotation files of the GPL570 platform, and gene expression levels were $\log_2$ transformed. This study has performed the quality control of the dataset via Uniform Manifold Approximation and Projection (UMAP), which is a dimension reduction method useful for visualizing clusters or groups of samples and relative proximities. The number of nearest neighbors used in the calculation is indicated in the plot (Konopka and Konopka, 2018).

The GEOquery package in Bioconductor is used to analyze GSE124145 dataset. The list of packages is Biobase, biomaRT, UMAP, and gplots packages in R studio (Davis and Meltzer, 2007; Durinck et al., 2005; Konopka and Konopka, 2018; Warnes et al., 2009). Benjamini-Hochberg technique is used to calculate the adjusted p-value, avoid Type I errors, and correct multiple testing. A hypergeometric model was performed for both the down-regulated and up-regulated DEGs in DAVID GO enrichment in categories and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway enrichment analysis (Huang et al., 2007; Kanehisa et al., 2016). Moreover, adjusting the statistical tests locally is done by calculation of a false discovery rate (FDR) (Benjamini and Hochberg, 1995; Dudoit et al., 2003; Hochberg and Tamhane, 1987).

2.2. Gene expression analysis

Most of the analyses were achieved in the R programming version 3.6.3. The design of the study is given in Fig. 1. This study aims to demonstrate the capability of NMF to uncover expressive biological evidence from malignant brain tumor RNA-microarray data. The novel application of the NMF bioinformatics pipeline was an effective method to elucidate commonalities and discrepancies between samples of the dataset. Differentially expressed genes are screened according to both fold change and $p-value$ criteria. Although methods to correct for multiple comparisons, such as Bonferroni correction, have long been applied, most of these techniques are not suitable for analyzing gene expression datasets (Tarca et al., 2006).

2.3. Non-negative matrix factorization

Given a target matrix V^{mxn}, NMF identifies non-negative matrices such that N^{mxr} and M^{rxn} (i.e., with all entries ≥ 0) to present the matrix decomposition as:

$V \approx NM$ #1

In practice, this study typically viewed N as a basis or metagenes matrix. The rank factorization is selected on the constraint $r \leq \min(m, n)$. The purpose of this selection is to explain and distinguish the details classified among V into r factors: the columns of N. Given a matrix $V^{m\times n}$, NMF finds two non-negative matrices $N^{m\times r}$ and $M^{r\times n}$(i.e., with all elements ≥ 0) to represent the decomposed matrix as:

$V \approx NM$,

For instance, by natural demanding of non-negative N and M to minimize the cost function:

$$\|V - NM\|_F, \text{ subject to } N \geq 0, M \geq 0$$

It is considered a gene expression dataset characterized by the expression levels of *m* genes (probes) in *n* samples of unique tissues, cells, cell lines, time points, or experiments. The number m of genes is usually from hundreds to thousands, and the n of experiments or patients is usually 100 for gene expression research. The gene expression datasets are presented by a matrix of expression *V* of size *NxM*, whose rows consist of m genes, expression levels, and columns of *n* samples. The goal is to factorize a small number of rows, i.e. rank, each defined as a positive linear combination of the target matrix *V*. The positive linear combination of metagenesis is described by the gene expression motif of the samples. To obtain a dimensional reduction of the microarray data and to evaluate the distinctions among samples, NMF was implemented utilizing the "NMF" package (Gaujoux and Seoighe, 2010) in R.

A critical issue is the decision of the factorization rank- *r*, which defines the number of metagenes used to approximate the target matrix. One of the most standard approaches is to process the cophenetic correlation coefficient. Brunet et al. proposed to choose the smallest *r* value at which this coefficient starts to decrease (Brunet et al., 2004). This study interpreted the *r* value as metagene profiles capturing gene expression patterns particular to different clinical traits.

Further, searching for genes with reasonably large coefficients in each biological and functional process may provide some benefits, assuming that additional genes can partake in more than one biological process. The *gene.score* scoring technique contributed by Kim and Park (Kim and Park, 2007) has been implemented to achieve this assumption. The most metagene-specific genes were extracted utilizing Kim and Park's scoring and screening approach. (Esposito et al., 2020; Ramanarayanan et al., 2011).

2.4. Identification of DEGs and clustering analysis of GSE124145

Gene expression values were extracted via the GEOquery package in Bioconductor (Davis and Meltzer, 2007) then converted to a base-2

Table 4
The GO and KEGG pathway enrichments of the common hub genes.

Category	Term	Genes	P-value	FDR
GOTERM_BP_FAT	GO:0007049 ~ cell cycle	PAX8, THRA, MAPK1	0.00099	0.003763553
GOTERM_BP_FAT	GO:0031663 ~ lipopolysaccharide-mediated signaling pathway	SCARB1, CCL5, MAPK1	0.00350	0.003275988
GOTERM_BP_FAT	GO:0048732 ~ gland development	DDR1, PAX8, THRA, MAPK1, EPHB3	0.00656	0.003376552
GOTERM_BP_FAT	GO:0042592 ~ homeostatic process	SCARB1, SLC46A1, THRA, RFC2, CCL5, MAPK1, SLC39A5, SLC39A13, ATP6V1E2	0.00719	0.003375988
GOTERM_BP_FAT	GO:0046916 ~ cellular transition metal ion homeostasis	SLC46A1, SLC39A5, SLC39A13	0.00931	0.025184143
GOTERM_BP_FAT	GO:0055076 ~ transition metal ion homeostasis	SLC46A1, SLC39A5, SLC39A13	0.01697	0.01002087
GOTERM_BP_FAT	GO:0000041 ~ transition metal ion transport	SLC39A5, SLC39A13, ATP6V1E2	0.01755	0.01230103
GOTERM_BP_FAT	GO:0030003 ~ cellular cation homeostasis	SLC46A1, CCL5, SLC39A5, SLC39A13, ATP6V1E2	0.01871	0.02056453
GOTERM_BP_FAT	GO:0035295 ~ tube development	DDR1, PAX8, THRA, MAPK1, EPHB3	0.01947	0.02076576
GOTERM_BP_FAT	GO:0006873 ~ cellular ion homeostasis	SLC46A1, CCL5, SLC39A5, SLC39A13, ATP6V1E2	0.02037	0.02077453
GOTERM_MF_FAT	GO:0072341 ~ modified amino acid binding	SCARB1, SLC46A1, TIMD4	0.01030	0.00109775
GOTERM_MF_FAT	GO:0005524 ~ ATP binding	DDR1, PXK, UBA7, RFC2, HSPA6, MAPK1, TTLL12, EPHB3, GNG11	0.01378	0.00139519
GOTERM_MF_FAT	GO:0032559 ~ adenyl ribonucleotide binding	DDR1, PXK, UBA7, RFC2, HSPA6, MAPK1, TTLL12, EPHB3, GNG11	0.01576	0.00146491
GOTERM_MF_FAT	GO:0008392 ~ arachidonic acid epoxygenase activity	CYP2A6, CYP2E1	0.03246	0.00175023
GOTERM_MF_FAT	GO:0008391 ~ arachidonic acid monooxygenase activity	CYP2A6, CYP2E1	0.03246	0.00177237
GOTERM_MF_FAT	GO:0035639 ~ purine ribonucleoside triphosphate binding	DDR1, PXK, UBA7, RFC2, HSPA6, MAPK1, TTLL12, EPHB3, GNG11	0.04234	0.00185748
GOTERM_MF_FAT	GO:0032550 ~ purine ribonucleoside binding	DDR1, PXK, UBA7, RFC2, HSPA6, MAPK1, TTLL12, EPHB3, GNG11	0.04341	0.00196678
GOTERM_MF_FAT	GO:0001883 ~ purine nucleoside binding	DDR1, PXK, UBA7, RFC2, HSPA6, MAPK1, TTLL12, EPHB3, GNG11	0.04377	0.00206671
GOTERM_CC_FAT	GO:0031253 ~ cell projection membrane	SCARB1, SLC46A1, GUCA1A	0.01161	0.00241032
GOTERM_CC_FAT	GO:1903561 ~ extracellular vesicle	DDR1, SCARB1, HSPA6, MAPK1, SLC39A5, CILP2, WFDC2, TMEM106A	0.03171	0.00335669
GOTERM_CC_FAT	GO:0005789 ~ endoplasmic reticulum membrane	CYP2A6, CYP2E1, ALG10, PIGX	0.03216	0.00427774
GOTERM_CC_FAT	GO:0043230 ~ extracellular organelle	DDR1, SCARB1, HSPA6, MAPK1, SLC39A5, CILP2, WFDC2, TMEM106A	0.32906	0.00543521
GOTERM_CC_FAT	GO:0042175 ~ nuclear outer membrane-endoplasmic reticulum membrane network	CYP2A6, CYP2E1, ALG10, PIGX	0.03339	0.00574228
GOTERM_CC_FAT	GO:0044432 ~ endoplasmic reticulum part	CYP2A6, CYP2E1, ALG10, PIGX	0.04210	0.00755629
GOTERM_CC_FAT	GO:0031226 ~ intrinsic component of plasma membrane	DDR1, SCARB1, BEST4, SLC39A5, EPHB3	0.04461	0.00887441
GOTERM_CC_FAT	GO:0005576 ~ extracellular region	DDR1, SCARB1, CCL5, MFAP3, HSPA6, MAPK1, SLC39A5,CILP2, WFDC2, TMEM106A, EPHB3	0.04475	0.00915997
GOTERM_CC_FAT	GO:0005576 ~ extracellular region	DDR1, SCARB1, CCL5, MFAP3, HSPA6, MAPK1, SLC39A5, CILP2, WFDC2, TMEM106A, EPHB3	0.04475	0.00925895
GOTERM_CC_FAT	GO:0005783 ~ endoplasmic reticulum	CYP2A6, CYP2E1, ALG10, PIGX	0.04595	0.00934579
GOTERM_CC_FAT	GO:0044421 ~ extracellular region part	DDR1, SCARB1, CCL5, HSPA6, MAPK1, SLC39A5, CILP2, WFDC2, TMEM106A	0.04627	0.00945447
KEGG_PATHWAY	hsa05164:Influenza A	CCL5, HSPA6, MAPK1	0.05383	0.01545990
KEGG_PATHWAY	hsa05216:Thyroid cancer	PAX8, MAPK1	0.06146	0.02600558
KEGG_PATHWAY	hsa05020:Prion diseases	PRNP, CCL5, MAPK1	0.07170	0.39374634

GO, Gene ontology; BP, biological function; MF, molecular function; CC, cellular component; KEGG, Kyoto Encyclopedia of Genes and Genomes

logarithmic scale in R. Normalization is done by dividing expression values by the sum of all expression values of the given array (Quackenbush, 2002). Clustering analysis of DEGs was done using the hclust() function (R Core Team, 2020) to associate the expression pattern of DEGs in each pairwise traits; hGBM – GSC_X01, hGBM – GSC_X03, hGBM - GBM_U251, GSC_X01- GSC_X03, GSC_X01- GBM_U251 and GSC_X03 – GBM_U251.

The experiment is designed to focus on all of the pairwise comparisons since we are interested in spotting DEGs in most tractable way. We do the pairwise comparisons between each trait by fixing one as the control group and then check for overlaps. We emphasize the statistical significance utilizing *t*-test under the threshold $p-value < 0.05$ and $|log_2(FC)| > 6.5$ to screen up-regulated and down-regulated DEGs. The study used fold change threshold value in log_2 scale between median and mean of the expression values of samples as provided in Table S1 to identify up-regulated and down-regulated DEGs.

2.5. GO terms and KEGG pathway analysis

Biomart package in R is used to convert probe-IDs of common hub genes to the gene symbols and names. The common genes were portrayed by their BP, molecular functions (MF), and cellular components (CC) with GO of the database for Annotation and DAVID, which stands for Visualization and Integrated Discovery (Huang et al., 2009). All categorized genes were carefully studied, and other parts like the annotation types, Universal Protein resource, and GO-terms in FAT were chosen to utilize DAVID and KEGG (Kanehisa et al., 2016).

2.6. PPI network

We built the PPI network of the common hub genes. NetworkAnalyst, reachable on the web, offers an exploration of PPI networks for particular genes utilizing STRING Interactome (Szklarczyk et al., 2016; Zhou et al., 2019). To broadly uncover the regulatory and molecular mechanisms in candidate hub genes, total RNAs is simply grouped into the hGBM and GSC_X01, GSC_X03, and GBM_U251 features. The DEGs from hGBM – GSC_X01, hGBM – GSC_X03, hGBM -GBM_U251, GSC_X01-GSC_X03, GSC_X01- GBM_U251, and GSC_X03 – GBM_U251 groups were studied to construct a PPI network with formerly narrated GO classification and enrichments. *Function Explorer* of NetworkAnalyst is utilized to implement functional enrichment analysis for the common hub nodes, which are specifically highlighted nodes via KEGG pathway databases that would result in the most significant pathways enrichment. A hypergeometric test is utilized to calculate the enrichment p-value < 0.05,

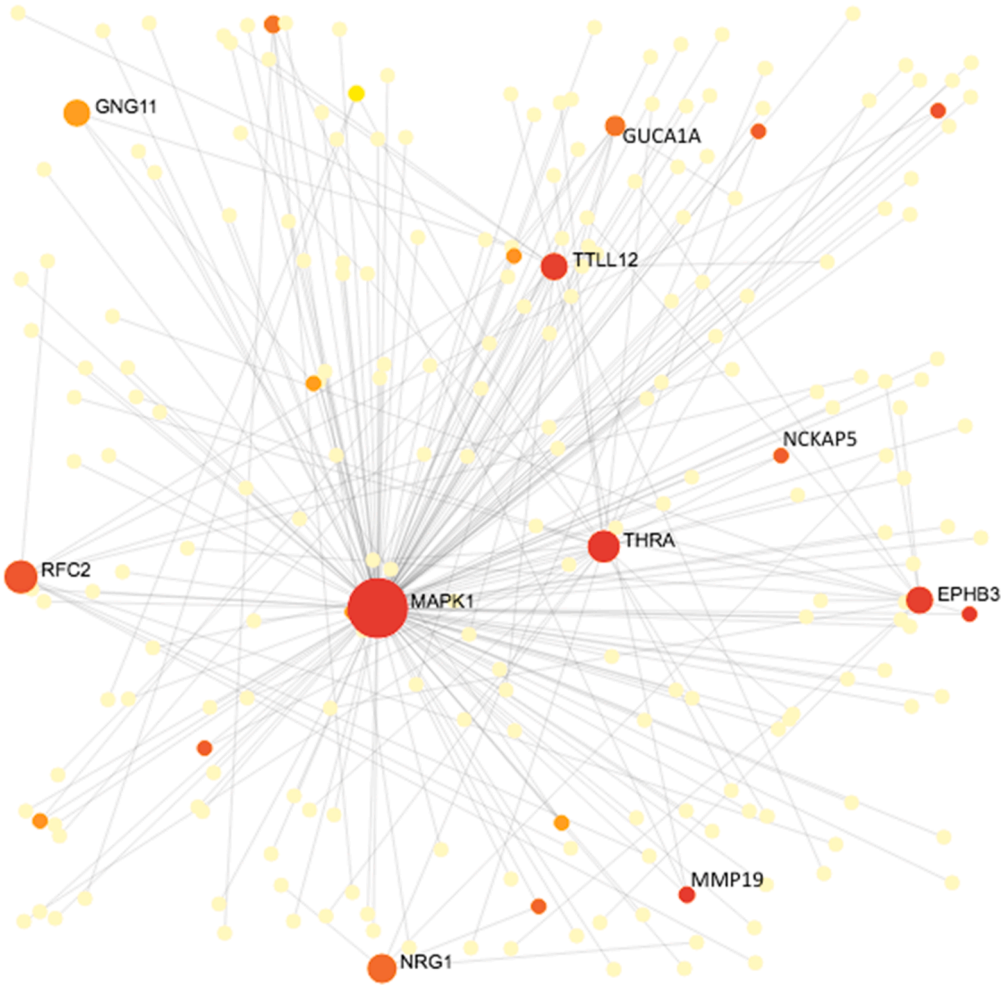

Fig. 6. A human PPI network for the common hub genes of the GSCs and GBM gene expression dataset.

Table 5
The top 15 common hub genes of PPI network of GBM gene expression dataset.

Gene ID	Genes	Nodes	Betweenness centrality	Expression	Fold Change
5594	MAPK1	152	28928.5	8.512	7.57
3310	HSPA6	29	6770.5	7.648	6.59
7067	THRA	23	4751.5	7.985	7.75
23,170	TTLL12	14	3133	10.631	8.53
2049	EPHB3	14	3133	8.666	8.59
5982	RFC2	9	1948	10.878	6.52
2978	GUCA1A	7	1980	3.988	9.23
11,099	PTPN21	6	739.5	5.113	10.78
7316	UBC	4	853	6.176	5.67
3084	NRG1	4	738	10.559	8.64
58,223	MMP19	3	493	11.453	6.95
2791	GNG11	3	1980	3.988	9.23
5573	NCKAP5	2	2151	8.273	8.95
3320	HSP90AA1	2	1980	2.950	5.23
2099	ESR1	2	1575	1.967	4.57

in which the complete corresponded proteins are also demonstrated.

2.7. Validation of common genes

This study selects the common hub genes (intersection of the meta-genes and DEGs sets) for validation and further analysis. To examine the portion of common hub genes in the stemness of GBM, the positively correlated genes in TCGA of the gene expression dataset through UALCAN database (Chandrashekar et al., 2017) at significance level (log rank p-value < 0.05) is studied. For deep analysis and validation, the log-rank test was employed to measure the survival analysis of GBM patients and the significance of the survival effect (Park, 2005).

3. Results

3.1. Quality control of gene expression data

For quality control (QC) of the GSE124145 gene expression dataset, UMAP, a dimension reduction method useful for visualizing clusters or groups of samples and relative proximities is used (Fig. 2). In Figure 3A and 3B, a boxplot of the non-normalized gene expression values and a density plot are presented with the groups (hGBM, GSC_X01, GSC_X03, and GBM_U251) of tissues, stem cells, and cell lines, respectively. The density plot Fig. 3B complements boxplot Fig. 3A in checking for data normalization before differential expression (DE) analysis.

3.2. The NMF analysis on GSE124145 gene expression dataset

In Fig. 4A, the function of factorization rank of r = 2,3,..,7 is selected as the emergence rank of the RSS survey. The optimum rank is detected by NMF at r = 4. In Fig. 4B, the consensus matrix plot validated the factorization rank as four, which describes the number of bases. The rank value affects the metagenes matrix defined by Brunet et al. 2007. Clear block diagonal patterns confirm the robustness of models with four metagenes modules, whereas a rank-5,6,7 factorization displays

Table 6
The KEGG pathways of the top 15 candidate hub genes of PPI network in GBM gene expression dataset.

Term	KEGG description	Total	Expected	Hits	P-Value	FDR
hsa04010	Mapk signaling pathway	135	0.293	8	3.94E-09	9.15E-06
hsa04727	Neurotrophin signaling pathway	119	0.261	7	2.92E-09	9.29E-07
hsa04725	Cholinergic synapse	112	0.246	6	8.74E-08	9.27E-06
hsa04728	Dopaminergic synapse	131	0.288	6	2.23E-07	1.52E-05
hsa04720	Long-term potentiation	67	0.147	5	2.39E-07	1.52E-05
hsa05214	Glioma	75	0.165	5	4.23E-07	2.24E-05
hsa04925	Aldosterone synthesis and secretion	98	0.215	5	1.61E-06	7.33E-05
hsa05205	Proteoglycans in cancer	201	0.442	6	2.78E-06	0.000111
hsa04114	Oocyte meiosis	125	0.275	5	5.38E-06	0.00019
hsa05200	Pathways in cancer	530	1.16	8	6.44E-06	0.000205
hsa05031	Amphetamine addiction	68	0.149	4	1.19E-05	0.000345
hsa04310	Wnt signaling pathway	158	0.347	5	1.69E-05	0.000431
hsa04971	Gastric acid secretion	75	0.165	4	1.76E-05	0.000431
hsa04012	ErbB signaling pathway	85	0.187	4	2.90E-05	0.000643
hsa04911	Insulin secretion	86	0.189	4	3.03E-05	0.000643

MAPK: mitogen-activated protein kinase; ErbB: erythroblastic oncogene B: KEGG, Kyoto Encyclopedia of Genes and Genomes.

Table 7
Five real hub genes by expression values in $\log_2$ base for each clinical feature.

Clinical Traits				
Hub Genes	hGBM	GBM_U251	GSC_X01	GSC_X03
GUCA1A	7.87859535	4.07658209	4.58879582	5.2724957
RFC2	9.03021097	9.83556542	8.80343733	8.40102803
GNG11	6.04000574	7.34968424	6.88022928	7.13599808
MMP19	9.78068367	11.191266	11.4270119	10.1253016
NRG1	7.8430097	9.38390535	8.24695019	7.66156598

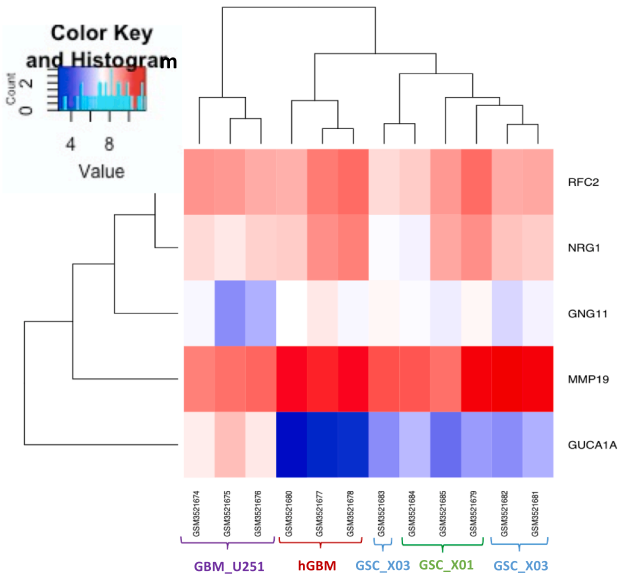

Fig. 7. Heatmap of the real hub genes expression values by the clinical traits.

enlarged diffusion (Fig. S1).

3.3. The GO and KEGG pathways enrichments of the four Metagenes

In Table 1, the four metagenes using the NMF method were detected according to significance level (p-value < 0. 05 and FDR < 0.05) of BP, CC, and MF of the GO pathway analysis. Moreover, KEGG pathway enrichments of the four metagenes are shown in Table 2.

It is demonstrated that the significant enrichments of the metagenes BP terms are homeostasis pathways, skeletal system morphogenesis, ion transport, stimulus, signaling, and cell communication pathways. The metagenes of CC terms are in general components of plasma membrane, extracellular region part, heterotrimeric G-protein complex, glutamatergic synapse, plasma membrane region, and endoplasmic reticulum part. The significant enrichment of the metagenes in MF are cytokine activity, transcription cofactor/factor activity, protein/actin binding, sequence-specific DNA binding, ion channel activity, and cell adhesion molecule binding. KEGG pathway enrichment of metagenes are mostly significant in hsa04724 (glutamatergic synapse), hsa04720 (long-term potentiation), hsa04713 (circadian entrainment), and hsa05214 (glioma).

3.4. Degs of paired clinical traits

Following data preprocessing and quality evaluation, the study revealed the expression values from the 12 samples in GSE124145. A total of 1985 DEGs (1658 up-regulated and 327 down-regulated as can be seen in Table S2-S7 in hGBM and GSC_X01, GSC_X03, and GBM_U251 clinical features, under the threshold of p-value < 0.05 and $|\log_2 FC| >$ 6.5 (Table 3) were screened for the subsequent analyses. The volcano plots of DEGs were shown in Fig. S2.

3.5. Identification of common hub genes

There were 450 genes in the four metagenes and 1985 DEGs in total. The overlapping number of genes in both sets was identified as 52 genes, which are the common genes (Fig. 5).

3.6. GO and KEGG pathway enrichment analysis of common hub genes

Table 4 shows the significant enrichments of the most candidate hub genes in BP terms are mostly enriched with GO:0007049 ~ cell cycle, GO:0031663 ~ lipopolysaccharide-mediated signaling pathway, GO:0048732 ~ gland development, GO:0042592 ~ homeostatic process, and GO:0046916 ~ cellular transition metal ion homeostasis. The significant enrichments GO terms in CC are GO:0031253 ~ cell projection membrane, GO:1903561 ~ extracellular vesicle, GO:0005789 ~ endoplasmic reticulum membrane, GO:0043230 ~ extracellular organelle, and GO:0042175 ~ nuclear outer membrane-endoplasmic reticulum membrane network. Moreover, the significant enrichment of the hub genes in MF contains GO:0072341 ~ modified amino acid binding, GO:0005524 ~ ATP binding, GO:0032559 ~ adenyl ribonucleotide binding, GO:0008392 ~ arachidonic acid epoxygenase activity, and GO:0008391 ~ arachidonic acid monooxygenase activity. Lastly, KEGG signaling pathway analysis reported that the hub genes were significantly enriched in hsa05164:influenza A, hsa05216:thyroid cancer, and hsa05020:prion diseases.

GO enrichment analysis of down-regulated and up-regulated DEGs of

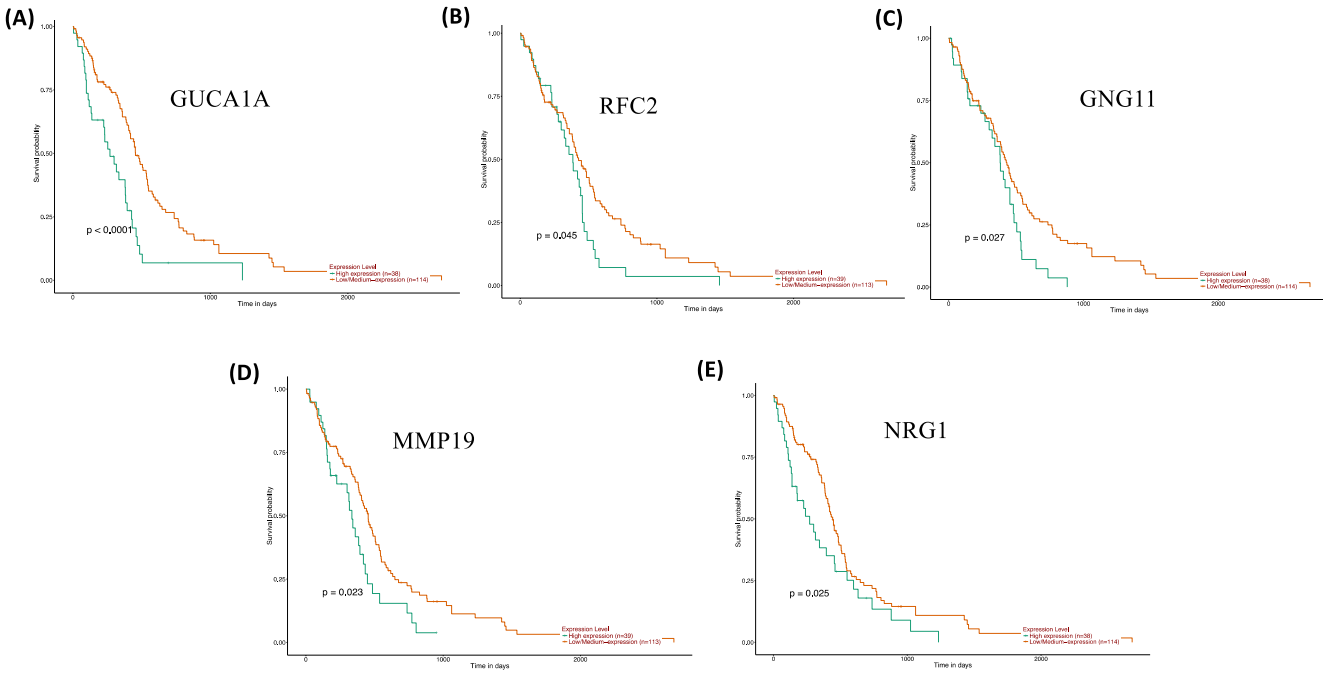

Fig. 8. Overall Survival (OS) analysis of GBM patients Kaplan- Meier analysis, log-rank test, P value < 0.05 of the real hub genes in the TCGA dataset via UALCAN. (A) GUCA1A, (B) RFC2, (C) GNG11, (D) MMP19, (E) NRG1. Orange lines represent low expression of the real hub genes, whereas green lines represent high expression of *n* patients.

categories BP, MF, and CC in each paired comparison of hGBM – GSC_X01, hGBM – GSC_X03, hGBM – GBM_U251, GSC_X01 –GSC_X03, GSC_X01 – GBM_U251, and GSC_X03 – GBM_U251 groups are listed in Table S2-S7.

3.7. PPI network

To achieve the analysis of protein–protein interactions, PPI network was built on the common hub genes list. A total of 256 nodes and 262 edges were screened from the PPI network (Fig. 6). Fifteen primary nodes with the highest degrees in the gene expression of GSE124145 dataset common hub genes are selected in Table 5. These were MAPK1, HSPA6, THRA, TTLL12, EPHB3, RFC2, GUCA1A, PTPN21, UBC, NRG1, MMP19, GNG11, NCKAP5, HSP90AA1, and ESR1. Genes with sequencing from yellow to red color shows the number of neighboring nodes and changes in proteins (genes) as illustrated in Fig. 6. The PPI network of common hub genes KEGG enrichment analysis picks the MAPK signaling pathway as the most significant (p-value < 0.05) pathway (Table 6). MAPK signaling pathway demonstrates an important role in many cancers involving GBM through hyperactivation and it is of concern in various biomarkers of tumorigenic progression such as migration, differentiation, proliferation, and survival. For PPI network of all the DEGs corresponding bi-comparison of clinical traits, see Fig. S3.

3.8. Validation of real hub genes

The overlapping common genes between key metagenes and DEGs were filtered and confirmed at expression levels and overall survival (OS) in TCGA datasets. After survival analysis by using GSE124145 and TCGA dataset, as illustrated in Table 7 and Fig. 7, we identified five real hub genes: Guanylate Cyclase Activator 1A (GUCA1A), Replication Factor C Subunit 2 (RFC2), G Protein Subunit Gamma 11 (GNG11), Matrix Metallopeptidase 19 (MMP19), and Neuregulin 1 (NRG1) strongly linked to the progression of GBM (Fig. 8).

4. Discussion

GBM is the most prevalent, destructive, and fatal brain tumor. Current treatment options, including surgery, chemotherapy, and radiotherapy, cannot fully treat the disease because the tumor is highly defiant to these treatments. GSCs have the self-renewal capacity and they are responsible for the tumor resistance in treating GBM. There is an urgent need to find out sensitive diagnostic and therapeutic targets for GBM. Our aim is to demonstrate the capability of NMF to uncover the biologically meaningful patterns from malignant brain tumor RNA-microarray data and discover reliable biomarkers for GBM.

In this study, first, we applied NMF analysis to GSE124145 dataset to evaluate the gene expression profile differences including GSCs, specifically U251 cell line, and a human GBM tissue sample. NMF is first described as a method for reduction dimension and feature identification on non-negative data (Lee and Seung, 1999). To detect gene expression patterns, NMF has been widely used (Esposito et al., 2020; Gaujoux et al., 2020). Its several advantages make NMF method more preferred than other clustering methods (k-means, hierarchical clustering, and self-organizing map algorithms). One of the most important advantages of NMF method is the ability to discover the specific genes for each population. In addition, NMF provides biologically interpretable results of microarray datasets.

This study identified the DEGs in the hGBM – GSC_X01, hGBM – GSC_X03, hGBM -GBM_U251, GSC_X01- GSC_X03, GSC_X01-GBM_U251, and GSC_X03 – GBM_U251 groups. After which, we chose the overlapping genes between the resulted metagenes utilizing NMF and identified DEGs; in total, 52 common hub genes were noted. Following a sequence of bioinformatics investigation, six real hub genes strongly related to the progression and prognosis of GBM were detected. The results suggested the development of therapeutic management, risk stratification, and prognosis prediction for GBM patients.

We detected significant enrichment of the most candidate hub genes in BP terms: GO:0007049 ~ cell cycle; the significant enrichment of the hub genes in MF contains GO:0072341 ~ modified amino acid binding, whereas the significant enrichment for GO terms in CC are GO:0031253 ~ cell projection membrane, GO:1903561 ~ extracellular vesicle,

GO:0005576 ~ extracellular region, and GO:0043230 ~ extracellular organelle. A previous study analyzing the GSE74304 and GSE124145 datasets also found out that one of the most enriched GO terms is extracellular region, which confirms our results (Wu et al., 2021).

We detect biological processes for each metagenes p-value < 0.05. The GO pathway analysis resulted in BP of Metagenes 1 mostly involved in GO:0001894 ~ tissue homeostasis, Metagenes 2 mostly involved in GO:0048705 ~ skeletal system morphogenesis, Metagenes 3 mostly involved in GO:0006811 ~ ion transport, and finally Metagenes 4 involved in GO:0071495 ~ cellular response to endogenous stimulus. Metagenes enriched for the different GO categories indicate that metagenes may affect different biological processes.

KEGG pathway enrichment analysis reported up-regulated DEGs that were involved in hsa04514:cell adhesion molecules (CAMs), hsa05202: transcriptional misregulation in cancer, hsa05014:amytrophic lateral sclerosis (ALS), and hsa04360:axon guidance. KEGG pathway enrichment analysis presented down-regulated genes that were mostly involved in hsa04151:PI3K-Akt signaling pathway (Daniel et al., 2018), hsa04728:dopaminergic synapse, and hsa04974:protein digestion and absorption. The constructed study of protein–protein interactions and KEGG pathway enrichment study indicated as hsa04010:MAPK signaling pathway, hsa04727:neurotrophin signaling pathway (Lawn et al., 2015; Zhu et al., 2017), hsa04725:cholinergic synapse (Yang et al., 2019), hsa04728:dopaminergic synapse (Wei et al., 2014), hsa04720: Long-term potentiation (Hu et al., 2015; Long et al., 2017), hsa05214: glioma in which hsa04141:protein processing in MAPK signaling pathway (Agarwal et al., 2015; Daniel et al., 2018; Krishna et al., 2021) was further studied because of a close association with GBM.

MAPK signaling pathway and PI3K-AKT signaling pathways were also among the most enriched KEGG pathways in the previous study, which used PCA to analyze the same dataset as in our study (Sakamoto et al., 2019). Based on the Cancer Genome Atlas project results, the phosphatidylinositol 3-kinase (PI3K) pathway is also an important signaling pathway in GBM (Uhm, 2009). Many roles of MAPK and PI3K-AKT signaling pathways, such as hyperactivation, proliferation, metabolism, survival, and migration, were shown in several cancer including GBM (Krishna et al., 2021). Targeting PI3K signaling pathway as a treatment approach for some cancers has been shown previously. Li et al., 2017 showed that the inhibition of MAPK signaling pathways can limit the glioma progression. However, there is not much clinical studies of PI3K and MAPK inhibitors for the treatment of GBM. Hence, the combination of PI3K and MAPK inhibitors could be effective therapeutic approaches for GBM. Our results support the idea that PI3K and MAPK signaling might be important therapeutic targets for GBM.

In this study, we implemented survival analysis to screen real hub genes under the cut-off p-value < 0.05. A total of five genes *(GUCA1A, RFC2, GNG11, MMP19, and NRG1)* were particularly outstanding. They were tightly linked with GBM prognosis likely to be potential biomarkers. Previous research also detected that those five real hub genes were involved in the process of the cell cycle, partaking in the tumor proliferation and formation. Our study identified different significant genes from the study that uses PCA to analyze the same dataset in which MYCN, DPP4 and MIF were reported as an important contributors of GSCs (Sakamoto et al., 2019).

It was proven that the mutation *GUCA1* (guanylyl cyclase-activating protein 1) is central to a severe dominantly inherited retinal degeneration (Buch et al., 2011; He et al., 2021; Hou et al., 2019), which is mostly active in signaling by the GPCR pathway (Cherry and Stella, 2014). Higher *RFC2* (replication factor subunit 2) levels, resulting in poor patient survival, were also noted in GBM patients and lower-grade gliomas (Ho et al., 2020).

G- protein family have many roles, including cell division and cell differentiation (Syrovatkina et al., 2016), and their roles in the progression of cancers have been shown in previous studies. GNG5 was potential biomarker for the gliomas (Zhang et al., 2021). Cai et al. reported that GNG11 (G protein subunit 11) expression was linked to poor prognosis (Cai et al., 2021). Our results contribute to the G-protein family members and they could be great biomarkers for GBM. Expression of *MMP19* (matrix metalloproteinase 19) was proven to be correlated with the WHO-grading of human malignant gliomas, which might be the cause of the growth of high-grade astrocytic tumors and might be candidate drug-targets. (Stojic et al., 2008; Wang et al., 2013). Neurogulins have roles in the development of nervous systems, and neuregulin family members are indicated in several cancer types (Cheng et al., 2019; Forster et al., 2011). NRG1 was found to promote malignancy in glioma and glioblastoma cells (Lin et al., 2020). Lin et al. also demonstrated NRG1 activates the MAPK signaling pathway, thus our findings strongly support that treatments targeting NRG1 and MAPK signaling pathway might successfully treat the GBM. Moreover, growing expression of NRG1 would converse the impacts of overexpression of miR-125a-3p on proliferation, apoptosis, and migration of glioblastoma cells reported by Yin et al, 2015. In addition, miR-125a-3p was relevant to the poor prognosis of GBM patients (Yin et al., 2015). Another recent study presented NRG1 to be differentially expressed in lower grade glioma and GBM in assessment to normal tissue (Zhao et al., 2021). All these genes were significantly up-regulated in GBM. Two GSCs (X01 and X02) with GBM tissue and U251 cell lines in GSE124145 in mRNA level have proven their vital role in angiogenesis. Further research is needed to discover the molecular functions of GUCA1A, RFL2, GNG11, MMP19, and NRG1 in the GBM to discover the treatment strategies of GBM. Our results would contribute to the future GBM studies to find out the molecular pathology of GBM.

Our study has some limitations. One of the limitations of this study is that NMF may not identify the biologically important genes with low expression levels. Another limitation is NMF method might be affected by batch effects in the microarray data. Although we validated the results in the TCGA dataset, the accuracy of the results requires molecular and cellular experiments. In this study, we showed that NMF is a useful method to find significant genes.

5. Conclusions

In this study, we focused on one of the GBM gene expression datasets in the database of GEO. In this study, *GUCA1A, RFC2, GNG11, MMP19, NCKAP5,* and *NRG1* were screened as the real hub genes for the upcoming molecular studies in GBM. These hub genes can be presented to the promising prospect of future research for therapeutic targets in GBM. The rest of the analysis in this study would help explore the causes of the gliomas, in particular GBM, underlying biological, cellular, and functional events. In this study, we showed that the combination of NMF and DEGs analyses is a useful method in finding significant genes at high resolution and interpreting the biological meaning of microarray data.

Funding

This work was funded by Deanship of Scientific Research at Jouf University under grant No (DSR-2021–01-0315).

7. Data availability

The GSE124145 dataset used and analyzed in this present study are available in the NIH GEO
(http://www.ncbi.nlm.nih.gov/geo) public repository.

CRediT authorship contribution statement

Sevinç Akçay: Conceptualization, Methodology, Software, Investigation, Writing – review & editing. **Emine Güven:** Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Funding acquisition. **Muhammad Afzal:** Investigation, Data curation, Writing – review & editing, Visualization. **Imran Kazmi:** Investigation, Data

curation, Writing – review & editing, Visualization.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work was funded by the Deanship of Scientific Research Jouf University, Saudi Arabia under the grant number (DSR-2021-01-0315).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.gene.2022.146395.

References

Agarwal, K., Saji, M., Lazaroff, S., Palmer, A.F., Ringel, M.D., Paulaitis, M.E., 2015. Analysis of exosome release as a cellular response to MAPK pathway inhibition. Langmuir 31, 5440–5448.

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. Roy. Stat. Soc.: Ser. B (Methodol.) 57, 289–300.

Berens, M.E., Rief, M.D., Loo, M.A., Giese, A., 1994. The role of extracellular matrix in human astrocytoma migration and proliferation studied in a microliter scale assay. Clin. Exp. Metast. 12, 405–415. https://doi.org/10.1007/BF01755884.

Boccarelli, A., Esposito, F., Coluccia, M., Frassanito, M.A., Vacca, A., Del Buono, N., 2018. Improving knowledge on the activation of bone marrow fibroblasts in MGUS and MM disease through the automatic extraction of genes via a nonnegative matrix factorization approach on gene expression profiles. J. Translational Med. 16, 1–16.

Brunet, J.-P., Tamayo, P., Golub, T.R., Mesirov, J.P., 2004. Metagenes and molecular pattern discovery using matrix factorization. PNAS 101, 4164–4169. https://doi.org/10.1073/pnas.0308531101.

Buch, P.K., Mihelec, M., Cottrill, P., Wilkie, S.E., Pearson, R.A., Duran, Y., West, E.L., Michaelides, M., Ali, R.R., Hunt, D.M., 2011. Dominant cone-rod dystrophy: a mouse model generated by gene targeting of the GCAP1/Guca1a gene. PLoS ONE 6, e18089.

Cai, Z., Yu, C., Li, S., Wang, C., Fan, Y., Ji, Q., Chen, F., Li, W., 2021. A Novel Classification of Glioma Subgroup, Which Is Highly Correlated With the Clinical Characteristics and Tumor Tissue Characteristics, Based on the Expression Levels of Gβ and Gγ Genes. Front. Oncol. 11, 2256. https://doi.org/10.3389/fonc.2021.685823.

Chandrashekar, D.S., Bashel, B., Balasubramanya, S.A.H., Creighton, C.J., Ponce-Rodriguez, I., Chakravarthi, B.V., Varambally, S., 2017. UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia 19, 649–658.

Cheng, Q., Huang, C., Cao, H., Lin, J., Gong, X., Li, J., Chen, Y., Tian, Z., Fang, Z., Huang, J., 2019. A Novel Prognostic Signature of Transcription Factors for the Prediction in Patients With GBM. Front. Genetics 10.

Cherry, A.E., Stella, N., 2014. G protein-coupled receptors as oncogenic signals in glioma: emerging therapeutic avenues. Neuroscience 222–236. https://doi.org/10.1016/j.neuroscience.2014.08.015.

Collisson, E.A., Sadanandam, A., Olson, P., Gibb, W.J., Truitt, M., Gu, S., Cooc, J., Weinkle, J., Kim, G.E., Jakkula, L., Feiler, H.S., Ko, A.H., Olshen, A.B., Danenberg, K. L., Tempero, M.A., Spellman, P.T., Hanahan, D., Gray, J.W., 2011. Subtypes of pancreatic ductal adenocarcinoma and their differing responses to therapy. Nat. Med. 17, 500–503. https://doi.org/10.1038/nm.2344.

Daniel, P.M., Filiz, G., Tymms, M.J., Ramsay, R.G., Kaye, A.H., Stylli, S.S., Mantamadiotis, T., 2018. Intratumor MAPK and PI3K signaling pathway heterogeneity in glioblastoma tissue correlates with CREB signaling and distinct target gene signatures. Exp. Mol. Pathol. 105, 23–31. https://doi.org/10.1016/j.yexmp.2018.05.009.

Davis, M.E., 2016. Glioblastoma: overview of disease and treatment. Clin. J. Oncol. Nursing 20, S2.

Davis, S., Meltzer, P., 2007. GEOquery: A bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics (Oxford, England) 23, 1846–1847. https://doi.org/10.1093/bioinformatics/btm254.

Devarajan, K., 2008. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. PLoS Comput. Biol. 4, e1000029 https://doi.org/10.1371/journal.pcbi.1000029.

Dirks, P.B., 2008. Brain tumour stem cells: the undercurrents of human brain cancer and their relationship to neural stem cells. Philosophical Trans. Roy. Soc. B: Biol. Sci. 363, 139–152.

Dudoit, S., Shaffer, J.P., Boldrick, J.C., 2003. Multiple hypothesis testing in microarray experiments. Statistical Sci. 71–103.

Durinck, S., Moreau, Y., Kasprzyk, A., Davis, S., De Moor, B., Brazma, A., Huber, W., 2005. BioMart and Bioconductor: a powerful link between biological databases and microarray data analysis. Bioinformatics 21, 3439–3440.

Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D., 1998. Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. 95, 14863–14868.

Esposito, F., Boccarelli, A., Del Buono, N., 2020. An NMF-Based Methodology for Selecting Biomarkers in the Landscape of Genes of Heterogeneous Cancer-Associated Fibroblast Populations. Bioinf. Biol. Insights 14. https://doi.org/10.1177/1177932220906827.

Forster, J.A., Paul, A.B., Harnden, P., Knowles, M.A., 2011. Expression of NRG1 and its receptors in human bladder cancer. Br. J. Cancer 104, 1135–1143. https://doi.org/10.1038/bjc.2011.39.

Frigyesi, A., Höglund, M., 2008. Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes. Cancer Inf. 6, CIN-S606.

Gaujoux, R., Seoighe, C., 2010. A flexible R package for nonnegative matrix factorization. BMC Bioinf. 11, 367. https://doi.org/10.1186/1471-2105-11-367.

Gaujoux, R., Seoighe, C., Gaujoux, M.R., 2020. Package 'NMF.'.

Hatfield, G.W., Hung, S., Baldi, P., 2003. Differential analysis of DNA microarray gene expression data. Mol. Microbiol. 47, 871–877.

He, J., Zeng, C., Long, Y., 2021. Establishment of an Immune-Related Gene Signature for Risk Stratification for Patients with Glioma. Comput. Math. Methods Med. 2021, 2191709. https://doi.org/10.1155/2021/2191709.

Ho, K.-H., Kuo, T.-C., Lee, Y.-T., Chen, P.-H., Shih, C.-M., Cheng, C.-H., Liu, A.-J., Lee, C.-C., Chen, K.-C., 2020. Xanthohumol regulates miR-4749-5p-inhibited RFC2 signaling in enhancing temozolomide cytotoxicity to glioblastoma. Life Sci. 254, 117807 https://doi.org/10.1016/j.lfs.2020.117807.

Hochberg, Y., Tamhane, A.C., 1987. Multiple comparison procedures. John Wiley & Sons Inc.

Hou, Z., Yang, J., Wang, H., Liu, D., Zhang, H., 2019. 201920192019A potential prognostic gene signature for predicting survival for glioblastoma patients. BioMed Res. Int.

Hu, G., Wei, B., Wang, L., Wang, L.e., Kong, D., Jin, Y., Sun, Z., 2015. Analysis of gene expression profiles associated with glioma progression. Mol. Med. Rep. 12, 1884–1890.

Huang, D.W., Sherman, B.T., Lempicki, R.A., 2009. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 37, 1–13.

Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C., 2007. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 35, W169–W175.

Jamail, I., Moussa, A., 2020. Current State-of-the-Art of Clustering Methods for Gene Expression Data with RNA-Seq, in. Pattern Recognition. IntechOpen.

Jiang, Y., Sun, A., Zhao, Y., Ying, W., Sun, H., Yang, X., Xing, B., Sun, W., Ren, L., Hu, B., 2019. Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. Nature 567, 257–261.

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462.

Kim, H., Park, H., 2007. Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis. Bioinformatics 23, 1495–1502. https://doi.org/10.1093/bioinformatics/btm134.

Konopka, T., Konopka, M.T., 2018. R-package: umap. Uniform Manifold Approximation and Projection.

Krishna, K.V., Dubey, S.K., Singhvi, G., Gupta, G., Kesharwani, P., 2021. MAPK pathway: Potential role in glioblastoma multiforme. Interdisciplinary Neurosurgery 23, 100901. https://doi.org/10.1016/j.inat.2020.100901.

Lai, A., Kharbanda, S., Pope, W.B., Tran, A., Solis, O.E., Peale, F., Forrest, W.F., Pujara, K., Carrillo, J.A., Pandita, A., 2011. Evidence for sequenced molecular evolution of IDH1 mutant glioblastoma from a distinct cell of origin. J. Clin. Oncol. 29, 4482.

Lan, X., Jörg, D.J., Cavalli, F.M., Richards, L.M., Nguyen, L.V., Vanner, R.J., Guilhamon, P., Lee, L., Kushida, M.M., Pellacani, D., 2017. Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. Nature 549, 227–232.

Lawn, S., Krishna, N., Pisklakova, A., Qu, X., Fenstermacher, D.A., Fournier, M., Vrionis, F.D., Tran, N., Chan, J.A., Kenchappa, R.S., Forsyth, P.A., 2015. Neurotrophin Signaling via TrkB and TrkC Receptors Promotes the Growth of Brain Tumor-initiating Cells *. J. Biol. Chem. 290, 3814–3824. https://doi.org/10.1074/jbc.M114.599373.

Lee, D.D., Seung, H.S., 1999. Learning the parts of objects by non-negative matrix factorization. Nature 401, 788–791.

Li, F., Yi, Y., Miao, Y., Long, W., Long, T., Chen, S., Cheng, W., Zou, C., Zheng, Y., Wu, X., 2019. N6-methyladenosine modulates nonsense-mediated mRNA decay in human glioblastoma. Cancer Res. 79, 5785–5798.

Lin, W., Ou, G., Lin, J., Yi, S., Yao, W., Pan, H., Zhao, W., 2020. Neuregulin 1 enhances cell adhesion molecule L1 like expression levels and promotes malignancy in human glioma. Oncol. Lett. 20, 326–336.

Long, H., Liang, C., Zhang, X., Fang, L., Wang, G., Qi, S., Huo, H., Song, Y., 2017. Prediction and Analysis of Key Genes in Glioblastoma Based on Bioinformatics. Biomed Res. Int. 2017, e7653101 https://doi.org/10.1155/2017/7653101.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Presented at the Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA, pp. 281–297.

Park, P.J., 2005. Gene Expression Data and Survival Analysis. Methods of Microarray Data Analysis. Springer 21–34.

Perry, A., Miller, C.R., Gujrati, M., Scheithauer, B.W., Zambrano, S.C., Jost, S.C., Raghavan, R., Qian, J., Cochran, E.J., Huse, J.T., 2009. Malignant gliomas with primitive neuroectodermal tumor-like components: a clinicopathologic and genetic study of 53 cases. Brain Pathol. 19, 81–90.

Quackenbush, J., 2002. Microarray data normalization and transformation. Nat. Genet. 32, 496–501.

R Core Team, 2020. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL. https://www.R-project.org/.

Ramanarayanan, V., Katsamanis, A., Narayanan, S., 2011. Automatic data-driven learning of articulatory primitives from real-time mri data using convolutive nmf with sparseness constraints. Presented at the Twelfth Annual Conference of the International Speech Communication Association.

Rau, A., Flister, M., Rui, H., Auer, P.L., 2019. Exploring drivers of gene expression in the Cancer Genome Atlas. Bioinformatics 35, 62–68. https://doi.org/10.1093/bioinformatics/bty551.

Sakamoto, D., Takagi, T., Fujita, M., Omura, S., Yoshida, Y., Iida, T., Yoshimura, S., 2019. Basic Gene Expression Characteristics of Glioma Stem Cells and Human Glioblastoma. Anticancer Res. 39, 597–607. https://doi.org/10.21873/anticanres.13153.

Stojic, J., Hagemann, C., Haas, S., Herbold, C., Kühnel, S., Gerngras, S., Roggendorf, W., Roosen, K., Vince, G.H., 2008. Expression of matrix metalloproteinases MMP-1, MMP-11 and MMP-19 is correlated with the WHO-grading of human malignant gliomas. Neurosci. Res. 60, 40–49. https://doi.org/10.1016/j.neures.2007.09.009.

Stupp, R., Hegi, M.E., Gorlia, T., Erridge, S.C., Perry, J., Hong, Y.-K., Aldape, K.D., Lhermitte, B., Pietsch, T., Grujicic, D., 2014. Cilengitide combined with standard treatment for patients with newly diagnosed glioblastoma with methylated MGMT promoter (CENTRIC EORTC 26071–22072 study): a multicentre, randomised, open-label, phase 3 trial. Lancet Oncol. 15, 1100–1108.

Swain, S., Banerjee, A., Bandyopadhyay, M., Satapathy, S.C., 2021. Dimensionality reduction and classification in hyperspectral images using deep learning, in: Machine Learning Approaches for Urban Computing. Springer, pp. 113–140.

Syrovatkina, V., Alegre, K.O., Dey, R., Huang, X.-Y., 2016. Regulation, signaling, and physiological functions of G-proteins. J. Mol. Biol. 428, 3850–3868.

Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., 2016. The STRING database in 2017: quality-controlled protein–protein association networks, made broadly accessible. Nucleic acids research gkw937.

Tamayo, P., Slonim, D., Mesirov, J., Zhu, Q., Kitareewan, S., Dmitrovsky, E., Lander, E.S., Golub, T.R., 1999. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. 96, 2907–2912.

Tang, M., Sun, J., Shimizu, K., Kadota, K., 2015. Evaluation of methods for differential expression analysis on multi-group RNA-seq count data. BMC Bioinf. 16, 360. https://doi.org/10.1186/s12859-015-0794-7.

Tarca, A.L., Romero, R., Draghici, S., 2006. Analysis of microarray experiments of gene expression profiling. Am. J. Obstet. Gynecol. 195 (2), 373–388.

Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J., Church, G.M., 1999. Systematic determination of genetic network architecture. Nat. Genet. 22, 281–285.

Tsai, C.-Y., Chiu, C.-C., 2010. A novel microarray biclustering algorithm. Int. J. Math. Computational Sci. 4, 533–539.

Turner, H., Bailey, T., Krzanowski, W., 2005. Improved biclustering of microarray data demonstrated through systematic performance tests. Comput. Stat. Data Anal. 48, 235–254.

Uhm, J., 2009. Comprehensive genomic characterization defines human glioblastoma genes and core pathways The Cancer Genome Atlas Research Network Nature 455: 1061–1068, 2008. Year Book of Neurology and Neurosurgery 2009, 117–118.

Vidman, L., Källberg, D., Rydén, P., 2019. Cluster analysis on high dimensional RNA-seq data with applications to cancer research-An evaluation study. PLoS ONE 14, e0219102.

Wang, L., Yuan, J., Tu, Y., Mao, X., He, S., Fu, G., Zong, J., Zhang, Y., 2013. Co-expression of MMP-14 and MMP-19 predicts poor survival in human glioma. Clin. Transl. Oncol. 15, 139–145.

Warnes, G.R., Bolker, B., Bonebakker, L., Gentleman, R., Huber, W., Liaw, A., Lumley, T., Maechler, M., Magnusson, A., Moeller, S., 2009. gplots: Various R programming tools for plotting data. R package version 2, 1.

Wei, B., Wang, L., Zhao, X., Jin, Y., Kong, D., Hu, G., Sun, Z., 2014. Co-mutated pathways analysis highlights the coordination mechanism in glioblastoma multiforme. Neoplasma 61, 424–432.

Wu, X., Xiao, S., Zhang, M., Yang, L., Zhong, J., Li, B., Li, F., Xia, X., Li, X., Zhou, H., 2021. A novel protein encoded by circular SMO RNA is essential for Hedgehog signaling activation and glioblastoma tumorigenicity. Genome Biol. 22, 1–29.

Yang, Q., Wang, R., Wei, B., Peng, C., Wang, L., Hu, G., Kong, D., Du, C., 2019. Gene and microRNA signatures are associated with the development and survival of glioblastoma patients. DNA Cell Biol. 38, 688–699.

Yin, F., Zhang, J.N., Wang, S.W., Zhou, C.H., Zhao, M.M., Fan, W.H., Fan, M., Liu, S., 2015. MiR-125a-3p Regulates Glioma Apoptosis and Invasion by Regulating Nrg1. PLoS ONE 10, e0116759. https://doi.org/10.1371/journal.pone.0116759.

Zhang, W., Liu, Z., Liu, B., Jiang, M., Yan, S., Han, X., Shen, H., Na, M., Wang, Y., Ren, Z., 2021. GNG5 is a novel oncogene associated with cell migration, proliferation, and poor prognosis in glioma. Cancer Cell Int. 21, 1–20.

Zhang, Y., Du, N., Ge, L., Jia, K., Zhang, A., 2012. A collective nmf method for detecting protein functional module from multiple data sources. In: Presented at the Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine, pp. 655–660.

Zhao, W., Ou, G., Lin, W., 2021. Integrative Analysis of Neuregulin Family Members-Related Tumor Microenvironment for Predicting the Prognosis in Gliomas. Front. Immunol. 12, 1784. https://doi.org/10.3389/fimmu.2021.682415.

Zhou, G., Soufan, O., Ewald, J., Hancock, R.E.W., Basu, N., Xia, J., 2019. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. Nucleic Acids Res. 47, W234–W241. https://doi.org/10.1093/nar/gkz240.

Zhu, J., Ye, J., Zhang, L., Xia, L., Hu, H., Jiang, H., Wan, Z., Sheng, F., Ma, Y., Li, W., 2017. Differential expression of circular RNAs in glioblastoma multiforme and its correlation with prognosis. Transl. Oncol. 10, 271–279.

Further reading

Stupp, R., Hegi, M.E., Gorlia, T., Erridge, S.C., Perry, J., Hong, Y.-K., Aldape, K.D., Lhermitte, B., PiAgarwal, K., Saji, M., Lazaroff, S., Palmer, A.F., Ringel, M.D., Paulaitis, M.E., 2015. Analysis of exosome release as a cellular response to MAPK pathway inhibition. Langmuir 31, 5440–5448.