Analysis of Gender-Specific Regulatory Mechanisms on the Oncogenesis and Prognosis of Glioblastoma Multiforme

Jianzhi Deng¹, Yuehan Zhou²* and Wenqing Long²

¹ Guilin University of Technology, Guilin 541004, China.
² Guilin Medical University, Guilin 541004, Guangxi, China.
Email: dengjianzhi@163.com, *Corresponding author: yuehanzhou2012@163.com

Abstract. In this paper, we focus on finding out the key mRNAs and the key regulatory mechanism of the different morbidity of glioblastoma multiforme (GBM) in different genders. 160 GBM patients’ RNA expressed data and clinical data were downloaded from the TCGA GBM project. 103 differential expressed mRNAs (DEmRNAs) between male and female were screened from the RNA matrix by clustering analysis. In the gene ontology (GO) analysis, the DEmRNAs were enriched in 26 GOterms, including 17 biological process (BP), 4 cellular component (CC), 5 molecular function (MF). Among them, AGTR2, CALCA, CALCB, CTSG, GCG, GCGR, HCRT, PRL and RXFP4 were found that enriched in signal pathway hsa04080. A multivariate COX model was constructed by these 9 DEmRNAs. And GCGR, HCRT and CTSG were found as the co-expression mRNAs. In the clinical data analysis, the kmplot survival curve (p = 1.63e-03) indicated that the co-expression mRNAs and the hsa04080 signal pathway would be built the potential regulatory mechanism for the treatment of GBM in different genders’ patients. And it was then proved by receiver operating characteristic (ROC) curve and area under the curve (AUC) value (AUC=0.766).

1. Introduction
GBM is one of the highly malignant brain disease. Former research for the GBM was shown that the morbidity and mortality of male was higher than female [1, 2]. At the same time, it could be known that genes and regulatory mechanisms were closely associated with diseases. In order to find out the relationship between the situation of gender difference and gene expression, we did the research by analyzing the downloaded GBM RNA-seq data from TCGA GBM project. The DEmRNAs between the different genders were screened by limma package [3]. And based on the DEmRNAs, the GO enriched analysis was done by Database for Annotation, Visualization and Integrated Discovery (DAVID) [4] while the KEGG pathway was enriched by analysis the Kyoto Encyclopedia of Genes and Genomes (KEGG) online database [5]. The univariate COX model (COX model, also be called, proportional hazards model) and multivariate COX model were constructed. For the purpose of finding the key genes, regulatory mechanism and prognosis, the DEmRNAs enriched in KEGG signal pathway were selected as the variates of the multivariate COX model[6]. And the co-expressed DEmRNAs were analyzed by Kaplan-Meier method and ROC curve.

2. Material and Methods

2.1. TCGA Data Download and RNA Matrix Merging
Our research was based on the downloaded GBM RNA data of the GBM project from TCGA online database (https://portal.gdc.cancer.gov) [7]. In this project, 168 tumor samples of 160 GBM patients were collected from 104 males and 56 females, including 109 males’ samples and 59 females’ ones.
Until the research was done in this paper, the 158 downloaded clinical data of the patients was shown that 129 patients were already dead while 29 patients were still alive. The downloaded RNA-seq files of the each single patient were contained protein coding, unitary pseudogene, processed pseudogene, processed transcript, mRNA, snRNA and so on. And it was separated into 2 groups according to the genders and merged into a whole RNA expression matrix.

### Table 1. Data number list

| Gene data | RNA Data number | Status data number |
|-----------|-----------------|--------------------|
| gender    | RNA Cases       | status Files       |
| count     | M | F | M | F | M | F | M | F | N |
| 104       | 56 | 109 | 59 | 85 | 44 | 17 | 12 | 2 |

*M: male; F: female; N: not known

2.2. mRNA Matrix Extracting and DEGs Screening

All the mRNA data in the RNA expression matrix were extracted and merged into a new mRNA expression matrix. By using the limma R package for the mRNA clustering analysis, the DEmRNAs between the two genders’ patients were screened according to the cutoff value of P-value < 0.05 and |LogFoldChange| > 2 [8]. The heatmap of the DEmRNAs was drawn in the clustering analysis[9]. It was shown that the mRNAs in the female patients were high expressed (red block) or low expressed (green block) to the male patients.

2.3. GO Enrichment and Pathway Enrichment

In the enrichment analysis, the screened DEmRNAs were analyzed by online database, DAVID and KEGG. The GO enrichments were focused on 3 GOterms, BP, CC and MF, by using the DAVID online tool (count >= 2, p < 0.1) [4]. According to the logFC value and the gene ID, the enriched signal pathway was found by the clusterprofiler R package and KEGG pathway online database (www.genome.jp). And the enriched DEmRNAs were marked by different colors in the signal pathway graph.

2.4. COX Model and Clinical Data Analysis

The DEmRNAs in the enriched pathway were selected for construction of the COX model. Firstly, the survival data and the DEmRNA expression of the cases were merged into a clinical matrix. Secondly, the DEmRNAs were used as the factors of the univariate COX model analysis. After that, the co-expression DEmRNAs analyzed by the multivariate COX model were made a survival risk matrix with the survival data thirdly.

The co-expression DEmRNAs and the survival risk matrix were analyzed by Kaplan-Meier plotter and ROC [10-12]. And the kmplot survival curve and ROC curve were drawn. The survival analysis of a 2-years or 5-year time was done based on the kmplot curve of high risk group and low risk group. The AUC value of ROC curve could be used to prove whether the COX model and the co-expression DEmRNAs were related to the disease.
Table 2. DEmRNAs list

| Expression | Genes |
|------------|-------|
| High       | OR7C2, RHBG, NRK, SLC10A2, AGTR2, LHX8, IL31RA, RNF17, RNF113B, SCN10A, AMELX, MRGPRX4, CDC83, DLK1, OCA2, MMP13, MAGEC1, TMPRSS2, CYP4F22, CTAG2, FAM150A, CYP24A1, CDC20B, LMX1A, PM20D1, HMGA2, GCSAML, FAM169B, OR1E1, AVIL, CFHR5, IGF2BP1, KANK4, OTX2, VSTM2A, CCBE1, BMP5, GCG, RXFP4, C1orf234, ISL1, COL24A1, COMP, C14orf39 |
| Low        | PCDHGA12, DAO, PGC, DNAH8, CERS3, CTSG, NPS, PRG4, CALC1B, PCDHGB4, SERPINA11, SRRM4, CLC, HES2, FGFBP2, RAX, MUC13, CR2, CALCA, LIPF, FGFI9, PIWIL1, RPS4Y2, PRL, KRT13, MMP12, HCR1T, ECE1L, SLC38A4, CLCA2, TRIM48, SPP2, SDR9C7, GCGR, IFNG, ANKRD30BL, PAX7, IZUMO3, TBL11Y, MAGEA4, KRT14, SPANXB1, EPYC, TMSB4Y, U82695.5, TSPY2, PCDH11Y, CXCL17, C1orf68, PDCL2, NLGN4Y, SRY, DDX3Y, USP9Y, RPS4Y1, EIF1AY, ZFY, KDM5D, UTY |

3. Results

3.1. DEmRNAs Screening

Our work was based on the differential expressed mRNAs between male’s and female’s. And the clustering analysis was used to screen the DEmRNAs. After the clustering analysis of the mRNA matrix with the limma method (cutoff value: P-value< 0.05 and |LogFoldChange| > 2) [13], 103 mRNAs in the male patients were differential expressed to the female patients (Table 2).

Compared to the male ones, 44 mRNAs were high expression in the female patients. And the rest 59 DEmRNAs were low expression. And the high expressed DEmRNAs ( red block ) and the low expressed DEmRNAs ( green block ) were shown as a heatmap in Figure 1. The tree map in heatmap was used to connect two genes or two subclass in the clustering analysis. In Figure 1, the blocks in one column were represented all the screened DEmRNAs from one single tumor sample while the blocks in the same line were represented the same DEmRNA in all the samples.

Figure 1. Heatmap of DEmRNAs

Figure 2. The barplot of top 5 GOterms
3.2. GO Enriched Analysis

When the DEmRNA list was uploaded to the DAVID bioinformatics online database, the enriched GO terms were output. The DEmRNAs were enriched in 26 GO terms, including 5 MF, 4 CC and 17 BP. 23 DEmRNAs were mainly enriched in cellular component, extracellular region (GO:0005576). The mainly enrichment GO terms were shown in Table 3 and Figure 2. The mainly enrichment GO terms (count > 5) were extracellular region (GO:0005576), multicellular organism development (GO:0007275), proteolysis (GO:0006508), oxidation-reduction process (GO:0055114), calcium ion binding (GO:0005509) proteinaceous extracellular matrix (GO:0005578) and extracellular space (GO:0005615).

Table 3. GO enrichment (count > 5)

| Term               | Genes                                                                 |
|--------------------|----------------------------------------------------------------------|
| GO:0005576         | FGF19, HCRT, CLCA2, VSTM2A, PRG4, MMP13, MMP12, GCG, CALCA, CXCL17, CALCB, AGTR2, NPS, FAM150A, COMP, IFNG, CFHR5, COL24A1, PRL, CTSG, SPP2, BMP5, LIPF |
| GO:0005615         | SERPINA11, PGC, DLK1, MMP13, GCG, CALCA, CXCL17, NLGN4Y, COMP, IFNG, CCBE1, EPYC, FGF2P2, MUC13, BMP5, CTSG |
| GO:0007275         | RNF17, C14orf39, AMELX, ZFY, OTX2, PIWIL1, RPS4Y1, HMGA2, CLC |
| GO:0006508         | Tmprss2, CLCA2, ECEL1, PGC, PM20D1, MMP13, MMP12, CTSG |
| GO:0055114         | CYP24A1, UTY, CYP4F22, DAO, SDR9C7, IZUMO3, KDM5D, LIPF |
| GO:0005509         | PCDHGA12, PCDH11Y, COMP, CCBE1, DLK1, MMP13, MMP12, PCDHGB4 |
| GO:0005578         | AMELX, COMP, CCBE1, COL24A1, EPYC, MMP13, MMP12 |

3.3. KEGG Pathway Enriched Analysis

The DEmRNAs were analyzed by the clusterprofiler R package and KEGG online database. One enriched signal pathway was found in our work. 9 DEmRNAs were enriched in hsa04080 (Neuroactive ligand-receptor interaction), AGTR2, CALCA, CALCB, CTSG, GCG, GCGR, HCRT, PRL and RXFP4. And the 9 DEmRNAs were marked by red or green color in the hsa04080 signal pathway (Figure 3). AGTR2, CALCA, CALCB, CTSG, GCG, GCGR, HCRT, PRL and RXFP4 were selected for the further analysis.

3.4. Survival Analysis

After the univariate COX model analysis, HR value and z value of all the DEmRNAs were computed. Based on the univariate COX model analysis and the mRNA clustering analysis, the gene list (Table 4) enriched in hsa04080 were constructed a multivariate COX model. The DEmRNAs, GCGR, HCRT and CTSG, were found as the co-expression mRNAs for the survival analysis (Table 5). And the other 6 DEmRNAs were ignored. The survival risk matrix was built for the kmplot survival analysis and ROC curve analysis.

Table 4. The DEmRNAs in hsa04080

| gene   | logFC  | HR   | z     |
|--------|--------|------|-------|
| AGTR2  | 3.631396 | 0.962914 | -0.32612 |
| CALCA  | -2.35145 | 1.05334 | 0.868111 |
| CALCB  | -2.24276 | 1.030752 | 0.561973 |
| CTSG   | -2.15647 | 1.189527 | 2.421903 |
| GCG    | 2.119445 | 1.003228 | 0.038103 |
| GCGR   | -3.03307 | 0.925335 | -1.27766 |
| HCRT   | -2.49641 | 1.148778 | 2.353796 |
| PRL    | -2.45603 | 0.973006 | -0.47919 |
| RXFP4  | 2.091579 | 0.950435 | -0.62204 |
Table 5. The co-expression DEmRNAs

| Gene | coef | se(coef) | z     | Pr(>|z|) |
|------|------|----------|-------|---------|
| GCGR | -0.1081 | 0.060365 | -1.79079 | 0.073327 |
| HCRT | 0.190898 | 0.059659 | 3.199838 | 0.001375 |
| CTSG | 0.197 | 0.073699 | 2.673025 | 0.007517 |

In the kmplot survival curve (Figure 4, \( p = 1.63 \times 10^{-3} \)), the clinical data was separated into high risk group and low risk group. It was shown that none of the patient clinical data was found in the high risk group in about 3-year time. And in a 2-year time, survival rate of the high risk group was about 20% lower than low risk group. In the ROC curve (Figure 5), the AUC value was 0.766. And the true positive rate was higher than the false positive rate.

4. Discussion

It could be known from the early research that the morbidity and mortality of GBM was different between male and female. In this paper, we downloaded 168 RNA-seq files of tumor samples from 160 GBM patients that collected by TCGA database. And we did the comparison of the mRNA expression between the 109 males’ samples and 59 females’ sample in order to found out the clue of the different morbidity between the different genders.

Figure 3. Signal pathway has04080
103 DEmRNAs were screened by the clustering analysis with limma R package and cutoff value: P-value < 0.05 and |LogFoldChange| > 2. Among them, 44 ones of DEmRNAs were high expression in the female patients while the other 59 DEmRNAs were low expression. It could be found by the DAVID database that the screened 103 DEmRNAs were enriched in 17 biological process, 4 cellular component and 5 molecular function. And the most GOterm, extracellular region (GO: 0005576), was enriched 23 DEmRNAs (Table 3). Extracellular region covers the host cell environment outside an intracellular parasite. It’s the external space of a cell.

And in the KEGG pathway enriched analysis, AGTR2, CALCA, CALCB, CTSG, GCG, GCGR, HCRT, PRL and RXFP4 were enriched in Neuroactive ligand-receptor interaction (hsa04080). It could be found in 3 of the top 5 enriched GOterms. HCRT, GCG, CALCA, CALCB, AGTR2, PRL and CTSG were enriched in extracellular region (GO: 0005576). GCG, CALCA and CTSG were enriched in extracellular space (GO: 0005615). CTSG was enriched in proteolysis (GO: 0006508).

Based on the 9 DEmRNAs, a multivariate COX model was constructed. And the co-expression mRNAs, GCGR, HCRT and CTSG, were selected by the COX model. And the 3 co-expression mRNAs were analyzed by Kaplan-Meier method and ROC curve. It could be learn from the survival curves in figure 4 that the survival rate of high risk group and low risk group was obvious difference in the 2-year time and 5-year time analysis. It was shown that the co-expression mRNAs model might be a potential biomarker for GBM. Although some of the points were in false positive rate area of the ROC curve, the ROC curve and AUC value were still provided the strong proof for the judgement in the statistical analysis.

By the analysis of the TCGA GBM project in our paper, it could be known that hsa04080 (Neuroactive ligand-receptor interaction) was a key signal pathway of the GBM related to the different genders. And the regulatory mechanism of hsa04080 and co-expression DEmRNAs in hsa04080 might be intervene the onset of GBM in the different genders’ patients.

5. Acknowledgements
This work was supported by National Natural Science Foundation of China (No. 81660031, 81360090), Guangxi Natural Science Foundation of China (2014GXNSFBA118151), and Guangxi key Laboratory Foundation of Embedded Technology and Intelligent System.

6. References
[1] M. Tian, W. Ma, Y. Chen, Y. Yu, D. Zhu, J. Shi, Y. Zhang, Impact of gender on the survival of patients with glioblastoma, *Biosci Rep*, vol. 38, no. 6, Dec 21, 2018.
[2] D. A. Silvestris, E. Picardi, V. Cesarini, B. Fosso, N. Mangraviti, L. Massimi, M. Martini, G. Pesole, F. Locatelli, Dynamic inosinome profiles reveal novel patient stratification and gender-specific differences in glioblastoma, *Genome Biol*, vol. 20, no. 1, pp. 33, Feb 13, 2019.
[3] M. E. Ritchie, B. Phipson, D. Wu, Y. Hu, C. W. Law, W. Shi, G. K. Smyth, limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleic Acids Res*, vol. 43, no. 7, pp. e47, Apr 20, 2015.

[4] G. Dennis, Jr., B. T. Sherman, D. A. Hosack, J. Yang, W. Gao, DAVID: Database for Annotation, Visualization, and Integrated Discovery, *Genome Biol*, vol. 4, no. 5, pp. P3, 2003.

[5] M. Kotera, M. Hirakawa, T. Tokimatsu, S. Goto, M. Kanehisa, The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals, *Methods Mol Biol*, vol. 802, pp. 19-39, 2012.

[6] Y. Z. jianzhi Deng, Big data analysis of differentially expressed mRNA in acute lymphoblastic leukemia patients with different genders, *ITME2019*, Qingdao, China 2019.

[7] M. Deng, J. Bragelmann, J. L. Schultze, Web-TCGA: an online platform for integrated analysis of molecular cancer data sets, *BMC Bioinformatics*, vol. 17, pp. 72, Feb 6, 2016.

[8] Y. Z. jianzhi Deng, Xiaohui Cheng, Bioinformatics analysis of differentially expressed genes and identification of AGER as a key gene in lung cancer, *BIBE2019*, 2019.

[9] B. B. Khomtchouk, J. R. Hennessy, MicroScope: ChIP-seq and RNA-seq software analysis suite for gene expression heatmaps, *BMC Bioinformatics*, vol. 17, no. 1, pp. 390, Sep 22, 2016.

[10] Y. T. van der Schouw, H. Straatman, A. L. Verbeek, ROC curves and the areas under them for dichotomized tests: empirical findings for logistically and normally distributed diagnostic test results, *Med Decis Making*, vol. 14, no. 4, pp. 374-81, Oct-Dec, 1994.

[11] C. van Walraven, S. Hawken, Competing risk bias in Kaplan-Meier risk estimates can be corrected, *J Clin Epidemiol*, vol. 70, pp. 101-5, Feb, 2016.

[12] G. Tripepi, F. Catalano, Kaplan-Meier analysis, *G Ital Nefrol*, vol. 21, no. 6, pp. 540-6, 2004.

[13] I. Diboun, L. Wernisch, C. A. Orengo, M. Koltzenburg, Microarray analysis after RNA amplification can detect pronounced differences in gene expression using limma, *BMC Genomics*, vol. 7, pp. 252, Oct 9, 2006.