Screening of differentially expressed lncRNAs for predicting Low Grade Glioma between different genders by Computational Biology

Jianzhi Deng 1,2, Meng Yao 2, Xiaohui Cheng 1, Yuehan Zhou 3, a
1 Guilin University of Technology, Guilin 541004, China
2 East China Normal University, Shanghai 200241, China
3 Guilin Medical University, Guilin 541004, Guangxi, China
a Corresponding author: yuehanzhou2012@163.com

Abstract. In this study, we aim to find out the high-risk gene prediction model of Low Grade Glioma (LGG) male patient. After analysis of over 500 LGG patient RNA data and clinical data from the Cancer Genome Atlas (TCGA) database, the regulatory network of the differentially expressed lncRNA (DElncRNA) and differentially expressed miRNA (DEmiRNA) was established. 8 DElncRNAs, IGF2-AS, TTTY14, LINC00305, XIST, LINC00276, ZFY-AS1, TTTY15 and C8orf49, and related survival data of the patients were analyzed by the multivariable COX regression model and Akaike Information Criterion (AIC). Furthermore, XIST, ZFY-AS1, IGF2-AS and C8orf49 were kept for ROC and AUC analysis. Based on the 4 DElncRNAs, closely related assessing risk score and risk level of the LGG patients, we drawn a survival kmplot curve (p=5.153e-06) and a ROC curve (AUC=0.599), and the true positive rate (TPR) and false positive rate (FPR) were shown in the ROC curve. We believe that the XIST, ZFY-AS1, IGF2-AS and C8orf49 model can give us an effective way for the diagnosis of LGG male patients.

1. INTRODUCTION
Low grade glioma (LGG) is a type of frequent and high invasive brain tumors. Some early research found that LGG was at higher risk in male than female. In this paper, we try to find out different biomarkers between the two genders and assess whether the different biomarker could be used to predict the high risk in male [1, 2]. Firstly, the differentially expressed genes were screening out from the downloaded RNA-seq collected by The Cancer Genome Atlas (TCGA) database [3, 4]. Secondly, the regulatory network of differentially expressed lncRNAs and differentially expressed miRNAs was established [5, 6]. Based on the DElncRNAs in the regulatory network and their closely clinical data, the COX model was used to integrate the significant clinical variables and key regulated DElncRNAs [7, 8]. The receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC) were done according to the result of the COX model analysis [9, 10]. It has proved that a model consists of 4 DElncRNAs, XIST, ZFY-AS1, IGF2-AS and C8orf49 is effective for LGG diagnosis prediction. And it has also been proved by the survival kmplot curve of the 4 DElncRNAs.

2. MATERIALS AND METHODS
2.1 Data download and rebuild
The RNA-seq, miRNA-seq and clinical data of LGG patients were downloaded from TCGA (portal.gdc.cancer.gov) database. In TCGA database, the RNA-seq or miRNA-seq of each patient were saved in one txt file. We got 528 RNA-seq files and 539 miRNA-seq files of the LGG patients. Among then, 289 RNA-seq files were male patients’ while the other 239 files were female’s. 288 ones of the 539 miRNA-seq files were male patients’, and the else were female LGG patients’. The lncRNAs and their expressions was screened from the RNA-seq files and merged into a lncRNA matrix. The miRNA matrix was merged by extracting the gene ID and its expression in every patient’s file.

The corresponding clinical survival data of all the LGG patients was downloaded in a zip file, including submitter_id, days_to_death, primary_diagnosis, vital_status and so on [3, 11]. The typical survival data, days_to_death and state, were extracted from the zip file and rebuilt a new survival data txt file.

2.2 DEGs screening and Regulatory analysis of lncRNA-miRNA
The research below was mainly based on the differentially expressed lncRNA (DElncRNA), and differentially expressed miRNA (DEmiRNA) between the male and female patients’ data which was screened from the matrix of lncRNA and miRNA by edgeR package with the threshold value: P-value < 0.05 and \(|\text{LogFoldChange}| > 2\) [12, 13].

The regulatory of lncRNA-miRNA was searched by mircode database (www.mircode.org) and the regulatory network was drawn by cytoscape software [14].

2.3 COX analysis, ROC curve and AUC analysis
The ROC curve AUC analysis was done by both up- and down-regulated DElncRNA that were built in the lncRNA-miRNA regulatory network. By using the multivariable COX regression model and AIC, only one gene among the co-expression DElncRNA was kept for further research.

Excluding the unavailable or redundant data, COX model was used to integrate the significant clinical variables and key regulated DElncRNAs [8]. And the risk-score of each patient was calculated by survival R package. According to the risk-score, the risk was divided into 2 levels, high (risk-score>1) and low (risk-score<0). The survival data and expression of the DElncRNAs were merged in one file. And the 5 years ROC curve AUC analysis was based on this file. It can be known from the ROC curve that the bigger the AUC is, the more accurate the predictive result is [9].

2.4 Kmplot survival analysis
The survival kmplot curve was also analyzed by the same lncRNA in the regulatory network using kaplan-meier way [15]. In the kmplot curve, the LGG patient’s samples were separated into high risk curve and low risk curve with the survival median value.

3. RESULTS
41 DElncRNAs and 16 DEMiRNAs between the two different genders were selected out from the TCGA downloaded LGG samples’ RNA-seq and miRNA-seq. In the male samples, 26 DElncRNAs and 11 DEMiRNAs were higher expression than the female samples while 15 DElncRNAs and 5 DEMiRNAs were lower expression.
These were 18 regulation pairs between 8 DElncRNAs (IGF2-AS, TTTY14, LINC00305, XIST, LINC00276, ZFY-AS1, TTTY15 and C8orf49) and 4 DEmiRNAs (hsa-mir-122, hsa-mir-372, hsa-mir-204 and hsa-mir-206) by screening the mircode database. Regulatory network of the 18 regulation pairs was shown in Figure 1.

All the DElncRNAs was analyzed by a multi-variable COX model. The 3 values of the 8 DElncRNAs, HR and p-value were listed in table 1.

Table 1. multi-variable COX model

| gene     | HR   | pvalue |
|----------|------|--------|
| IGF2-AS  | 1.105239 | 0.020896 |
| LINC00276 | 0.944154 | 0.627493 |
| LINC00305 | 1.292216 | 0.217352 |
| TTTY14   | 1.012889 | 0.604603 |
| TTTY15   | 1.01152  | 0.566609 |
| XIST     | 1.002708 | 0.87443  |
| ZFY-AS1  | 1.092709 | 0.093997 |
| C8orf49  | 1.230466 | 0.000407 |

According to the COX model and the AIC value, TTTY15, TTTY14, LINC00305 and LINC00276 were excluded because they were co-expressed with the rest 4 DElncRNA, XIST, ZFY-AS1, IGF2-AS and C8orf49. And the result of the COX analysis was listed in table 2. The 4 DElncRNA related risk-score and risk level of the LGG patients was also calculated in COX analysis.

Table 2. the COX model value of the ROC analysis

| gene         | coef   | z      | Pr(>|z|) |
|--------------|--------|--------|---------|
| XIST         | 0.064669 | 2.172229 | 0.029838 |
| ZFY-AS1      | 0.238582 | 2.787637 | 0.005309 |
| IGF2-AS      | 0.085628 | 1.881481 | 0.059906 |
| C8orf49      | 0.196818 | 3.302374 | 0.000959 |

The ROC curve and AUC analysis was based on XIST, ZFY-AS1, IGF2-AS and C8orf49, and drawn in a ROC curve (figure 1), including TPR and FPR. The AUC value is 0.599. The clinical survival data of the 4 DElncRNAs were drawn a kmplot curve by analyzing in kaplan-meier method. Compare with female, KM survival curve (p=5.153e-06) of the 4 dysregulated DElncRNAs is high-risk in male and shown in Figure 2.

4. DISCUSSION

LGG is a complicate tumor that has a higher incidence in male than in female. And it has already known from the recently work. In this work, we focus on the difference between male and female in genetic level. By screening the DElncRNA and DEmiRNA from over 500 LGG samples in TCGA, the
regulatory network was constructed in 18 regulation pairs by 8 DElncRNAs (IGF2-AS, TTTY14, LINC00305, XIST, LINC00276, ZFY-AS1, TTTY15 and C8orf49) and 4 DEmiRNAs (hsa-mir-122, hsa-mir-372, hsa-mir-204 and hsa-mir-206). By using the multivariable COX model, the key regulated DElncRNAs were found out, based on the clinical data and the 8 DElncRNAs. XIST, ZFY-AS1, IGF2-AS and C8orf49 were kept for ROC and AUC analysis after the redundant data was excluded. The 4 DElncRNA related risk-score was calculated in the 5 years ROC curve AUC analysis, and shown a ROC curve (AUC=0.599) and survival kmplot curve (p=5.153e-06).

In conclusion, the above research shown that the XIST, ZFY-AS1, IGF2-AS and C8orf49 model is effective for the prediction of LGG male patients.
ACKNOWLEDGMENTS
This work was supported by the 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.

References
[1] R. Durmaz et al., Efficacy of prognostic factors on survival in patients with low grade glioma, Turk Neurosurg, vol. 18, no. 4, pp. 336-44, Oct 2008.
[2] F. Fang et al., [Factors associated with the surgical efficacy and prognosis of seizures in patients with low-grade glioma], Zhonghua Yi Xue Za Zhi, vol. 96, no. 13, pp. 1031-4, Apr 5 2016.
[3] K. Tomczak, P. Czerwinska, and M. Wiznerowicz, The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge, Contemp Oncol (Pozn), vol. 19, no. 1A, pp. A68-77, 2015.
[4] G. S. Krasnov et al., CrossHub: a tool for multi-way analysis of The Cancer Genome Atlas (TCGA) in the context of gene expression regulation mechanisms, Nucleic Acids Res, vol. 44, no. 7, p. e62, Apr 20 2016.
[5] Y. Pan, G. Liu, D. Wang, and Y. Li, Analysis of IncRNA-Mediated ceRNA Crosstalk and Identification of Prognostic Signature in Head and Neck Squamous Cell Carcinoma, Front Pharmacol, vol. 10, p. 150, 2019.
[6] L. Salmena, L. Poliseno, Y. Tay, L. Kats, and P. P. Pandolfi, A ceRNA hypothesis: the Rosetta Stone of a hidden RNA language?, Cell, vol. 146, no. 3, pp. 353-8, Aug 5 2011.
[7] C. Lusivika-Nzinga, H. Selinger-Leneman, S. Grabar, D. Costagliola, and F. Carrat, Performance of the marginal structural cox model for estimating individual and joined effects of treatments given in combination, BMC Med Res Methodol, vol. 17, no. 1, p. 160, Dec 4 2017.
[8] Z. Ai, L. Li, R. Fu, J. M. Lu, J. D. He, and S. Li, Integrated Cox's model for predicting survival time of glioblastoma multiforme, Tumour Biol, vol. 39, no. 4, p. 1010428317694574, Apr 2017.
[9] P. Liao, H. Wu, and T. Yu, ROC Curve Analysis in the Presence of Imperfect Reference Standards, Stat Biosci, vol. 9, no. 1, pp. 91-104, Jun 2017.
[10] A. N. Kamarudin, T. Cox, and R. Kolamunnage-Dona, Time-dependent ROC curve analysis in medical research: current methods and applications, BMC Med Res Methodol, vol. 17, no. 1, p. 53, Apr 7 2017.
[11] Y. Zhu, P. Qiu, and Y. Ji, TCGA-assembler: open-source software for retrieving and processing TCGA data, Nat Methods, vol. 11, no. 6, pp. 599-600, Jun 2014.
[12] H. Varet, L. Brillet-Gueguen, J. Y. Coppee, and M. A. Dillies, SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data, PLoS One, vol. 11, no. 6, p. e0157022, 2016.
[13] C. W. Law et al., RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR, F1000Res, vol. 5, 2016.
[14] A. Jeggari, D. S. Marks, and E. Larsson, miRcode: a map of putative microRNA target sites in the long non-coding transcriptome, Bioinformatics, vol. 28, no. 15, pp. 2062-3, Aug 1 2012.
[15] F. Pan, S. Cui, W. Wang, A. Gu, and L. Jiang, Survival analysis for lung adenosquamous carcinoma patients with brain metastasis, J Cancer, vol. 9, no. 20, pp. 3707-3712, 2018.