Research Paper

Identification of HMG-box family establishes the significance of SOX6 in the malignant progression of glioblastoma

Lan Jiang[1,2], Hui Yang[1,2], Tianbing Chen[1,2], Xiaolong Zhu[1,2], Jingjing Ye[1,2], Kun Lv[1,2]

[1]Central Laboratory, Yijishan Hospital of Wannan Medical College, Wuhu 241001, China
[2]Key Laboratory of Non-coding RNA Transformation Research of Anhui Higher Education Institution, Wannan Medical College, Wuhu 241001, China

Correspondence to: Kun Lv; **email:** lvkun@yjsyy.com

Keywords: lncRNA, ceRNA, HMG-box, SOX6, glioblastoma
Received: December 7, 2019 **Accepted:** March 30, 2020 **Published:** May 10, 2020

ABSTRACT

Glioblastoma multiforme (GBM) is the most malignant neuroepithelial primary brain tumor and its mean survival time is 15 months after diagnosis. This study undertook to investigate the genome-wide and transcriptome-wide analyses of human high mobility group box (HMG-box) TF (transcript factor) families / HOX, TOX, FOX, HMG and SOX gene families, and their relationships to GBM. According to the TCGA-GBM profile analysis, differentially expressed HOX, FOX, HMG and SOX gene families (62 DEmRNA) were found in this study. We also analyzed DEmRNA (HMG-box related genes) co-expressed eight DElncRNA in GBM, and constructed a ceRNA network analysis as well. We constructed 50 DElncRNA-DEmiRNA-DEmRNA (HMG-box related genes) pairs between GBM and normal tissues. Then, risk genes SOX6 and SOX21 expression were correlated with immune infiltration levels in GBM. SOX6 also had a strong association with MAPT, GSK3B, FYN and DPYSL4, suggesting that they might be functional members in GBM.

INTRODUCTION

Glioblastoma multiforme (GBM) is the most malignant neuroepithelial primary brain tumor [1]. For GBM patients, its mean survival time is 15 months after diagnosis [2]. HMG-box (high mobility group box) domains are associated with the HMG-box proteins which influence DNA-dependent processes (transcription, replication, and DNA repair) and require changing the conformation of chromatin [3].

The HMG-box gene family is a family of TF-encoding genes which include a DNA-binding homeobox domain [4], such as HOX, FOX, SOX, HMG, and TOX genes. There were many studies on HMG-box genes in gliomas. HOX gene family was highly expressed in GBM cancer stem cells compared with parental lines, and HOX-PBX inhibition was a potential therapeutic target for GBM patients [5], and HOXD10 was targeted by hsa-miRNA-23a to inhibit glioma cell invasion [6]. Sex-determining region Y (SRY)-related high mobility group box of genes was abbreviated as SOX genes [7]. Using human glioma-initiating cell (GIC) lines (GIC1 and GIC2) created from anaplastic oligodendroglioma (AO) and GBM, both GIC1 and GIC2 expressed SOX2 and SOX3, and neither GIC line expressed SOX1 [8]. The gliogenesis of GBM was dependent on SoxD (SOX5, SOX6 and SOX7) and SoxE (SOX8, SOX9 and SOX10) [9]. SOX6 was specifically expressed by IgGs in GBM [10]. The moderate expression of SOX10 and SOX11 was linked to glioma, whereas the over-expression of them were associated with GBM [11]. SOX9 expression is connected to a poor prognosis of GBM patients and with resistance to temozolomide [12]. SOX2 / SOX21 axis could function as a tumor suppressor during glioma genesis [13]. However, SRY, SOX12, SOX15, SOX18, and SOX30 have not been reported to be associated with GBM. FOXM1 overexpression promoted clonogenic growth of GBM cells. FOXG1 and SOX2 via transcriptional control of core cell cycle and epigenetic regulators to fuel unconstrained self-renewal in GBM stem cells [14]. SOX9 and FOXG1 co-regulated a subset of EGFR [15]. Hsa-miR-338-5p also played a tumor suppressor role in glioma by binding FOXD1 [16].

Non-coding RNA plays an important role in post-transcriptional control of many animals [17, 18]. Numerous miRNAs could also bind and regulate SOX genes in GBM. SOX5 was over-expressed in GBM tissues, SNHG12-miR-195-SOX5 feedback loop could regulate the glioma cells' malignant progression [19]. MiR-143, miR-253, miR-452 and miR-145 could down-regulate SOX2 in GBM, whereas miR-145 worked as a tumor-suppressive RNA by targeting SOX9 in human glioma cells [20]. SOX7 inhibited GBM tissue and was regulated by several miRNAs, such as miR-595 [21], miR-24 [22], miR-128 [23] and miR-616 [24].

However, transcriptomic- and genomic- wide systematic studies of HMG-box families in GBM is lacking. In order to better solve this problem, integrated analysis of HMG-box related gene family in GBM based on data gathered from GEO and TCGA database. We expected to find the DE-HMG-box and related non-coding RNA in GBM, and discovered the potential drug or disease target for GBM. Our findings provided new insights into the molecular role and phylogeny of the HMG-box families in GBM.

RESULTS

Transcriptomic identification of DEGs between GBM and normal tissues

By obtaining data from TCGA database, we re-analyze the transcriptomic profiles of TCGA-GBM dataset, and 174 samples (169 GBM tissues and 5 normal tissues) were chosen to obtain DEmRNA (differentially expressed mRNA) and DElncRNA (differentially expressed lncRNA). GBM miRNAs expressed profiles were downloads from GEO database (GSE90603). There were 123 HMG-box genes that exist in TCGA-GBM. Through the analysis of the TCGA datasets, it was found that partial SOX, FOX, HOX, TOX and HMG gene families (a total of 62 genes) were significantly differentially expressed (|logFC| > 1 and q-value < 0.05) in our study, relative expression heatmap visualization was drawn in Supplementary Figure 1. Starting from the left, the first 5 datasets were normal tissues, and the remaining 169 were GBM tissues. Differentially expressed HMG-box genes were displayed via volcano plot (Figure 1). Only five HMG-box DEGs (FOXP1, FOXO4, SOX7, FOXP2 and

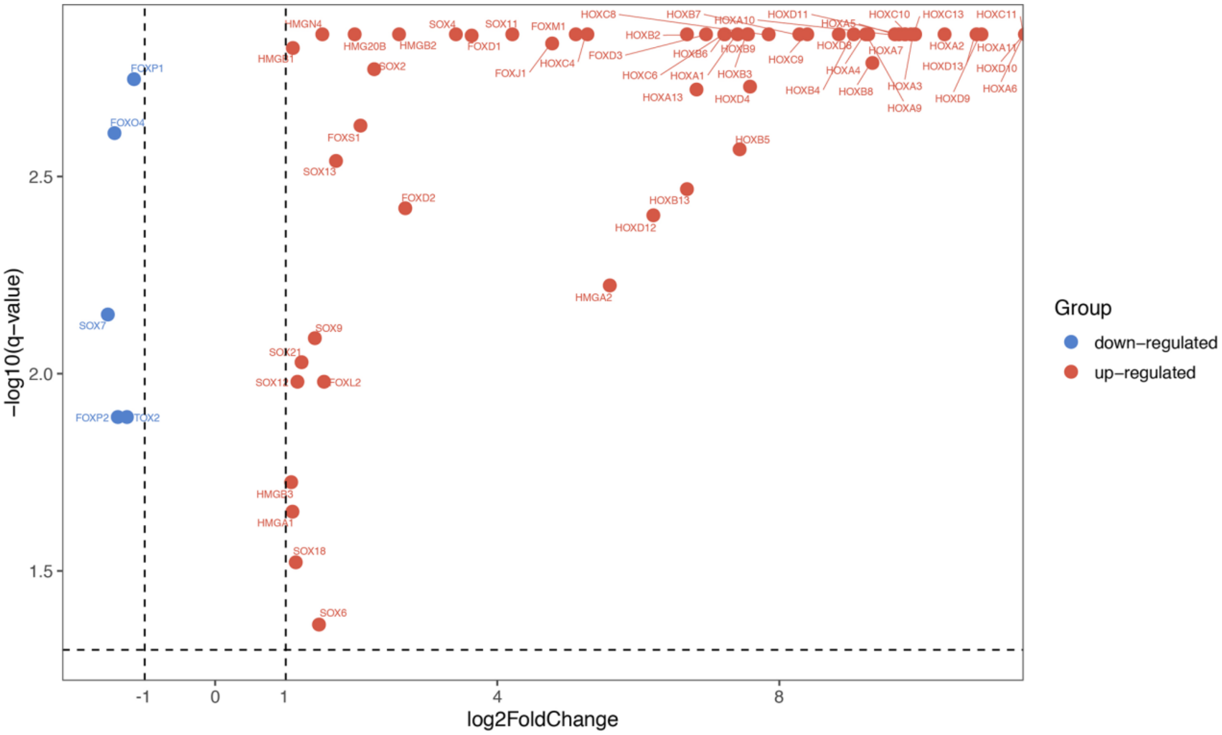


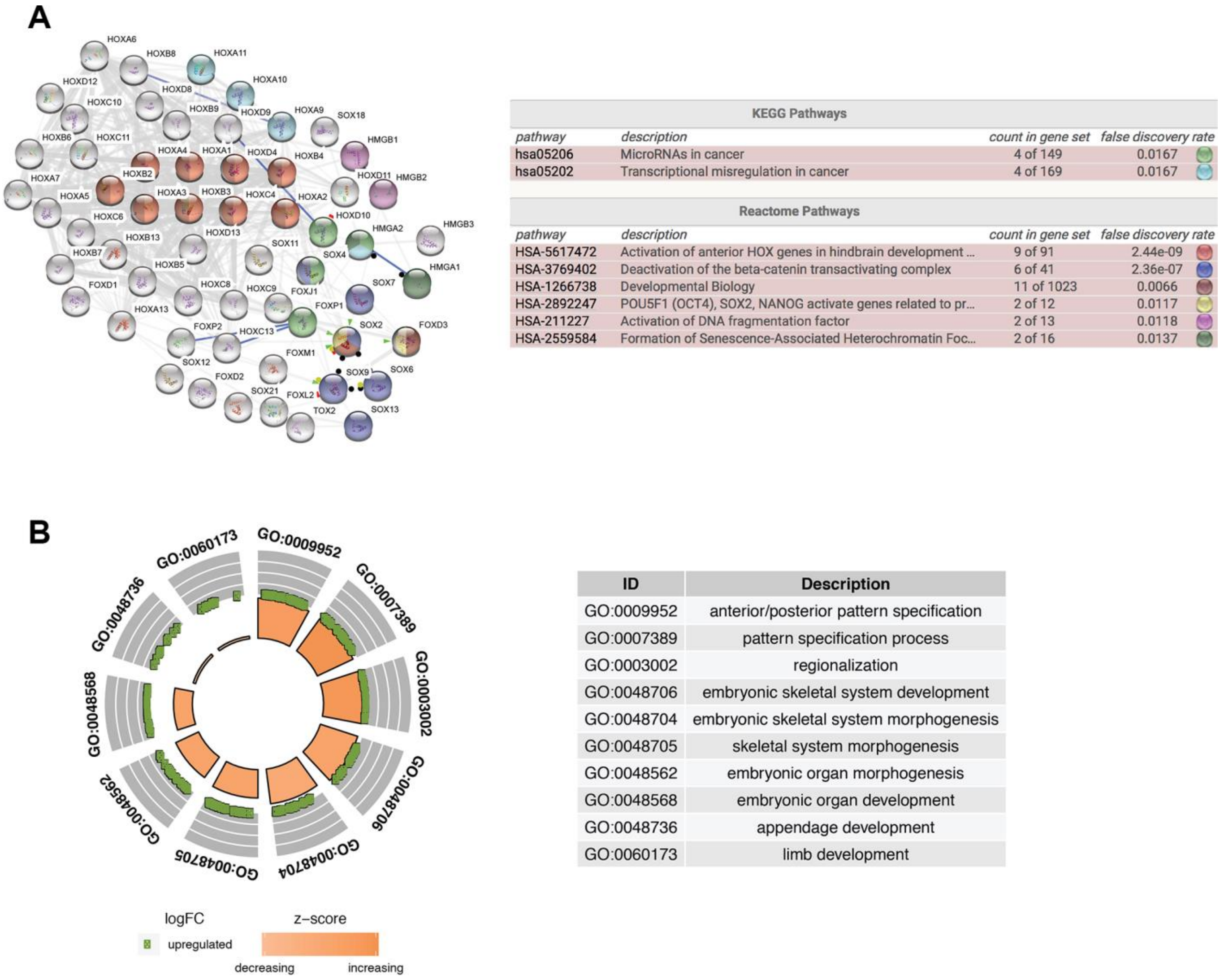
Figure 1. Differentially expressed HMG-box genes were displayed via volcano chart.

TOX2) were downregulated between GBM and normal tissues, the others were up-regulated.

To further explore the function of isolated DE-HMG-box related genes, these 62 DEmRNAs were entered into clusterProfiler for GO enrichment. PPI network and Reactome and KEGG pathway enrichment analyses were built by STRING database and also presents the most significant enriched pathways of DE-HMG-box related genes in GBM (Figure 2A). Moreover, HMGA2, HOX9-11 were enriched in "Transcriptional mis-regulation in cancer" in the KEGG pathway. From Reactome pathway results, we found these genes enriched in six pathways, such as HOXA1-4, HOXB2-4, HOXC4, SOX2, and FOXD3 were enriched in "developmental biology", HMGB1-2 were enriched in "Activation of DNA fragmentation factor" (Figure 2A). The top ten GO enrichment analysis results (q-value < 0.05) were shown in Figure 2B, the most significantly enriched in "GO:0009952: anterior/posterior pattern specification" (Figure 2B).

Identification of HMG-box DEGs co-expressed lncRNAs

According to the median risk score, GBM patients in TCGA were divided into high- and low-risk groups. We performed principle component analysis (PCA) graphs on the HMG-box related DEmRNA, co-expressed DElncRNA and risk DElncRNA (Figure 3A–3C), green dots present low risk, and red dots present high risk in GBM patients. The eight HMG-box related lncRNAs heatmap employed in constructing the risk scoring model and survival information were shown in



KEGG Pathways				
pathway	*description*	*count in gene set*	*false discovery rate*	
hsa05206	MicroRNAs in cancer	4 of 149	0.0167	●
hsa05202	Transcriptional misregulation in cancer	4 of 169	0.0167	●

Reactome Pathways				
pathway	*description*	*count in gene set*	*false discovery rate*	
HSA-5617472	Activation of anterior HOX genes in hindbrain development ...	9 of 91	2.44e-09	●
HSA-3769402	Deactivation of the beta-catenin transactivating complex	6 of 41	2.36e-07	●
HSA-1266738	Developmental Biology	11 of 1023	0.0066	●
HSA-2892247	POU5F1 (OCT4), SOX2, NANOG activate genes related to pr...	2 of 12	0.0117	●
HSA-211227	Activation of DNA fragmentation factor	2 of 13	0.0118	●
HSA-2559584	Formation of Senescence-Associated Heterochromatin Foc...	2 of 16	0.0137	●

ID	Description
GO:0009952	anterior/posterior pattern specification
GO:0007389	pattern specification process
GO:0003002	regionalization
GO:0048706	embryonic skeletal system development
GO:0048704	embryonic skeletal system morphogenesis
GO:0048705	skeletal system morphogenesis
GO:0048562	embryonic organ morphogenesis
GO:0048568	embryonic organ development
GO:0048736	appendage development
GO:0060173	limb development

Figure 2. Functional enrichment analysis of differentially expressed HMG-box related genes. (**A**) Integrative analysis of PPI network and pathway enrichment analysis (KEGG and Reactome). (**B**) The top ten of GO enrichment analysis.

Figure 3D, 3E. The hazard ratio of eight risk lncRNAs is shown in the forest plot (Figure 3F). Of these eight lncRNAs, six were detected as high risk (BNC2-AS1, AC018450.1, MIR222HG, AC005005.3, AC025171.1, AGAP2-AS1, coefficient > 0), while two were supportive (SOX21-AS1, ZEB1-AS1, coefficient < 0). We also found that the overall survival time of patients in the high-risk group was lower than that in the low-risk group (p-value <1.604e-08, Figure 3G).

A total of 147 DElncRNA (q-value < 0.05) were gained as well, of which 44 DElncRNA were up-regulated and 103 DElncRNA down-regulated. The association networks that included the DE-HMG-box gene families and their related co-expressed DElncRNA were constructed (Figure 4). The resulting lncRNA-mRNA association network had 68 interfaces between 38 lncRNAs and 27 mRNAs. The network showed that SOX6 was proposed to be the target of nine lncRNAs, FOXO4 was targeted by seven lncRNAs, and three mRNAs (HOXD4, SOX11, and SOX6) targeted AP002360.3.

CeRNA network construction

We collected TCGA-GBM profiles (lncRNAs and mRNAs) and GEO data GSE90603 (miRNAs) in GBM through computational analysis to estimate potential relationships based on the ceRNA hypothesis to further understand their function. We found that 401

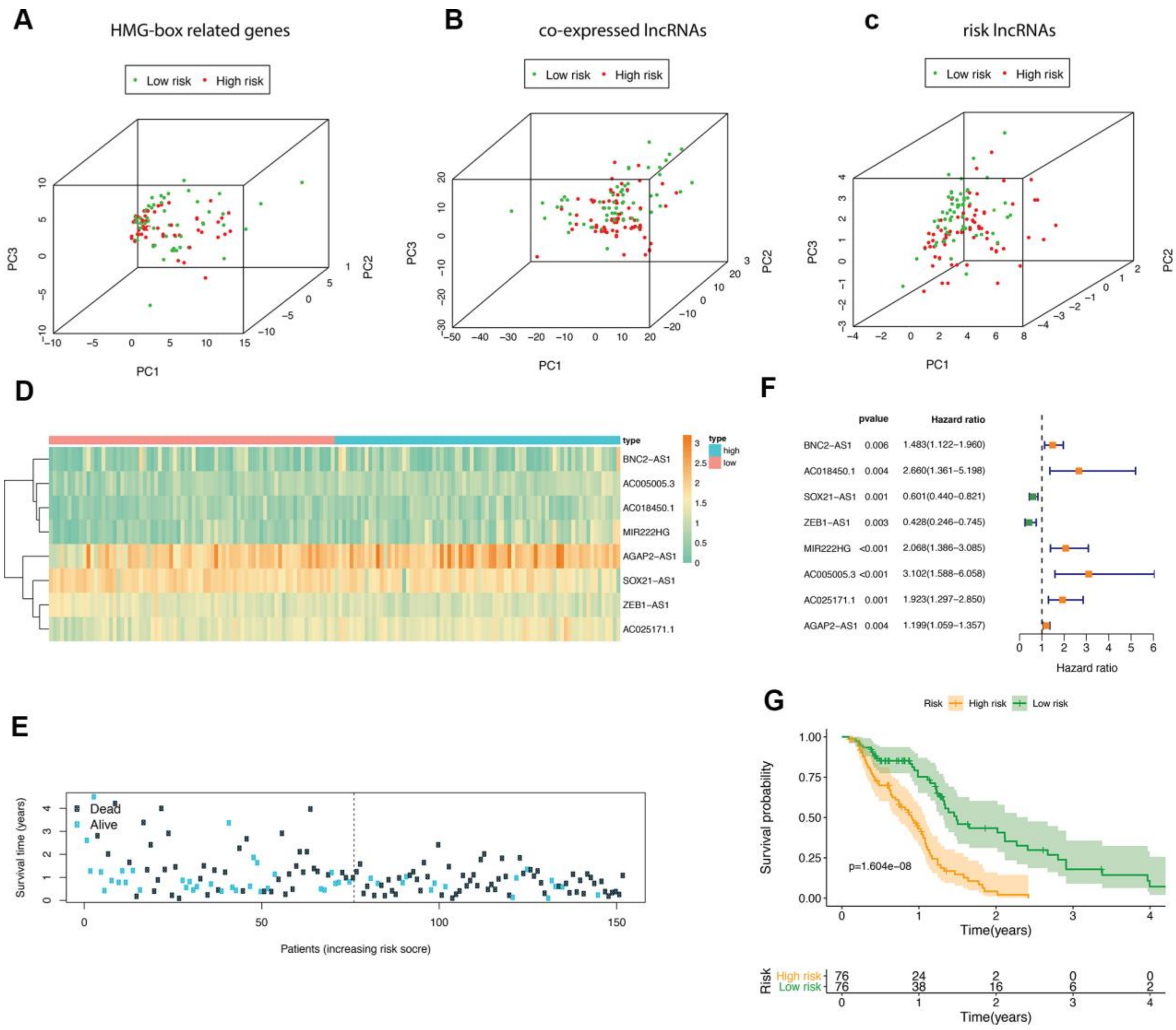


Figure 3. The analyses of HMG-box related DEmRNAs co-expressed lncRNA. Principle component analysis (PCA) of HMG-box related DEmRNAs was shown in (**A**), PCA analysis for co-expressed lncRNA in (**B**); PCA analysis for risk lncRNA was shown in (**C**). (**D**) Heatmap of risk lncRNA among high and low risk groups. (**E**) The distribution of co-expressed lncRNA survival status and survival time in model group. (**F**) Forest plot drawing for the independent prognostic value of risk lncRNAs extracted from univariate Cox regression analysis. (**G**) The survival curves of GBM patients in model group.

DEmiRNA (differentially expressed miRNA) is between seven normal tissues and sixteen GBM tissues (282 down-regulated and 119 up-regulated DEmiRNA). Using miRCode through miRNA response elements, eleven specific DEmiRNAs (three down-regulation and eight up-regulation) were detected to bind with sixteen DElncRNAs (fifteen down-regulation and one up-regulation).

In order to improve the prediction accuracy, TargetScan, SeedVicious and miRanda databases were combined to predict nine candidate DEmRNA targets for DEmiRNA. Cytoscape software was used to visualize a ceRNA network comprising sixteen lncRNAs, eleven miRNAs, and nine mRNAs based on the interactions between lncRNAs, miRNAs, and mRNAs (Figure 5).

Risk score performance, comparison and combination of gene-signature

To confirm the performance of the risk score in determining the survival rate of GBM patients, we used a model based on the prognostic dual genes (SOX6 and SOX21) signature to score the risk for each GBM patient. Risk genes (SOX6 and SOX21) expression levels were positively correlated with the infiltration levels of dendritic cells (p-value = 7.524E-08) and macrophages (p-value = 0.012) (Figure 6A). ROC curve analysis of five-year survival rate was used to evaluate the projection potential of two HMG-box-related genes. The area under the curve (AUC) of the prognostic model based on the properties of the two genes had a total survival time of 0.625 at 60 months (Figure 6B). Patients were categorized as high risk (n = 76) or low risk (n = 76), with the median risk being used as the cutoff value for survival analysis. Kaplan-Meier analysis showed that the overall survival curves of the two groups were significantly different (p-value = 1.478e-03, Figure 6C). Each patient's risk score (Figure 6D), survival status (Figure 6E), and spread of gene expression levels of both genes (Figure 6F) were also analyzed. In order to evaluate the performance of HMG-related genes as markers, we obtained two gene markers (SOX21, HR: 0.970 (95% CI: 0.942–0.999)); SOX6, HR: 0.906 (95%) to predict the prognosis of GBM patients through forest distribution maps. CI: 0.827-0.993)) (Figure 6G). Given the increasing association between immunological feature and prognosis in GBM cancer, we further explored the correlation between SOX6 and SOX21 in GBM. We explored whether SOX6 and SOX21 expression were correlated with immune infiltration levels in GBM. We measured the correlations of SOX6 and SOX21 expression with immune infiltration levels in GBM from TIMER. SOX6 expression level has significant positive correlations with infiltrating levels of purity, and significant negative correlation with dendritic cells in GBM, whereas the SOX21 expression level has significant negative correlation with neutrophil in GBM (Figure 6H). Subsequently, we further investigated the correlation between SOX6 and SOX21 gene expression in GBM patients, and the results showed that there was a significant positive correlation between SOX6 and SOX21 expression (Figure 6I). Regarding prognosis, Kaplan-Meier curves illustrated that GBM with SOX6-high had a worse prognosis than that with SOX6-low (p = 0.023), and with SOX21-high had a worse prognosis than that with SOX21-low (p = 6.17E-06) (Figure 6J). We also combined the clinical information to visualize the expression profiles of SOX6 and SOX21 and found that there is a significant difference in SOX6 only in terms of age composition (Figure 6K). These findings strongly implied that SOX6 might play a specific role in immune infiltration in different subtypes of GBM.

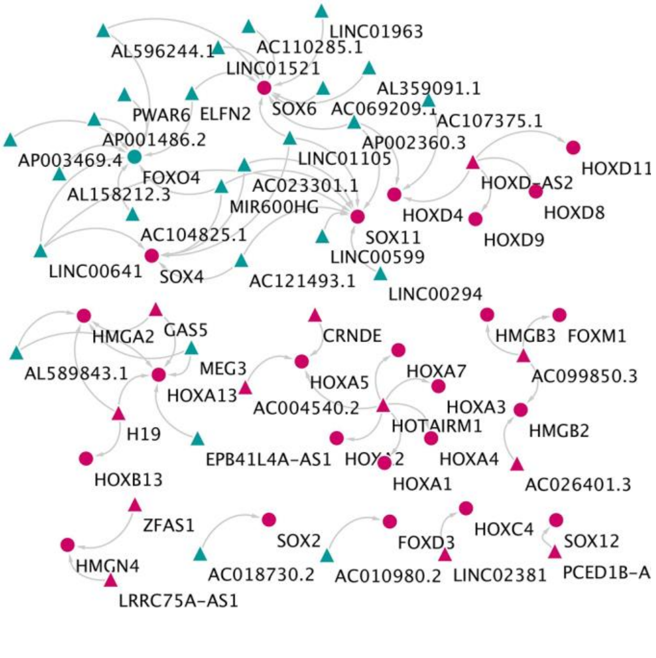


Figure 4. The network of lncRNA and HMG-box related genes co-expression. The triangles indicate lncRNAs, and circles mean mRNAs. The color green means down-regulated genes, and red means up-regulated genes.

Data from the Human Protein Atlas database showed that immunohistochemistry staining of SOX6 protein was higher in GBM cancer tissue compared with normal tissue (Figure 7).

Systematic analysis of SOX gene family and the importance of SOX6 in GBM

As a result, a total of 81 SOX members were identified in our study and divided into nine groups. Generally, the SoxA group contains SRY, SoxB1 has three members (SOX1, SOX2, SOX3), SoxB2 has two members (SOX14, SOX21), SoxC has three members (SOX4, SOX11, SOX12), SoxD contains SOX5, SOX6, SOX13, SoxE contains SOX8, SOX9, SOX10, SoxF contains SOX7, SOX17, SOX18, SoxH has SOX30, SoxG contains SOX15. As shown in Supplementary Figure 2A, the phylogenetic tree was constructed based on the SOX proteins using FastTree. We generated a graph to show the SOX protein structures by GSDS. According to the result of MEME suite, we found that there was a conserved and core motif (motif 1) in all SOX proteins, which is HMG-box domain and contains 79 amino acid residues (Supplementary Figure 2B and Supplementary Figure 2C). All motifs' logos were shown in Supplementary Figure 2B and 2C. It's noteworthy that motif 10, 2, 5 only appeared in SoxD group (SOX5, SOX6, and SOX13), they might be the domain to identify SoxD group. The SOX protein secondary structures showed in Supplementary Figure 2D and Supplementary Table 1. The secondary structures of SOX proteins were predicted by SOPM, PHD and PREDATOR methods on the NPS@, Network Protein Sequence Analysis website. For example, SOX6 was predicted to contain 38.12% α-helix, 4.45% β-sheet, and 54.89% random coil, respectively (Supplementary Table 1). By examining the properties of SOX genes for each of the four species (*Homo sapiens*, *Mus musculus*, *Coturnix japonica*, and *Gallus gallus*), the grand average of hydropathicity (GRAVY) value for those SOX genes mainly ranged from -1.080 - -0.206, which were higher than Mus musculus -1.984 - -0.207 (Supplementary Figure 2E and Supplementary Table 2). We found that the length of amino acids varied among species ranging from 204 - 817 nt. The distribution of molecular weight (Mol. Wt., kDa) for SOX genes ranged from 23.88 - 90.72. The isoelectric point (pI) of the SOX genes was from 4.91 - 9.96. According to the results of chromosome location, a total of 81 SOX gene members were mapped to the 14 chromosomes (Supplementary Figure 3).

SOX6 belongs to the SoxD group, based on the high expression in GBM, we used the GEPIA2 to obtain the top 200 co-expressed genes (Spearman's correlation >= 0.68). Then, co-expressed genes network was constructed by STRING, and re-drawn by Cytoscape (Figure 8A). To analyze the biological classification and pathway of co-expressed genes, we used Cytoscape's plugin ClueGO app for functional enrichment analyses (p-value <= 0.05). GO analysis indicated that the biological processes including tau-protein kinase activity (FYN, TTBK1, and GSK3B), intermediate filament cytoskeleton organization (FYN, DCAF1 and RAF1), negative regulation of extrinsic apoptotic signaling pathway via death domain receptors (DCAF1, RAF1 and GSK3B), histone H4 acetylation (KMT2A, MSL2 and EPC1), positive regulation of protein localization to synapse (NLGN1, NLGN2 and MAPT), microtubule polymerization or depolymerization (KIF2A and CLASP2) (Figure 8B). Consistent with enrichment of the respective cellular component and proposed molecular function (Figure 8C, 8D). Collectively, these data suggest an essential role of SOX6 in regulating cell survival and death mechanisms in GBM cancer.

The most important module was obtained using MCODE plugin (Figure 8A). We found MAPT, GSK3B, FYN and DPYSL4 as co-expressed hub genes. Hierarchical clustering of the hub genes was performed using the UCSC online tool (Figure 9A), indicating the concordant expression pattern across four genes. SOX6 compared with MAPT, GSK3B, FYN and DPYSL4 had the highest correlation coefficients (Spearman's = 0.648, 0.765, 0.693 and 0.642) in GBM compared with other tumors (Figure 9B). This data demonstrated that SOX6 had a strong association with MAPT, GSK3B, FYN and DPYSL4, suggesting that they may be functional partners in GBM.

DISCUSSION

GBM is an aggressive primary malignant brain tumor, and has one of the worst 5-year survival rates among all

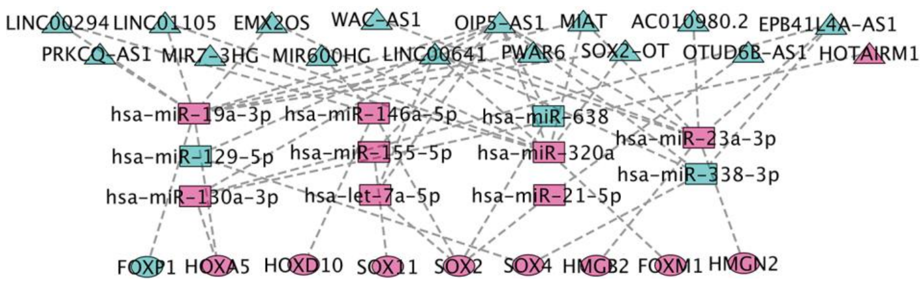


Figure 5. CeRNA network construction. The triangles represent lncRNAs, and circles mean mRNAs. The color green means down-regulated genes, and red means up-regulated genes.

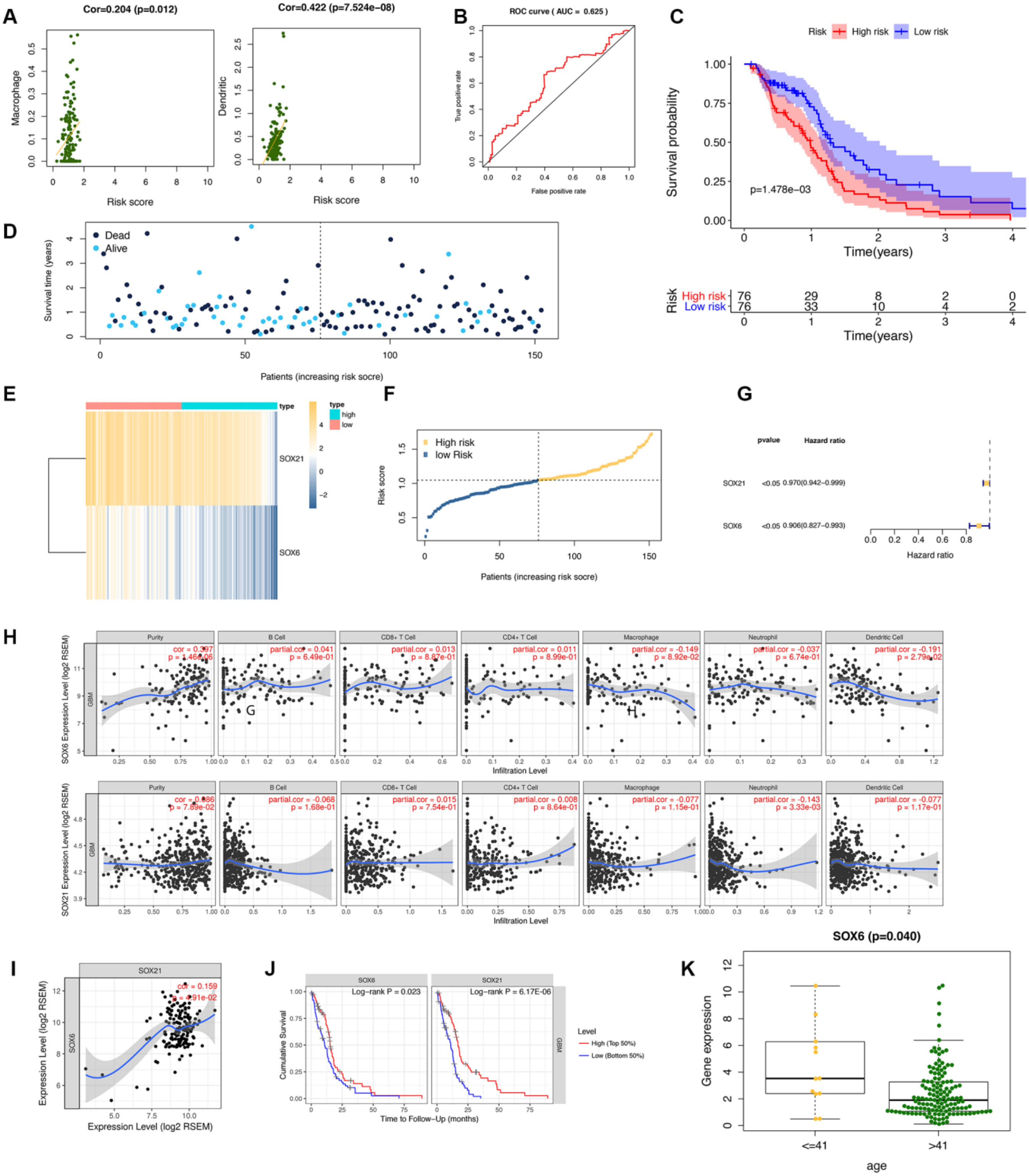


Figure 6. Diagnostic studies and survival analysis in risk genes (SOX6 and SOX21). (**A**) The correlation between the immune-infiltration abundance in risk genes (SOX6 and SOX21). (**B**) The receiver operating characteristic (ROC) curve for this model. (**C**) The survival curve is based on dividing the sample according to the median value of the risk value. (**D**–**F**) showed survival status of risk SOX6 and SOX21 among high and low risk groups. (**G**) Forest plot drawing for the independent prognostic value of risk HMG-box related obtained from univariate Cox regression analysis. (**H**) The correlation between the immune-infiltration abundance and the SOX6/SOX21 expression value. (**I**) The correlation of mRNA (SOX6, and SOX21) expression values in GBM by TIMER. (**J**) Survival curves of SOX6 and SOX21 in GBM. (**K**) The expression SOX6 between age of <=41 years and > 41 years.

cancers after diagnosis [25]. It is the first time to report the role of HMG-box gene families and their related DElncRNA and DEmiRNA in GBM. Here, we presented a description of GBMs based on the integration of the genomic and transcriptomic profiles of the HMG-box gene families (such as SOX, HOX, FOX, HMG and TOX gene families).

Through significant differential expression analysis, we found that only 62 of the 123 HMG-box genes were significantly differentially expressed in GBM, and only five genes were down-regulated and 57 were up-regulated. From the expression distribution, this showed that most DE-HMG-box genes were specifically and highly expressed in GBM and had a very important role in enhancing cancer cells growth. From the PPI network and functional pathway results in this study, we found that partial HOX genes were correlated with transcriptional misregulation in cancer, activation of anterior HOX genes in hindbrain development during early embryogenesis, and developmental genetics. DE HMG-box families were closely linked to glioma-related tumors.

Mechanisms utilizing lncRNA have been shown to take part in various types of cancer. However, a comprehensive analysis of the differential expression profiles of DE HMG-box genes co-expressed lncRNA network in GBM has been lacking. Using multivariate Cox and risk score methods, we detected an eight-lncRNA signature which was able to classify GBM patients into the high-risk group and low-risk group with significantly different overall survival (p-value = 1.604e-08). Comparing our functional analyses with DE HMG-box genes, we discovered that eight-lncRNA might participate to GBM via development biology. Then, we further made a DE-lncRNA-DEmiRNA-DE-HMG-box network to expose the HMG-box related ceRNA mechanism in GBM. For example, in our ceRNA network, we found MIR200HG-hsa-miR-146a-5p-SOX2 / HOXD10 axis in GBM, down-regulated lncRNA MIR200HG could competitively bound miRNAs (up-regulated), thereby indirectly encouraging targeted SOX2 / HOXD10 upregulation. In HMDD (the Human microRNA Disease Database) v3.2 [26], we observed that the miRNAs are linked to glioblastoma or glioma, and found several connections in miRNA-

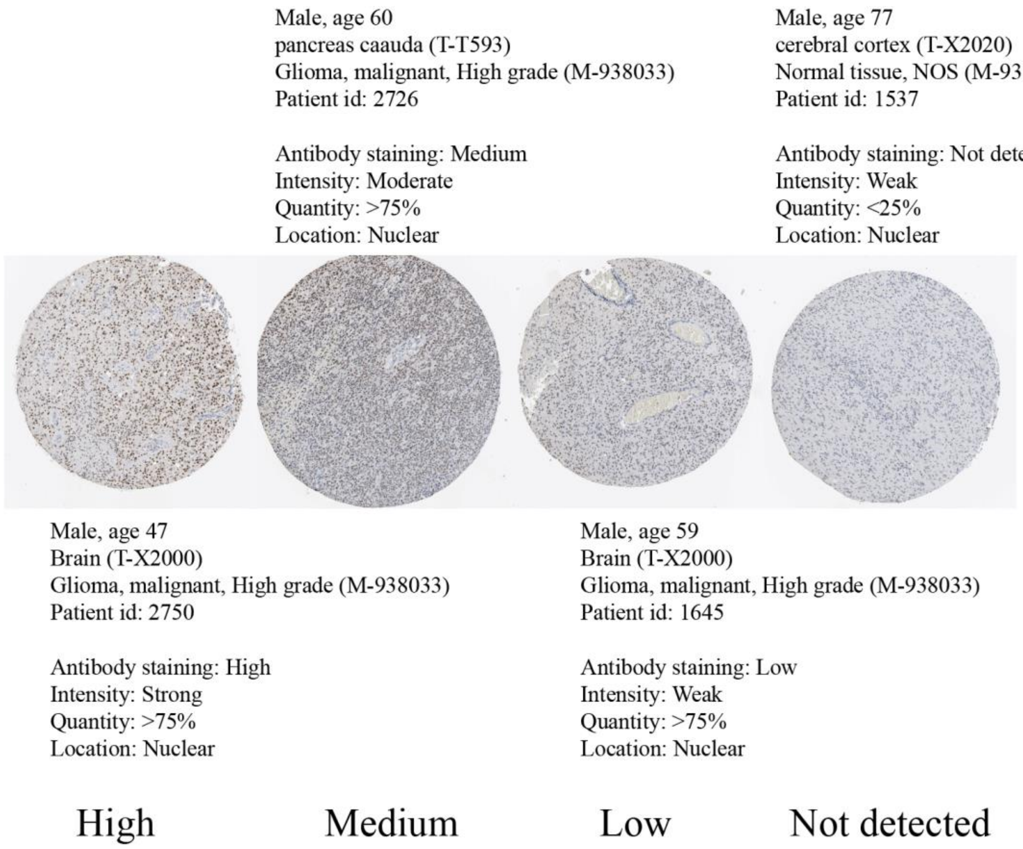


Figure 7. Immunohistochemical staining of glioma tissue taken from the Human Protein Atlas showing SOX6-negative tissue (male, age 77) and high SOX6 (male, age 47), medium SOX6 (male, age 60), and low SOX6 (male, age 59) expressing glioma tissue.

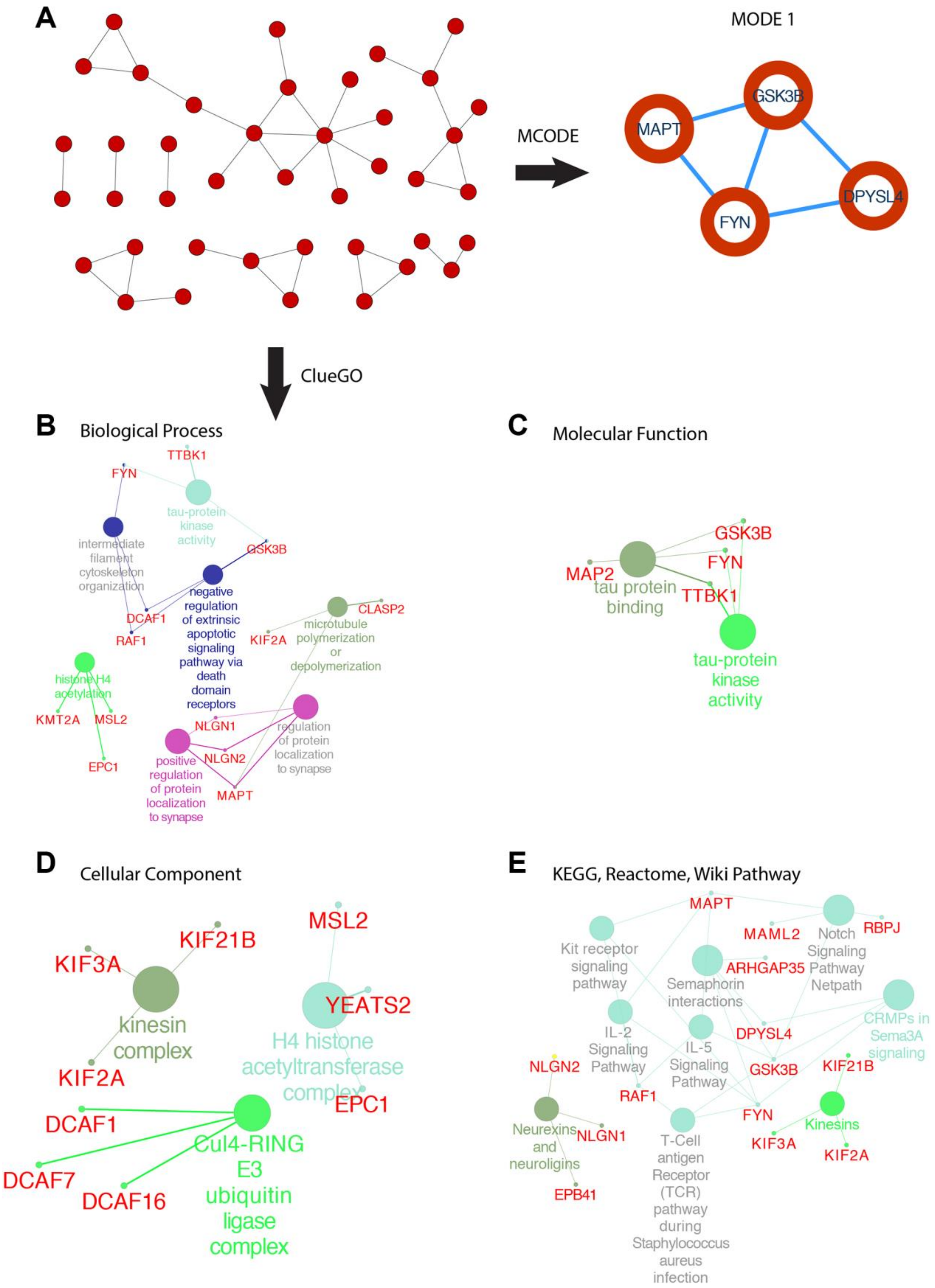

A
MODE 1
MCODE
GSK3B
MAPT
DPYSL4
FYN
ClueGO
B
Biological Process
TTBK1
FYN
tau-protein kinase activity
intermediate filament cytoskeleton organization
GSK3B
negative regulation of extrinsic apoptotic signaling pathway via death domain receptors
DCAF1
RAF1
CLASP2
KIF2A
microtubule polymerization or depolymerization
histone H4 acetylation
KMT2A
MSL2
EPC1
NLGN1
regulation of protein localization to synapse
positive regulation of protein localization to synapse
NLGN2
MAPT
C
Molecular Function
GSK3B
FYN
MAP2
tau protein binding
TTBK1
tau-protein kinase activity
D
Cellular Component
MSL2
KIF21B
KIF3A
YEATS2
kinesin complex
H4 histone acetyltransferase complex
KIF2A
EPC1
DCAF1
DCAF7
DCAF16
Cul4-RING E3 ubiquitin ligase complex
E
KEGG, Reactome, Wiki Pathway
MAPT
MAML2
Notch Signaling Pathway Netpath
RBPJ
Kit receptor signaling pathway
ARHGAP35
Semaphorin interactions
CRMPs in Sema3A signaling
IL-2 Signaling Pathway
IL-5 Signaling Pathway
DPYSL4
GSK3B
NLGN2
KIF21B
RAF1
FYN
Kinesins
Neurexins and neuroligins
NLGN1
KIF3A
KIF2A
T-Cell antigen Receptor (TCR) pathway during Staphylococcus aureus infection
EPB41

Figure 8. PPI network of SOX6 positive correlation genes and functional analysis of hub genes. (**A**) PPI network of SOX6 positive correlation genes and hub genes were found by MCODE in Cytoscape. (**B**) GO enrichment of co-expressed genes in biological process, (**C**) molecular function, (**D**) cellular component. (**E**) KEGG, Reactome, Wiki pathway enrichment analyses by ClueGO in Cytoscape.

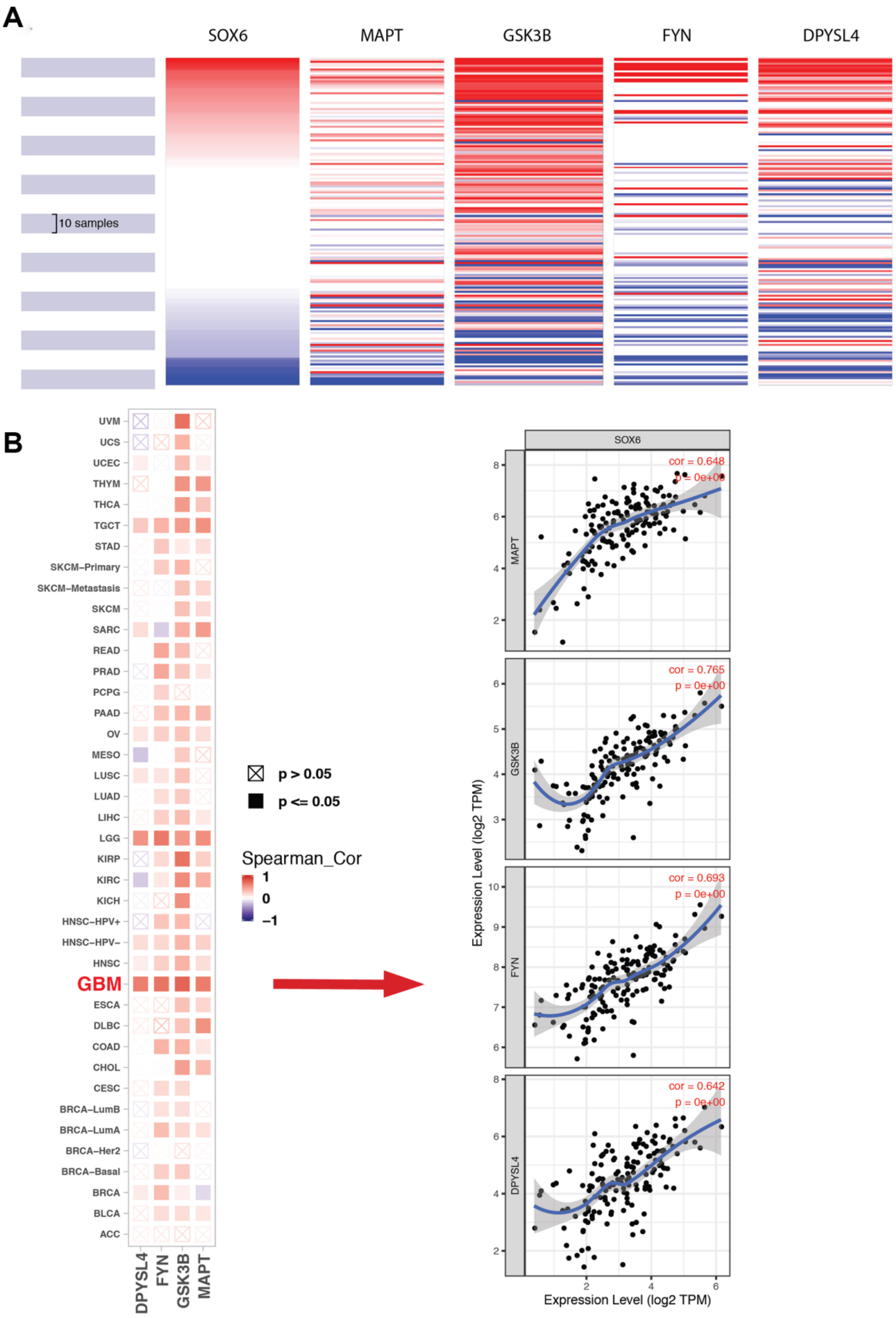


Figure 9. Expression of SOX6 and hub genes. (**A**) The hierarchical clustering of hub genes was generated by the UCSC online database. (**B**) The correlation between SOX6 and DPYSL4 (or FYN, or GSK3B, or MAPT) in different tumors, the correlation of co-expression in GBM was shown on the right by TIMER online browser.

mRNA binding. The upregulation of hsa-miR-146a and inactivation of NF-κB signaling induced the sensitization of human glioblastoma cells to TMZ-induced apoptosis by curcumin [27]. Hsa-miRNA-155 targeted FOXO3a could promote cell proliferation and invasion in glioma [28]. In this study, hsa-miRNA-155a-3p could combine OIP5-AS, and reduce the influence to SOX11, down-regulated MIAT, AC010980.2, and EPB41L4A-AS could target up-regulated hsa-miR-23a-3p, and formed a ceRNA network with up-regulated HMGB2 and HMGN2. As previously reported, Circ_PTN performed as sponge of miR-122, and activated SOX6 expression in glioma cells [29].

We systematically identified the HOX, FOX, SOX, TOX and HMG genes in humans, and also took SOX gene family as an example to comprehensive analysis of the SOX gene family from the phylogeny and protein structure. Although the absence of 3D structural data and the low accuracy of secondary structure prediction, we characterized secondary structures of SOX proteins to identify possible structural consequences of amino-acid substitutions, which indicated that these evolutionary changes have altered SOX protein function in some way.

In the current study, we initially screened out two DEmRNA (SOX6 and SOX21) of the SOX gene family that were found to be related to the clinical outcome of the GBM patients TCGA database. This study observed that up-regulated SOX6 could be targeted by eight down-regulated DElncRNAs in GBM tissues. SoxD group included SOX5, SOX6 and SOX13, we detected that SOX6 and SOX13 were over-expressed in GBM patients. SOX6 was expressed most frequently (6/7 or 86%) in GBM by RT-QPCR experiment [30], and expressed in the nuclei by immunohistochemical experiment [31]. GBM patients with low SOX6 expression present higher survival rates than those with high SOX6. FOXC1 could play the essential role in brain tumor biology and patients with GBM [32]. High expression of GATA2 connected with poor prognosis in GBM patients and promoted GBM progression by EGFR pathway [33]. SOX6 and SOX13 could co-interact with FOXC1 and GATA2, which might lead to aggressive the brain tumors. Hsa-miR-335-5p and hsa-let-7b-5p were significantly suppressed in GBM tissues [34], which targeted SOX13 in current study. We speculated that SOX6 might regulate GBM indirectly.

All the mechanism studies are aimed at finding drug targets, better prevention and treatment of diseases, there have been many studies such as a large number of alkylating anticancer agents and mutagens, and might be related to DNA replication [35–37]. As the HMG-box proteins influence DNA-dependent processes (transcription, replication, and DNA repair) [3], DE HMG-box genes and related DE-lncRNA / DE-miRNA in GBM, might serve as a potential drug target for DNA loss repair. Existing studies have shown that SOX5, SOX9 and SOX6 could be used as a drug target for INSULIN and DEXAMETHASONE, for the treatment of neurological diseases [38, 39], indicating that the combination of drugs and genes can reach the blood-brain barrier, so whether they could also be used as a drug target for solid tumor GBM, needs further research and mining by our research group.

CONCLUSIONS

In summary, our study obtained the identification of HMG-box families and established a ceRNA network in GBM by TCGA and GEO dataset, presenting them as potential therapeutic targets for the treatment of GBM. Comparing our functional analyses with DE HMG-box genes, we identified that eight-lncRNA might contribute to GBM via development biology. SOX6 and SOX21 might represent a prognostic biomarker and potential therapeutic target to improve the diagnosis and treatment of GBM. SOX6 had a strong association with MAPT, GSK3B, FYN and DPYSL4 and might be functional partners in GBM. Our study provided useful information for further exploration of GBM. Moreover, more GEO datasets should be integrated to assess and reduce the bias during the analysis process, further experiments should validate in vitro and in vivo to make the role of these key genes and pathways clear in the development of GBM.

MATERIALS AND METHODS

Database selection and gene expression analysis of GBM

The gene expression datasets of GBM were downloaded from public database. (1) The Cancer Genome Atlas (TCGA) [40] (TCGA-GBM), which collected 169 GBM tissues and 5 normal tissues, were used to screen differentially expressed mRNAs (DEmRNAs) and differentially expressed lncRNAs (DElncRNAs) between GBM and normal tissues. (2) And the Gene Expression Omnibus (GEO) dataset GSE90603 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE90603), the platform for GSE90603 datasets was GPL21572, which contained 7 samples from non-tumor samples and 16 GBM tumor samples [41], were selected to analyze differentially expressed miRNAs (DEmiRNAs) between GBM and normal tissues.

Data analyses were performed by R packages, GEO dataset was downloaded by "GEOquery" [42] and TCGA-GBM data was gained from website, differentially expressed genes (DEG) were filtered out by “limma” [43], | logFC | > 1 and q-value < 0.05. “Ggplot2” [44] and “pheatmap” [45] were used to draw diagrams. Within the HMG-box DEGs, we performed the functional enrichment analysis of Gene Ontology (GO) function using “clusterProfiler” [46]. Survival analyses, the correlations with immune infiltration levels in GBM were performed by TIMER [47]. Immunohistochemical staining of glioma tissue extracted from the Human Protein Atlas [48].

Analysis of DElncRNA-DEmRNA co-expression network

DElncRNA-DEmRNA (both genes with fold change > 2 and q-value < 0.05) co-expression network was built to determine the relationships in GBM (the absolute value of Pearson correlation coefficient > 0.5 and p-value < 0.001). The co-expression relationships were visually represented as the co-expression network using Cytoscape v3.7.2 [49].

Single and multivariate factor cox analysis, ROC and survival curve plot

In order to predict the DEmRNA co-expressed lncRNA connected to survival, we performed the single factor Cox analysis by R package “survival” [50], risk model was calculated as previously reported [51]. According to the best risk model obtained by multivariate Cox analysis, the survival score was performed, and the average number of risk scores of each sample of TCGA-GBM data was also calculated. Above-average patients belong to the high-risk group, and below-average patients belong to the low-risk group. The Kaplan-Meier method was used to draw the survival curves of the two groups.

CeRNA network construction

We then used DElncRNA, DEmiRNA and DEmRNA in this study to construct lncRNA-miRNA-mRNA associations, as previously reported [18, 52]. (1) DElncRNA-DEmiRNA interactions were described by miRcode 11 [53]. (2) Using 3’UTR regions for targeting, only DEmRNAs predicted by TargetScan [54], SeedVicious [55] and miRanda (score ≥150, MFE (minimum free energy) <−20 Kcal/mol) [56], were considered as target mRNAs, online software Venny v2.1.0 (https://bioinfogp.cnb.csic.es/tools/venny/) was used to scan the targeting DEmRNAs, and then performed the DEmiRNA-DEmRNA pairs. (3) According to the above DElncRNA-DEmiRNA and DEmiRNA-DEmRNA interactions, the visualization of the DElncRNA-DEmiRNA-DEmRNA network was built by using Cytoscape v3.7.2 [49].

Example: Genome-wide retrieval and identification of SOX gene family

The genomic and protein data of human was downloaded from NCBI (ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/001/405/GCF_000001405.39_GRCh38.p13). Then, we downloaded the HMM file of HMG-box domain with InterPro ID (IPR009071) from Pfam v32.0 [57] and ran HMMER v3.2.1 [58] to obtain the SOX genes from the complete genome with e-value cutoff 1.8e-21 as previously reported [7].

SOX genes were mapped on chromosomes by Idiographica v2.4 [59]. The theoretical molecular weight (kDa), pI (isoelectric points), amino acids length and GRAVY (Grand Average of Hydropathy) values were evaluated using the ExPASy ProtParam platform (http://web.expasy.org/protparam/) [60], and then drawn violin plots by “easyGgplot2” [61] to illustrate the comparative relationship in human, mice, chicken and quail [62].

After finding and downloading SOX sequences, we used MAFFT v7.429 [48] to align the SOX genes, and constructed ML (maximizing the tree’s likelihood) tree by FastTree v2.1 [49]. Gene structures were drawn by GSDS v2.0 [50]. Motifs reported on SOX protein data via MEME v5.0.5 [51]. SOX secondary structure was built by Secondary structure by NPS@: Network Protein Sequence Analysis online service [52].

Data mining for SOX6 and co-expressed hub genes

GEPIA2 (Gene Expression Profiling Interactive Analysis) [68] was used to analyze the gene expression correlation by TCGA-GBM data. The Spearman method was used to find the correlation coefficient. PPI network was constructed by STRING v11 [69]. Cytoscape’s plugin CluGO was used for functional enrichment analyses (GO, KEGG, Reactome, and Wiki pathway) [70]. Cytoscape’s plugin MCODE was applied for finding linked regions based on topology (MCODE score > 5, degree cutoff =2, node score cutoff = 0.2, max depth = 100, k-score =2). We plotted the heat map of SOX6 and co-expressed hub genes by University of California Santa Cruz (UCSC) browser [71], and the correlation of these genes was drawn by TIMER [47].

AUTHOR CONTRIBUTIONS

Data curation, Lan Jiang, Xiaolong Zhu and Kun Lv; Formal analysis, Lan Jiang, Xiaolong Zhu and Jingjing Ye; Funding acquisition, Hui Yang, Tianbing Chen and Kun Lv; Validation, Lan Jiang, Hui Yang and Tianbing

Chen; Writing – original draft, Lan Jiang and Kun Lv; Writing – review and editing, Lan Jiang.

ACKNOWLEDGMENTS

We thank Aging Editor-in-Chief Mikhail Blagosklonny and our anonymous reviewers for their helpful comments and guidance throughout the review process.

CONFLICTS OF INTEREST

The authors declare no conflicts of interests.

FUNDING

This project was supported by the National Natural Science Foundation of China (grant nos. 81701557, 81802503, 81870017, 81772180 and 81901519), Key Projects of Natural Science Research of Universities in Anhui Province (grant no. KJ2018A0265). Anhui Provincial Natural Science Foundation (grant no. 1908085QH380). Research Project Training Fund of Wannan Medical College (grant no. WK2018ZF02).

REFERENCES

1. Bunda S, Heir P, Metcalf J, Li AS, Agnihotri S, Pusch S, Yasin M, Li M, Burrell K, Mansouri S, Singh O, Wilson M, Alamsahebpour A, et al. CIC protein instability contributes to tumorigenesis in glioblastoma. Nat Commun. 2019; 10:661. https://doi.org/10.1038/s41467-018-08087-9 PMID:30737375

2. Reiß S, Tomiuk S, Kollet J, Drewes J, Brück W, Jungblut M, Bosio A. Characterization and classification of glioblastoma multiforme using the novel multiparametric cyclic immunofluorescence analysis system MACSima. AACR. 2019. https://doi.org/10.1158/1538-7445.AM2019-245

3. Štros M, Launholt D, Grasser KD. The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. Cell Mol Life Sci. 2007; 64:2590–606. https://doi.org/10.1007/s00018-007-7162-3 PMID:17599239

4. Chen SL, Qin ZY, Hu F, Wang Y, Dai YJ, Liang Y. The Role of the *HOXA* Gene Family in Acute Myeloid Leukemia. Genes (Basel). 2019; 10:621. https://doi.org/10.3390/genes10080621 PMID:31426381

5. Morgan R, Primon M, Shnyder S, Short S, Kaur B, Hong B, Bagwan I, Rogers W, Pandha HS. Targeting of HOX-PBX binding in glioblastoma multiforme as a novel therapeutic treatment. AACR. 2019. https://doi.org/10.1158/1538-7445.AM2019-5215

6. Hu X, Chen D, Cui Y, Li Z, Huang J. Targeting microRNA-23a to inhibit glioma cell invasion via HOXD10. Sci Rep. 2013; 3:3423. https://doi.org/10.1038/srep03423 PMID:24305689

7. Jiang L, Bi D, Ding H, Wu X, Zhu R, Zeng J, Yang X, Kan X. Systematic Identification and Evolution Analysis of *Sox* Genes in *Coturnix japonica* Based on Comparative Genomics. Genes (Basel). 2019; 10:314. https://doi.org/10.3390/genes10040314 PMID:31013663

8. Takezaki T, Hide T, Takanaga H, Nakamura H, Kuratsu J, Kondo T. Essential role of the Hedgehog signaling pathway in human glioma-initiating cells. Cancer Sci. 2011; 102:1306–12. https://doi.org/10.1111/j.1349-7006.2011.01943.x PMID:21453386

9. Tian R, Wang J, Yan H, Wu J, Xu Q, Zhan X, Gui Z, Ding M, He J. Differential expression of miR16 in glioblastoma and glioblastoma stem cells: their correlation with proliferation, differentiation, metastasis and prognosis. Oncogene. 2017; 36:5861–73. https://doi.org/10.1038/onc.2017.182 PMID:28628119

10. Ueda R, Iizuka Y, Yoshida K, Kawase T, Kawakami Y, Toda M. Identification of a human glioma antigen, SOX6, recognized by patients' sera. Oncogene. 2004; 23:1420–27. https://doi.org/10.1038/sj.onc.1207252 PMID:14691456

11. Mondal SK, Sen MK. An in-silico characterization of Sry-related HMG box C (SOXC) in humans and mouse. Meta Gene. 2019; 19:235–45. https://doi.org/10.1016/j.mgene.2018.12.012

12. Xu X, Wang Z, Liu N, Cheng Y, Jin W, Zhang P, Wang X, Yang H, Liu H, Zhang Y, Tu Y. Association between SOX9 and CA9 in glioma, and its effects on chemosensitivity to TMZ. Int J Oncol. 2018; 53:189–202. https://doi.org/10.3892/ijo.2018.4382 PMID:29749469

13. Caglayan D, Lundin E, Kastemar M, Westermark B, Ferletta M. Sox21 inhibits glioma progression in vivo by forming complexes with Sox2 and stimulating aberrant differentiation. Int J Cancer. 2013; 133:1345–56. https://doi.org/10.1002/ijc.28147 PMID:23463365

14. Bulstrode H, Johnstone E, Marques-Torrejon MA, Ferguson KM, Bressan RB, Blin C, Grant V, Gogolok S, Gangoso E, Gagrica S, Ender C, Fotaki V, Sproul D, et al. Elevated FOXG1 and SOX2 in glioblastoma enforces neural stem cell identity through transcriptional control of cell cycle and epigenetic regulators. Genes Dev. 2017; 31:757–73.

https://doi.org/10.1101/gad.293027.116 PMID:28465359

15. Liu F, Hon GC, Villa GR, Turner KM, Ikegami S, Yang H, Ye Z, Li B, Kuan S, Lee AY, Zanca C, Wei B, Lucey G, et al. EGFR mutation promotes glioblastoma through epigenome and transcription factor network remodeling. Mol Cell. 2015; 60:307–18. https://doi.org/10.1016/j.molcel.2015.09.002 PMID:26455392

16. Ma XL, Shang F, Ni W, Zhu J, Luo B, Zhang YQ. MicroRNA-338-5p plays a tumor suppressor role in glioma through inhibition of the MAPK-signaling pathway by binding to FOXD1. J Cancer Res Clin Oncol. 2018; 144:2351–66. https://doi.org/10.1007/s00432-018-2745-y PMID:30225541

17. Jiang L, Bi D, Ding H, Ren Q, Wang P, Kan X. Identification and comparative profiling of gonadal microRNAs in the adult pigeon (*Columba livia*). Br Poult Sci. 2019; 60:638–48. https://doi.org/10.1080/00071668.2019.1639140 PMID:31343256

18. Jiang L, Wang Q, Yu J, Gowda V, Johnson G, Yang J, Kan X, Yang X. miRNAome expression profiles in the gonads of adult *Melopsittacus undulatus.* PeerJ. 2018; 6:e4615. https://doi.org/10.7717/peerj.4615 PMID:29666766

19. Liu X, Zheng J, Xue Y, Qu C, Chen J, Wang Z, Li Z, Zhang L, Liu Y. Inhibition of TDP43-mediated SNHG12-miR-195-SOX5 feedback loop impeded malignant biological behaviors of glioma cells. Mol Ther Nucleic Acids. 2018; 10:142–58. https://doi.org/10.1016/j.omtn.2017.12.001 PMID:29499929

20. Xu YR, Yang WX. SOX-mediated molecular crosstalk during the progression of tumorigenesis. Semin Cell Dev Biol. 2017; 63:23–34. https://doi.org/10.1016/j.semcdb.2016.07.028 PMID:27476113

21. Hao Y, Zhang S, Sun S, Zhu J, Xiao Y. MiR-595 targeting regulation of SOX7 expression promoted cell proliferation of human glioblastoma. Biomed Pharmacother. 2016; 80:121–26. https://doi.org/10.1016/j.biopha.2016.03.008 PMID:27133048

22. Xiuju C, Zhen W, Yanchao S. SOX7 inhibits tumor progression of glioblastoma and is regulated by miRNA-24. Open Med (Wars). 2016; 11:133–37. https://doi.org/10.1515/med-2016-0026 PMID:28352781

23. Yang J, Hu F, Fu X, Jiang Z, Zhang W, Chen K. MiR-128/SOX7 alleviates myocardial ischemia injury by regulating IL-33/sST2 in acute myocardial infarction. Biol Chem. 2019; 400:533–44. https://doi.org/10.1515/hsz-2018-0207 PMID:30265647

24. Bai QL, Hu CW, Wang XR, Shang JX, Yin GF. MiR-616 promotes proliferation and inhibits apoptosis in glioma cells by suppressing expression of SOX7 via the Wnt signaling pathway. Eur Rev Med Pharmacol Sci. 2017; 21:5630–37. https://doi.org/10.26355/eurrev_201712_14006 PMID:29271996

25. Aguila B, Morris AB, Spina R, Bar E, Schraner J, Vinkler R, Sohn JW, Welford SM. The Ig superfamily protein PTGFRN coordinates survival signaling in glioblastoma multiforme. Cancer Lett. 2019; 462:33–42. https://doi.org/10.1016/j.canlet.2019.07.018 PMID:31377205

26. Huang Z, Shi J, Gao Y, Cui C, Zhang S, Li J, Zhou Y, Cui Q. HMDD v3.0: a database for experimentally supported human microRNA-disease associations. Nucleic Acids Res. 2019; 47:D1013–17. https://doi.org/10.1093/nar/gky1010 PMID:30364956

27. Wu H, Liu Q, Cai T, Chen YD, Wang ZF. Induction of microRNA-146a is involved in curcumin-mediated enhancement of temozolomide cytotoxicity against human glioblastoma. Mol Med Rep. 2015; 12:5461–66. https://doi.org/10.3892/mmr.2015.4087 PMID:26239619

28. Osei-Sarfo K, Gudas LJ. Retinoids induce antagonism between FOXO3A and FOXM1 transcription factors in human oral squamous cell carcinoma (OSCC) cells. PLoS One. 2019; 14:e0215234. https://doi.org/10.1371/journal.pone.0215234 PMID:30978209

29. Chen C, Deng L, Nie DK, Jia F, Fu LS, Wan ZQ, Lan Q. Circular RNA Pleiotrophin promotes carcinogenesis in glioma via regulation of microRNA-122/SRY-box transcription factor 6 axis. Eur J Cancer Prev. 2020; 29:165–73. https://doi.org/10.1097/CEJ.0000000000000535 PMID:31609809

30. Lee MH, Son EI, Kim E, Kim IS, Yim MB, Kim SP. Expression of cancer-testis genes in brain tumors. J Korean Neurosurg Soc. 2008; 43:190–93. https://doi.org/10.3340/jkns.2008.43.4.190 PMID:19096642

31. Ueda R, Yoshida K, Kawakami Y, Kawase T, Toda M. Immunohistochemical analysis of SOX6 expression in human brain tumors. Brain Tumor Pathol. 2004; 21:117–20. https://doi.org/10.1007/BF02482186 PMID:15696972

32. Robertson E, Perry C, Doherty R, Madhusudan S. Transcriptomic profiling of Forkhead box transcription factors in adult glioblastoma multiforme. Cancer Genomics Proteomics. 2015; 12:103–12. PMID:25977169

33. Wang Z, Yuan H, Sun C, Xu L, Chen Y, Zhu Q, Zhao H, Huang Q, Dong J, Lan Q. GATA2 promotes glioma progression through EGFR/ERK/Elk-1 pathway. Med Oncol. 2015; 32:87. https://doi.org/10.1007/s12032-015-0522-1 PMID:25707769

34. Wu HM, Wang HD, Tang Y, Fan YW, Hu YB, Tohti M, Hao XK, Wei WT, Wu Y. Differential expression of microRNAs in postoperative radiotherapy sensitive and resistant patients with glioblastoma multiforme. Tumour Biol. 2015; 36:4723–30. https://doi.org/10.1007/s13277-015-3121-z PMID:25758051

35. Kou Y, Koag MC, Cheun Y, Shin A, Lee S. Application of hypoiodite-mediated aminyl radical cyclization to synthesis of solasodine acetate. Steroids. 2012; 77:1069–74. https://doi.org/10.1016/j.steroids.2012.05.002 PMID:22583912

36. Kou Y, Koag MC, Lee S. N7 methylation alters hydrogen-bonding patterns of guanine in duplex DNA. J Am Chem Soc. 2015; 137:14067–70. https://doi.org/10.1021/jacs.5b10172 PMID:26517568

37. Kou Y, Koag MC, Lee S. Structural and kinetic studies of the effect of guanine N7 alkylation and metal cofactors on DNA replication. Biochemistry. 2018; 57:5105–16. https://doi.org/10.1021/acs.biochem.8b00331 PMID:29957995

38. Kupcsik L, Stoddart MJ, Li Z, Benneker LM, Alini M. Improving chondrogenesis: potential and limitations of SOX9 gene transfer and mechanical stimulation for cartilage tissue engineering. Tissue Eng Part A. 2010; 16:1845–55. https://doi.org/10.1089/ten.tea.2009.0531 PMID:20067399

39. Renard E, Porée B, Chadjichristos C, Kypriotou M, Maneix L, Bigot N, Legendre F, Ollitrault D, De Crombrugghe B, Malléin-Gérin F, Moslemi S, Demoor M, Boumediene K, Galéra P. Sox9/Sox6 and Sp1 are involved in the insulin-like growth factor-I-mediated upregulation of human type II collagen gene expression in articular chondrocytes. J Mol Med (Berl). 2012; 90:649–66. https://doi.org/10.1007/s00109-011-0842-3 PMID:22215151

40. Tomczak K, Czerwińska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn). 2015; 19:A68–77. https://doi.org/10.5114/wo.2014.47136 PMID:25691825

41. Guo X, Luo Z, Xia T, Wu L, Shi Y, Li Y. Identification of miRNA signature associated with BMP2 and chemosensitivity of TMZ in glioblastoma stem-like cells. Genes Dis. 2019. https://doi.org/10.1016/j.gendis.2019.09.002

42. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007; 23:1846–47. https://doi.org/10.1093/bioinformatics/btm254 PMID:17496320

43. Smyth GK. Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor. Springer; 2005. pp. 397–420. https://doi.org/10.1007/0-387-29362-0_23

44. Wickham H. ggplot2. Wiley Interdiscip Rev Comput Stat. 2011; 3:180–85. https://doi.org/10.1002/wics.147

45. Kolde R, Kolde MR. Package ‘pheatmap. R Package; 2015. p. 1.

46. Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. OMICS. 2012; 16:284–87. https://doi.org/10.1089/omi.2011.0118 PMID:22455463

47. Li T, Fan J, Wang B, Traugh N, Chen Q, Liu JS, Li B, Liu XS. TIMER: a web server for comprehensive analysis of tumor-infiltrating immune cells. Cancer Res. 2017; 77:e108–10. https://doi.org/10.1158/0008-5472.CAN-17-0307 PMID:29092952

48. Uhlen M, Zhang C, Lee S, Sjöstedt E, Fagerberg L, Bidkhori G, Benfeitas R, Arif M, Liu Z, Edfors F, Sanli K, von Feilitzen K, Oksvold P, et al. A pathology atlas of the human cancer transcriptome. Science. 2017; 357:eaan2507. https://doi.org/10.1126/science.aan2507 PMID:28818916

49. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003; 13:2498–504. https://doi.org/10.1101/gr.1239303 PMID:14597658

50. Therneau TM, Lumley T. Package ‘survival. R Top Doc; 2015. p. 128.

51. Liu Z, Li M, Hua Q, Li Y, Wang G. Identification of an eight-lncRNA prognostic model for breast cancer using WGCNA network analysis and a Cox-proportional hazards model based on L1-penalized estimation. Int J Mol Med. 2019; 44:1333–43. https://doi.org/10.3892/ijmm.2019.4303 PMID:31432096

52. Yao Y, Zhang T, Qi L, Liu R, Liu G, Wang J, Song Q, Sun C. Comprehensive analysis of prognostic biomarkers in lung adenocarcinoma based on aberrant lncRNA-miRNA-mRNA networks and Cox regression models. Biosci Rep. 2020; 40:40. https://doi.org/10.1042/BSR20191554 PMID:31950990

53. Jeggari A, Marks DS, Larsson E. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. Bioinformatics. 2012; 28:2062–63. https://doi.org/10.1093/bioinformatics/bts344 PMID:22718787

54. Agarwal V, Bell GW, Nam JW, Bartel DP. Predicting effective microRNA target sites in mammalian mRNAs. Elife. 2015; 4:e05005. https://doi.org/10.7554/eLife.05005 PMID:26267216

55. Marco A. SeedVicious: analysis of microRNA target and near-target sites. PLoS One. 2018; 13:e0195532. https://doi.org/10.1371/journal.pone.0195532 PMID:29664927

56. John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. PLoS Biol. 2004; 2:e363. https://doi.org/10.1371/journal.pbio.0020363 PMID:15502875

57. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer EL, Hirsh L, Paladin L, et al. The Pfam protein families database in 2019. Nucleic Acids Res. 2019; 47:D427–32. https://doi.org/10.1093/nar/gky995 PMID:30357350

58. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. Nucleic Acids Res. 2018; 46:W200–04. https://doi.org/10.1093/nar/gky448 PMID:29905871

59. Kin T, Ono Y. Idiographica: a general-purpose web application to build idiograms on-demand for human, mouse and rat. Bioinformatics. 2007; 23:2945–46. https://doi.org/10.1093/bioinformatics/btm455 PMID:17893084

60. Gasteiger E, Hoogland C, Gattiker A, Wilkins MR, Appel RD, Bairoch A. Protein identification and analysis tools on the ExPASy server. The proteomics protocols handbook. Springer; 2005. pp. 571–607. https://doi.org/10.1385/1-59259-890-0:571

61. Kassambara A. (2014). easyGgplot2: Perform and customize easily a plot with ggplot2. R package version 1.0. 0.9000.

62. Crutchley B, Wang Z, Wang Q, Klinke DJ. How to use R to generate Violin Plots to assist in sensitivity analysis of a cancer model. Proceedings of the West Virginia Academy of Science. 2018; 90.

63. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K. MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res. 2019; 47:W5–10. https://doi.org/10.1093/nar/gkz342 PMID:31062021

64. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS One. 2010; 5:e9490. https://doi.org/10.1371/journal.pone.0009490 PMID:20224823

65. Hu B, Jin J, Guo AY, Zhang H, Luo J, Gao G. GSDS 2.0: an upgraded gene feature visualization server. Bioinformatics. 2015; 31:1296–97. https://doi.org/10.1093/bioinformatics/btu817 PMID:25504850

66. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME Suite. Nucleic Acids Res. 2015; 43:W39–49. https://doi.org/10.1093/nar/gkv416 PMID:25953851

67. Combet C, Blanchet C, Geourjon C, Deléage G. NPS@: network protein sequence analysis. Trends Biochem Sci. 2000; 25:147–50. https://doi.org/10.1016/S0968-0004(99)01540-6 PMID:10694887

68. Tang Z, Kang B, Li C, Chen T, Zhang Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. Nucleic Acids Res. 2019; 47:W556–60. https://doi.org/10.1093/nar/gkz430 PMID:31114875

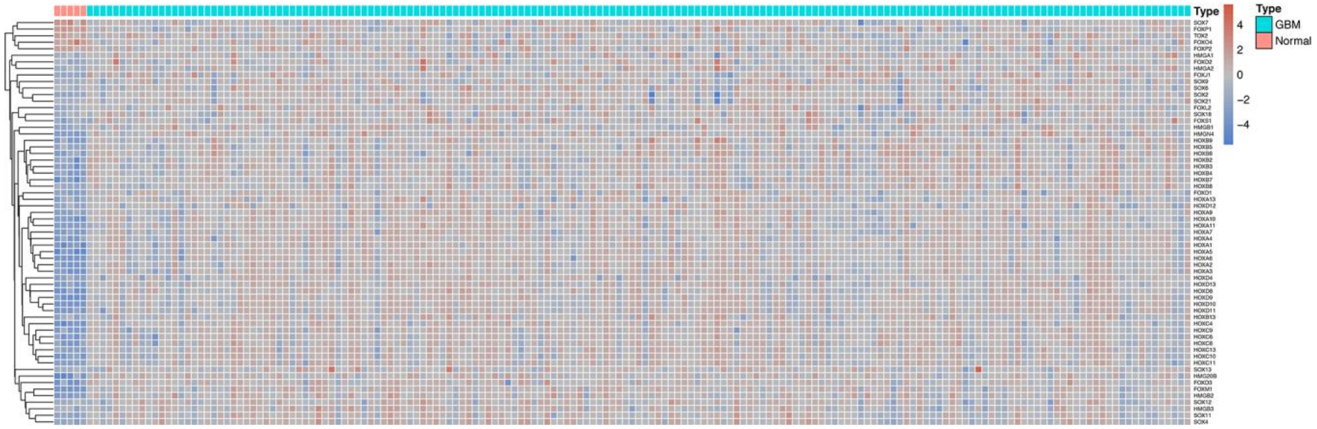
69. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, Simonovic M, Doncheva NT, Morris JH, Bork P, Jensen LJ, Mering CV. STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res. 2019; 47:D607–13. https://doi.org/10.1093/nar/gky1131 PMID:30476243

70. Lee IY, Ho JM, Chen MS. (2005). CLUGO: a clustering algorithm for automated functional annotations based on gene ontology. Fifth IEEE International Conference on Data Mining (ICDM'05): IEEE), pp. 4.

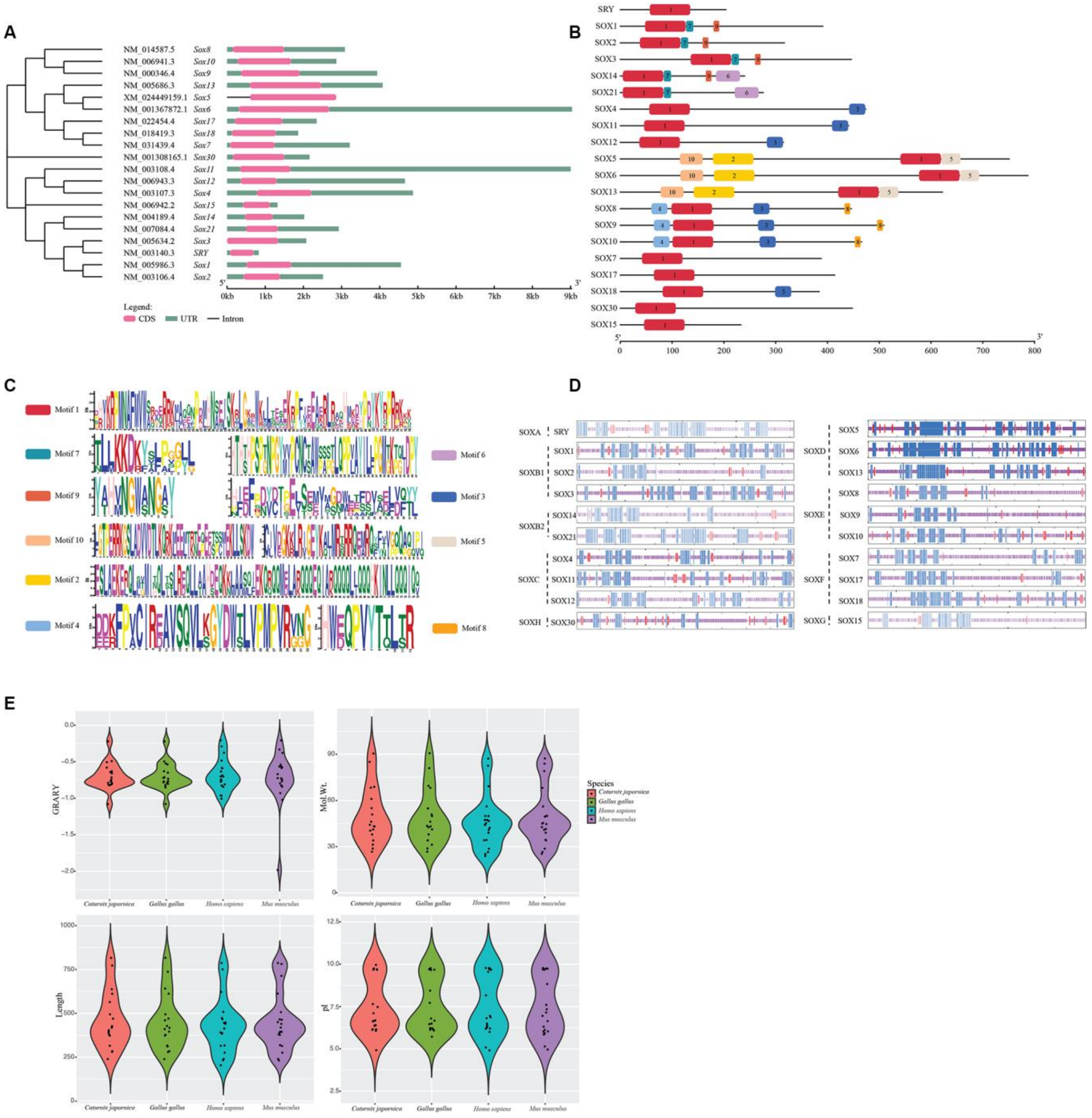
71. Lee CM, Barber GP, Casper J, Clawson H, Diekhans M, Gonzalez JN, Hinrichs AS, Lee BT, Nassar LR, Powell CC, Raney BJ, Rosenbloom KR, Schmelter D, et al. UCSC Genome Browser enters 20th year. Nucleic Acids Res. 2020; 48:D756–61.
https://doi.org/10.1093/nar/gkz1012
PMID:31691824

SUPPLEMENTARY MATERIALS

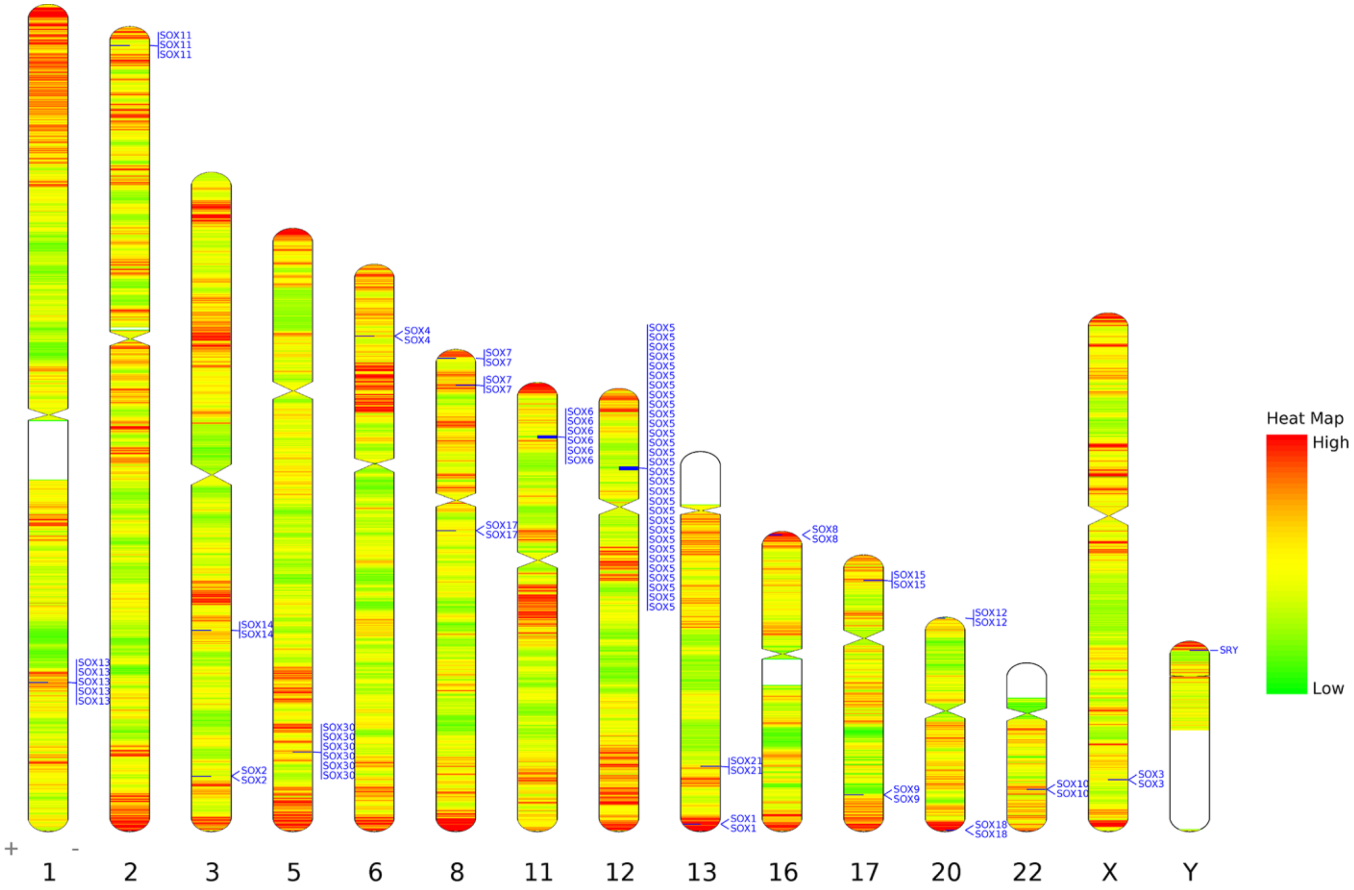
Supplementary Figures



Supplementary Figure 1. Heatmap for HMG-box related genes between GBM and normal tissue in TCGA. Starting from the left, the first 5 datasets were normal tissues, and the remaining 169 were GBM tissues.



Supplementary Figure 2. Structures analyses of SOX gene family in humans. (**A**) Phylogenetic and protein structures analyses of SOX gene family in humans. (**B** and **C**) SOX motif prediction. (**D**) SOX protein secondary structures in humans. The α-helix, β-sheet and disordered loop regions are drawn in blue, red and purple, respectively. (**E**) Protein properties for SOX genes identified from *Homo sapiens, Mus musculus, Coturnix japonica*, and *Gallus gallus*.



Supplementary Figure 3. Distribution of the SOX gene family on human chromosomes.

Supplementary Tables

Supplementary Table 1. The composition of SOX protein secondary structures in humans.

Group	Type	Alpha helix (%)	Beta sheets (%)	Random coil (%)
SOXA	SRY	36.27	3.92	56.37
	SOX1	33.76	4.6	58.57
	SOX2	15.14	3.15	79.18
	SOX3	31.39	3.14	62.33
	SOX14	21.25	2.92	73.33
	SOX21	37.68	4.35	55.8
	SOX4	24.05	4.64	67.51
	SOX11	28.12	5.22	62.59
	SOX12	28.25	1.9	66.67
	SOX5	35.02	5.46	56.86
	SOX6	38.12	4.45	54.89
	SOX13	30.87	4.18	63.34
	SOX8	15.7	3.81	79.37
	SOX9	14.93	1.77	81.73
	SOX10	19.74	6.01	71.03
	SOX7	22.94	0.77	74.23
	SOX17	21.26	2.17	74.4
	SOX18	28.91	2.34	67.97
SOXH	SOX30	11.83	7.81	77.68
SOXG	SOX15	19.31	1.72	78.54

Supplementary Table 2. Predictions for protein properties for SOX from four species.

pI	GRAVY	Length	Mol.Wt.	Species
9.7	-0.375	391	39.02	*Homo_sapiens*
6.19	-0.825	466	49.91	*Homo_sapiens*
4.91	-0.702	441	46.68	*Homo_sapiens*
5.08	-0.96	315	34.12	*Homo_sapiens*
6.25	-0.766	622	69.23	*Homo_sapiens*
9.68	-0.585	240	26.49	*Homo_sapiens*
9.78	-0.843	233	25.25	*Homo_sapiens*
6	-0.633	414	44.12	*Homo_sapiens*
8.16	-0.589	384	40.89	*Homo_sapiens*
9.74	-0.742	317	34.31	*Homo_sapiens*
9.74	-0.206	276	28.58	*Homo_sapiens*
9.78	-0.29	446	45.21	*Homo_sapiens*
6.81	-0.701	448	49.88	*Homo_sapiens*
6.87	-0.483	474	47.26	*Homo_sapiens*
6.38	-0.745	751	82.58	*Homo_sapiens*
6.95	-0.809	787	87.24	*Homo_sapiens*
6.2	-0.687	388	42.20	*Homo_sapiens*
6.49	-0.77	446	47.31	*Homo_sapiens*
6.31	-1.007	509	56.14	*Homo_sapiens*
9.55	-0.968	204	23.88	*Homo_sapiens*
9.7	-0.379	391	39.05	*Mus_musculus*
9.74	-0.724	319	34.41	*Mus_musculus*
9.78	-0.332	450	45.44	*Mus_musculus*
7.15	-0.539	440	45.04	*Mus_musculus*
5.92	-0.767	714	79.13	*Mus_musculus*
6.95	-0.824	787	87.32	*Mus_musculus*
6.01	-0.671	380	41.49	*Mus_musculus*
6.64	-0.796	464	49.88	*Mus_musculus*
6.31	-1.021	507	56.08	*Mus_musculus*
6.12	-0.827	466	49.95	*Mus_musculus*
4.96	-0.727	395	42.63	*Mus_musculus*
5.14	-0.929	314	34.08	*Mus_musculus*
6.03	-0.745	613	68.17	*Mus_musculus*
9.68	-0.587	240	26.52	*Mus_musculus*
9.68	-0.842	231	25.31	*Mus_musculus*
5.85	-0.569	419	44.65	*Mus_musculus*
7.6	-0.565	377	40.90	*Mus_musculus*
9.74	-0.207	276	28.61	*Mus_musculus*
8.83	-0.575	782	83.94	*Mus_musculus*
7.34	-1.984	395	49.49	*Mus_musculus*
9.7	-0.495	373	37.93	*Gallus_gallus*
9.77	-0.78	312	34.10	*Gallus_gallus*
9.66	-0.712	316	34.01	*Gallus_gallus*
9.68	-0.63	240	26.67	*Gallus_gallus*
9.74	-0.22	280	28.80	*Gallus_gallus*
7.71	-0.65	428	43.11	*Gallus_gallus*

5.72	-0.818	396	43.50	*Gallus_gallus*
8.44	-0.72	285	31.26	*Gallus_gallus*
6.13	-0.733	737	80.99	*Gallus_gallus*
6.66	-0.825	817	90.72	*Gallus_gallus*
6.13	-0.801	612	68.18	*Gallus_gallus*
6.46	-0.788	470	50.83	*Gallus_gallus*
6.23	-1.078	494	54.85	*Gallus_gallus*
6.2	-0.851	461	49.86	*Gallus_gallus*
6.21	-0.524	410	42.60	*Gallus_gallus*
6.49	-0.766	418	46.23	*Gallus_gallus*
6.79	-0.769	377	41.22	*Gallus_gallus*
6.09	-0.536	642	69.47	*Gallus_gallus*
9.7	-0.505	373	37.92	*Coturnix_japornica*
9.96	-0.641	404	44.22	*Coturnix_japornica*
9.68	-0.715	316	34.03	*Coturnix_japornica*
9.68	-0.63	240	26.67	*Coturnix_japornica*
9.74	-0.218	280	28.77	*Coturnix_japornica*
7.1	-0.658	427	43.11	*Coturnix_japornica*
4.92	-0.77	396	43.18	*Coturnix_japornica*
7.65	-0.805	285	31.27	*Coturnix_japornica*
6.15	-0.746	773	85.08	*Coturnix_japornica*
6.66	-0.817	816	90.57	*Coturnix_japornica*
6.13	-0.81	612	68.32	*Coturnix_japornica*
6.37	-0.82	470	50.90	*Coturnix_japornica*
6.16	-1.08	495	55.04	*Coturnix_japornica*
7.47	-0.788	565	60.97	*Coturnix_japornica*
6.68	-0.58	387	40.39	*Coturnix_japornica*
6.4	-0.791	418	46.22	*Coturnix_japornica*
6.62	-0.759	377	41.16	*Coturnix_japornica*
6.07	-0.49	638	68.85	*Coturnix_japornica*