Journal Pre-proofs

Database

Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Gliomas

Zheng Zhao, Ke-Nan Zhang, Qiangwei Wang, Guanzhang Li, Fan Zeng, Ying Zhang, Fan Wu, Ruichao Chai, Zheng Wang, Chuanbao Zhang, Wei Zhang, Zhaoshi Bao, Tao Jiang

PII: S1672-0229(21)00045-0
DOI: https://doi.org/10.1016/j.gpb.2020.10.005
Reference: GPB 500

To appear in: *Genomics, Proteomics & Bioinformatics*

Received Date: 1 July 2019
Revised Date: 1 October 2020
Accepted Date: 26 December 2020

Please cite this article as: Z. Zhao, K-N. Zhang, Q. Wang, G. Li, F. Zeng, Y. Zhang, F. Wu, R. Chai, Z. Wang, C. Zhang, W. Zhang, Z. Bao, T. Jiang, Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Gliomas, *Genomics, Proteomics & Bioinformatics* (2021), doi: https://doi.org/10.1016/j.gpb.2020.10.005

This is a PDF file of an article that has undergone enhancements after acceptance, such as the addition of a cover page and metadata, and formatting for readability, but it is not yet the definitive version of record. This version will undergo additional copyediting, typesetting and review before it is published in its final form, but we are providing this version to give early visibility of the article. Please note that, during the production process, errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

© 2021 The Authors. Production and hosting by Elsevier B.V. on behalf of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Gliomas

Zheng Zhao[1,#], Ke-Nan Zhang[1,#], Qiangwei Wang[1,2,#], Guanzhang Li[1], Fan Zeng[1], Ying Zhang[1], Fan Wu[1], Ruichao Chai[1], Zheng Wang[3], Chuanbao Zhang[3], Wei Zhang[3], Zhaoshi Bao[1,3,*], Tao Jiang[1,3,4,5,*]

[1]*Beijing Neurosurgical Institute, Capital Medical University, Beijing 100070, China*

[2]*Department of Neurosurgery, The Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou 310009, China*

[3]*Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing 100070, China*

[4]*Center of Brain Tumor, Beijing Institute for Brain Disorders, Beijing 100069, China*

[5]*China National Clinical Research Center for Neurological Diseases, Beijing 100070, China*

[#]Equal contribution

[*]Corresponding authors.

E-mail: taojiang1964@163.com (Jiang T), bzsjoe1985@163.com (Bao ZS)

Running title: *Zhao et al / Chinese Glioma Genome Atlas*

Total references: 32

Total figures: 3

Total tables: 1

Total supplementary figures: 0

Total supplementary tables: 0

Journal Pre-proofs

Total supplementary files: 0

Journal Pre-proofs

Abstract

Glioma is the most common and malignant intracranial tumors in adults. Recent studies have revealed the significance of **functional genomics** for the development of glioma pathophysiological researches and treatments. However, the access to comprehensive genomic data and analytical platforms is often limited. Here, we developed the **Chinese Glioma Genome Atlas** (CGGA), a user-friendly data portal for the storage and interactive exploration of cross-omics data, including nearly 2000 primary and recurrent glioma samples from **Chinese cohorts**. Currently, open access is provided to whole-exome sequencing data (286 samples), mRNA sequencing (1018 samples) and microarray data (301 samples), DNA methylation microarray data (159 samples), and microRNA microarray data (198 samples), and to detailed clinical information (age, gender, chemoradiotherapy status, WHO grade, histological type, critical molecular pathological information, and survival data). In addition, we have developed several analytical tools for users to analyze the mutation profiles, mRNA/microRNA expression, and DNA methylation profiles and to perform survival and gene correlation analyses of specific glioma subtypes. This **database** removes the barriers for researchers, providing rapid and convenient access to high-quality functional genomic data resources for biological research and clinical applications. CGGA is available at http://www.cgga.org.cn.

KEYWORDS: Glioma; Functional genomics; Chinese Glioma Genome Atlas; Chinese cohorts; Database

Journal Pre-proofs

Introduction

Glioma is the most common intracranial malignant tumors in adults. According to a multi-center cross-sectional study on brain tumors in China, the age-standardized prevalence of primary brain tumors is approximately 22.52 per 100,000 for all populations, with glioma accounting for 31.1% [1−3]. Despite advances in current treatment standards, the survival rate of patients with glioma has not obviously improved in decades, especially for aggressive gliomas (associated with a poor median survival time of 14.4 months) [4,5]. According to the 2016 World Health Organization (WHO) classification of central nervous system (CNS) tumors, glioma is classified from grade II to grade IV by not only histological characteristics but also several molecular pathological features, *e.g.*, *IDH* mutation and chromosome 1p/19q co-deletion status [6]. Clinically, most lower-grade gliomas (LGG) progress to glioblastoma (grade IV, GBM) in less than 10 years [6−8]. Glioma recurrence or malignant progression is likely for several reasons: (1) infiltrative tumor cells cannot be completely removed by neurosurgical resection [9,10]; (2) residual tumor cells cannot be effectively suppressed by limited postoperative treatment options [3,11,12]; (3) multiple lesions may progress sequentially [13,14]; (4) tumor cell cloning occurs rapidly under chemotherapy and/or radiotherapy [7,15]; (5) tumor cells readily adapt to the immunosuppressive tumor microenvironment [16,17]; (6) research is hindered by limited data resources. Therefore, it is essential to collect clinical specimens and provide genomic sequencing data to the research community.

Recently, high-throughput technologies have been extended to characterize genomic status including but not limited to DNA methylation modification, genetic alteration, and gene expression regulation. In the cancer research community, major large-scale projects, such as The Cancer Genome Atlas (TCGA, which includes 516 LGGs and 617 GBMs as of October 18, 2019) [18] and the International Cancer Genome Consortium (ICGC, which includes 80 adult GBMs and 50 pediatric GBMs (excluding the TCGA samples) as of April 3, 2019) [19,20], have generated an unparalleled amount of functional genomic data. These projects have changed our understandings of cancers and led to breakthroughs in diagnosis, treatments, and prevention. Importantly, they have provided opportunities for discovery and validation to researchers worldwide. However, the data generated by these projects are often difficult to access, analyze, and visualize, especially for

researchers with little bioinformatics skills. These limitations have greatly hindered the use of functional genomics data to obtain novel findings of significance for drug development and clinical treatments. Although several webservers, *e.g.*, cBioportal [21,22] and GlioVis [23], have been constructed to analyze multi-dimensional glioma data, they ignore the heterogeneity in tumors, as data obtained from recurrent glioma samples and subtype analyses are lacked.

Here, we introduce the Chinese Glioma Genome Atlas (CGGA, http://www.cgga.org.cn), an open-access and easy-to-use platform for the interactive exploration of multi-dimensional functional genomic datasets collected from nearly 2000 glioma samples from Chinese cohorts. The database currently contains a wide range of data derived from whole-exome sequencing (WES, 286 samples), mRNA sequencing (1018 samples) and microarray (301 samples), DNA methylation microarray (159 samples), and microRNA microarray analyses (198 samples) as well as comprehensive clinical data. Furthermore, we developed various online tools to browse mutational landscape profiles, mRNA/microRNA expression profiles, and DNA methylation profiles, and to perform survival and correlation analyses of specific subtypes. We believe that the website removes the barriers for researchers who need fast and convenient access to high-quality functional genomic data resources.

Database implementation

In CGGA, all data were organized using MySQL 14.14 based on relational schema, which will be supported in future CGGA updates. The website code was written based on Java Server Pages using the Java Servlet framework. The website is deployed on the Tomcat 6.0.44 web server and runs on a CentOS 5.5 Linux system. JQuery was used to generate, render and manipulate data for visualization. The 'Analyze' module was realized by Perl and R script. The CGGA website has been fully tested in Google Chrome and Safari browsers. The design of CGGA is displayed in **Figure 1**.

Database content and usage

Database content

The CGGA database is designed to archive functional genomic data and to allow the interactive exploration of multidimensional datasets from both primary and recurrent gliomas in Chinese

cohorts. The database is available at http://cgga.org.cn. Currently, CGGA contains WES (286 samples), mRNA sequencing (a total of 1018 samples, with batch 1 comprising 693 samples and batch 2 comprising 325 samples), microarray (301 samples), DNA methylation microarray (159 samples), and microRNA microarray (198 samples) data and detailed clinical data (including age, gender, chemoradiotherapy status, WHO grade, histological type, critical molecular pathological information and survival data). Detailed statistical information of each dataset is provided in **Table 1**. Out-house sequencing data from TCGA (702 samples) and REMBRANDT (475 samples) can be acquired on the download page. We have organized web interfaces of CGGA according to the four main functional features: (i) Home, (ii) Analyze, (iii) Tools, and (iv) Download. In what follows, we provide an example of how to use CGGA.

The homepage

On the 'Home' page, CGGA provides a statistical table of all collected datasets, including dataset name, data type, number of samples in each subgroup, clinical data and analysis purpose. For instance, we have performed mRNA sequencing on 1,018 glioma samples and obtained 693 samples in batch 1 and 325 samples in batch 2 (including 282 primary LGGs, 161 recurrent LGGs, 140 primary GBMs and 109 recurrent GBMs in batch 1 and 144 primary LGGs, 38 recurrent LGGs, 85 primary GBM, 24 recurrent GBMs and 30 secondary GBMs in batch 2). Significantly, CGGA is the first database to archive functional genomic data for both recurrent LGGs and GBMs. In addition, users can view the results of the analysis of each dataset by clicking on hyperlinks on the 'Home' page. The 'Download' and 'Help' pages can be accessed directly from the 'Home' page.

The analyses and results

To facilitate analysis of the CGGA data, especially for bioinformatics beginners, we developed four online modules in the 'Analzse' tab (**Figure 2**). 'WEseq data', 'mRNA data', 'methylation data', and 'microRNA data' are included for analyzing the WES, mRNA expression, DNA methylation and microRNA expression data, respectively (Figure 2A). A key feature of CGGA is its ease of use. In the example below, we illustrate the usage of the 'Analzse' tab in CGGA.

On the 'WEseq data' page, users can visualize the mutational profile of a gene set of interest and perform a survival analysis of a specific gene of interest in specific glioma subtype (Figure 2B). In

Journal Pre-proofs

the 'Oncoprint' section, users are guided to a) input a gene set of interest, for example, *IDH1, TP53* and *ATRX*; and b) select a dataset of interest, for example, 'All Grade Gliomas'. Based on user input, the tool automatically generates results and displays to the users. In these results, each case or patient data are presented in columns, each row corresponds to a gene, and different kinds of mutations are marked in colors and a heatmap is presented below the table depicting clinical information (Figure 2C). The 'Oncoprint' section can be very useful for visualizing the mutational profile of a gene set of interest in a specific glioma subtype and intuitively validate mutational frequency and mutual exclusivity or co-occurrence for a gene pair. In the above example, the mutations in gene *IDH1* (47%), *TP53* (46%) and *ATRX* (30%) were the most common mutations in all grade gliomas. In the 'Survival' section, users can input a specific gene (*e.g.*, *IDH1*) and select a dataset (*e.g.*, 'Primary LGG') to investigate the association of gene mutation with survival. Consistent with previous studies [24], primary LGG cases with *IDH1* mutation enjoy better overall survival than wild-type *IDH1* patients ($P < 0.0001$, Figure 2D, Left). The results from the 'WEseq data' section can be exported in PDF format. To ensure repeatability, the input data (Figure 2D, Middle) and R code (Figure 2D, Right) are provided, enabling users to reproduce the figure with customized options according to their own demands.

On the 'mRNA data' page, users can perform the distribution of gene expression, correlation and survival analyses for a specific gene in a specific glioma subtype (Figure 3A). Three mRNA datasets are available to users, including two batches of RNA-seq datasets (batch 1: 693 samples; batch 2: 325 samples) and one microarray dataset (301 samples). In the 'Distribution' section, users can display one gene distribution pattern for each glioma subtype by selecting a dataset (*e.g.*, 'mRNAseq_325') and inputting a gene name of interest (*e.g.*, *ADAMTSL4*).

Upon hovering the mouse over each point, the expression level and clinical information of each case appear in a pop-up window. The results show the gene expression pattern in each glioma subtype classified by clinical information. In our illustrative case, similar to our previous studies [25], the *ADAMTSL4* gene was shown to be differentially expressed according to the WHO 2016 classification based on the *IDH* mutation and/or 1p/19q co-deletion status and WHO grade (Figure 3B). In addition, a unique feature of the CGGA dataset is the inclusion of recurrent gliomas. This module allows users to infer whether a gene may be a candidate factor that drives malignant progression if it is differentially expressed in primary and recurrent gliomas. In the 'Correlation'

section, the user can validate the co-expression pattern by selecting a dataset (*e.g.*, 'mRNAseq_325') and entering a gene pair (*e.g.*, *ADAMTSL4* and *CD274*). As a result, the co-expression patterns in each glioma subtype are displayed with the results of Pearson's correlation and the *P* value (Figure 3C). In the 'Survival' section, users can perform survival analysis based on gene expression by selecting a dataset (*e.g.*, 'mRNAseq_325') and inputting a gene (*e.g.*, *ADAMTSL4*). In our illustrative case, all primary glioma patients with low *ADAMTSL4* expression enjoyed better overall survival than those with high ADAMTSL4 expression ($P < 0.0001$, Figure 3D Left; $P = 0.00023$, Figure 3D Middle; $P = 0.0036$, Figure 3D Right). The above results from the 'mRNA data' section are consistent with the results of our previous study [25]. Similar to the 'mRNA data' page, on 'methylation data' page and the 'microRNA data' page, users can view the methylation/miRNA distribution and perform correlation and survival analyses.

Further analyses can be accomplished in the 'Tools' section, such as differential expression analysis, cluster analysis and correlation analysis. An expression matrix can be downloaded and rearranged by the user, and the user can upload an input matrix following the instructions. The resulting graph can be downloaded in PDF format.

Data acquisition

All the datasets can be downloaded on the 'Download' page by researchers. Each data type is saved as the gene and/or probe level and is then combined with available clinical data, including basic clinical information, survival and therapy information. The sequencing raw data can be accessed from the National Genomics Data Center (NGDC) by online application.

Method

Clinical specimen collection

Glioma tissues and corresponding genomic data and patient follow-up information were obtained from Beijing Tiantan Hospital at Capital Medical University, Tianjin Medical University General Hospital, Sanbo Brain Hospital at Capital Medical University, the Second Affiliated Hospital of Harbin Medical University, the First Affiliated Hospital of Nanjing Medical University, and the First Hospital of China Medical University. According to the central pathology reviews of

independent neuropathologists, all the subjects were consistently diagnosed with glioma and were then further classified according to the 2007/2016 WHO classification system. The specimens were collected under IRB KY2013-017-01 and frozen in liquid nitrogen within 5 min of resection.

Data processing for whole-exome sequencing data

Genomic DNA from each tumor and matched blood sample was extracted and assessed for integrity by 1% agarose gel electrophoresis. The DNA was subsequently fragmented and subjected to quality control, and then pair-end libraries were prepared. The Agilent SureSelect kit v5.4 (Santa Clara, CA) was used for target capture. Sequencing was performed on an Illumina HiSeq 4000 platform (San Diego, CA) using pair-end sequencing strategy. Valid DNA sequencing data were mapped to the reference human genome (UCSC hg19) using Burrows-Wheeler Aligner (v0.7.12-r1039, bwa mem) [26] with default parameters. Then, SAMtools (V1.2) [27] and Picard (V2.0.1, Broad Institute, Cambridge, MA) were used to sort the reads by coordinates and mark duplicates. Statistics such as sequencing depth and coverage were calculated based on the resultant BAM files. SAVI2 was used to identify somatic mutations (including single nucleotide variations and short insertion/deletions) as previously described [7,8]. Briefly, in this pipeline, SAMtools mpileup and bcftools (V0.1.19) [28] were used to perform variant calling; then, the preliminary variant list was filtered to remove positions with insufficient sequencing depth, positions with only low-quality reads, and positions that were biased toward either strand. Somatic mutations were identified and evaluated by an Empirical Bayesian method. In particular, mutations with a mutation allele frequency in tumors significantly higher than that in normal controls were selected.

Data processing for mRNA sequencing data

Prior to library preparation, total RNA was isolated using the Qiagen RNeasy Mini Kit (Cat No. 74104, Dusseldorf, Germany) according to the manufacturer's instructions. Pestle and Qiagen QIAshredder (Cat No. 79654, Dusseldorf, Germany) were used to disrupt and homogenize frozen tissue. RNA intensity was evaluated using Agilent 2100 Bioanalyzer (Santa Clara, CA), and only high-quality samples with an RNA Integrity Number (RIN) value greater than or equal to 6.8 were used to construct the sequencing library. Typically, 1 μg of total RNA was used with the Illumina TruSeq RNA library preparation kit (RS-122-2001, San Diego, CA) in accordance with low-

throughput protocols, except for the use of SuperScript III reverse transcriptase (Invitrogen 18080044, Carlsbad, CA) to synthesize first strand cDNA. After PCR enrichment and purification of adapter-ligated fragments, the concentration of DNA with adapters was determined with 7500 Fast Real-Time PCR Systems (Applied Biosystems, Carlsbad, CA) using primers QP1 5′-AATGATACGGCGACCACCGA-3′ and QP2 5′-CAAGCAGAAGACGGCATACGAGA-3′. The length of the DNA fragment was measured using a 2100 Bioanalyzer with a median insert size of 200 nucleotides. Then, RNA-seq libraries were sequenced using the Illumina HiSeq 2000, 2500 or 4000 Sequencing System (San Diego, CA). The libraries were prepared using the paired-end strategy witha read length of 101 bp, 125 bp or 150 bp. Base-calling was performed by the Illumina CASAVA V1.8.2 pipeline. RNA-seq mapping and quantification were performed by STAR (V2.5.2b) [29] and RSEM (V1.2.31) software [30]. Briefly, the reads were aligned to the human genome reference (GENCODE v19, hg19) with STAR, and then sequencing read counts for each GENCODE gene were calculated using RSEM. The expression levels of different samples were merged into a fragments per kilobase transcriptome per million fragments (FPKM) matrix. We defined the expressed gene only if FPKM is larger than 0 in half of the samples. We retained only the expressed genes in the mRNA expression profile.

Data processing for mRNA microarray data

A rapid hematoxylin & eosin stain for frozen sections was applied to each sample to assess the tumor cell proportion before RNA extraction. RNA was extracted from only those samples with >80% tumor cells. Total RNA was extracted from frozen tumor tissue with the Ambion mirVana miRNA Isolation kit (AM1560, Austin, TX) as described previously [31]. The NanoDrop ND-1000 spectrophotometer (Wilmington, DE) was applied to evaluate the quality and concentration of the extracted total RNA, and the Agilent 2100 Bioanalyzer (Agilent, Santa Clara, CA) was used to assess RNA integrity. Then, the qualified RNA was collected for further procedures. cDNA and biotinylated cRNA were synthesized and hybridized to the Agilent Whole Human Genome Array according to the manufacturer's instructions. Finally, the array-generated data were analyzed by the Agilent G2565BA Microarray Scanner System and Agilent Feature Extraction software (V9.1). GeneSpring GX11.0 was applied to calculate probe intensity.

Data processing for methylation microarray data

A hematoxylin and eosin-stained frozen section was prepared for assessment of the percentage of tumor cells before RNA extraction. Only samples with greater than 80% tumor cells were selected. Genomic DNA was isolated from frozen tumor tissues using the QIAGEN QIAamp DNA Mini Kit (Cat No. 51304, Dusseldorf, Germany) according to the manufacturer's protocol. DNA concentration and quality were measured using the NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Houston, TX). We used the Illumina Infinium HumanMethylation27 Bead-Chip (Illumina Inc.). The Bead-Chip contains 27,578 highly informative CpG sites covering more than 14,000 human RefSeq genes. This array allows researchers to interrogate all of these sites per sample at a single nucleotide resolution. Bisulfite modification of DNA, chip processing and data analysis were performed following the manufacturer's manual at the Wellcome Trust Centre for Human Genetics Genomics Lab, Oxford, UK. The array results were analyzed the BeadStudio software (Illumina, San Diego, CA).

Data processing for microRNA microarray data

Total RNA was extracted from frozen tissues by using the mirVana miRNA Isolation Kit (Ambion, Inc., Austin, Tex), and its concentration and quality were determined with the NanoDrop ND-1000 spectrophotometer (Technologies, Wilmington, Del). microRNA expression profiling was performed using the human v2.0 miRNA Expression BeadChip (Illumina, Inc., San Diego, California) with 1146 miRNAs covering 97% of the miRBase 12.0 database according to the manufacturer's instructions.

Discussion and perspectives

The current version of CGGA is the first release of the database, which includes multi-dimensional functional genomic glioma data, *e.g.*, whole-exome sequencing, mRNA and microRNA expression, and DNA methylation data, for nearly 2,000 samples from Chinese cohorts. Considering the significance of these data for glioma research, we have decided to make CGGA publicly available for worldwide researchers. To the best of our knowledge, CGGA is the first database archiving functional genomic data of both recurrent LGGs and GBMs. In addition, CGGA provides online

interactive functionalities, including mutational profile, gene expression distribution pattern, correlation and survival analyses. Phenotype-focused exploration, differential expression analysis, and cluster analysis can be performed by uploading rearranged gene matrixes and online tools. These features will be convenient for generating and validating novel findings of biological significance for bioinformatics beginners.

However, the current version of CGGA is still nascent. The visitor-interactive functionalities will be improved in future updates. Unlike TCGA, CGGA lacks neuroimaging data, a limitation of the database; these data will be uploaded in the near future. In addition to addressing these shortcomings, several future directions for our CGGA database are planned. First, relying on the Beijing Neurosurgical Institute, Beijing Tiantan Hospital and Chinese Glioma Cooperative Group (CGCG) Research Network, we will continue collecting glioma tissue samples, performing cross-omics sequencing/microarray analyses, and updating the database regularly. In addition, we plan to provide single-cell sequencing data that match a subset of patients in the existing cohort. Furthermore, we will improve the integrity of the molecular pathological data by retrospectively checking medical records or redetecting pathological slices.

In summary, CGGA provides access to multi-omics sequencing data on Chinese cohorts for the global research community. It provides a user-friendly interface for obtaining integrated datasets, performing intuitive visualized analysis, and downloading these datasets. CGGA greatly reduces the barriers to access of complex functional genomic data for glioma researchers, allowing them to use functional genomic data to important biological insights and identify potential clinical applications.

Ethical statements

All research performed was approved by the Beijing Tiantan Hospital Capital Medical University Institutional Review Board (IRB) and was conducted according to the principles of the Helsinki Declaration. All patients received written informed consent.

Data availability

All data referred to in this article can be found online at http://cgga.org.cn/. The raw sequence data reported in this article have been deposited in the Genome Sequence Archive [32] at the National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation (GSA: HRA000071, HRA000073, and HRA000074), and are publicly accessible at http://bigd.big.ac.cn/gsa-human.

CRediT author statement

Zheng Zhao: Methodology, Software, Writing - Original Draft, Visualization. **Ke-Nan Zhang:** Methodology, Formal analysis, Investigation, Writing - Review & Editing. **Qiangwei Wang:** Formal analysis, Investigation, Data Curation, Writing - Review & Editing. **Guanzhang Li:** Investigation, Data Curation. **Fan Zeng:** Investigation, Data Curation. **Ying Zhang:** Data Curation. **Fan Wu:** Resources, Data Curation. **Ruichao Chai:** Resources, Data Curation. **Zheng Wang:** Resources, Data Curation. **Chuanbao Zhang:** Data Curation. **Wei Zhang:** Conceptualization, Validation, Project administration, Funding acquisition. **Zhaoshi Bao:** Conceptualization, Supervision. **Tao Jiang:** Conceptualization, Resources, Supervision, Project administration, Funding acquisition.

Competing interests

The authors have declared no competing interests.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (NSFC) fund (Nos. 81702460 and 81802994).

ORCID

0000-0001-8945-9632 (Zheng Zhao);

0000-0001-7270-569X (Ke-Nan Zhang);

Journal Pre-proofs

0000-0002-7308-049X (Qiangwei Wang);

0000-0002-0353-5751 (Guanzhang Li);

0000-0001-5351-2155 (Fan Zeng);

0000-0002-7613-6188 (Ying Zhang);

0000-0001-9256-0176 (Fan Wu);

0000-0003-3451-8871 (Ruichao Chai);

0000-0003-1687-6990 (Zheng Wang);

0000-0003-2615-4190 (Chuanbao Zhang);

0000-0001-7800-3189 (Wei Zhang);

0000-0003-4922-4470 (Zhaoshi Bao);

0000-0002-7008-6351 (Tao Jiang).

References

[1] Jiang T, Tang GF, Lin Y, Peng XX, Zhang X, Zhai XW, et al. Prevalence estimates for primary brain tumors in China: a multi-center cross-sectional study. Chin Med J (Engl) 2011;124:2578−83.

[2] Zhao Z, Meng F, Wang W, Wang Z, Zhang C, Jiang T. Comprehensive RNA-seq transcriptomic profiling in the malignant progression of gliomas. Sci Data 2017;4:170024.

[3] Jiang T, Mao Y, Ma W, Mao Q, You Y, Yang X, et al. CGCG clinical practice guidelines for the management of adult diffuse gliomas. Cancer Lett 2016;375:263−73.

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

[5] Van Meir EG, Hadjipanayis CG, Norden AD, Shu HK, Wen PY, Olson JJ. Exciting new advances in neuro-oncology: the avenue to a cure for malignant glioma. CA Cancer J Clin 2010;60:166−93.

[6] Louis DN, Perry A, Reifenberger G, von Deimling A, Figarella-Branger D, Cavenee WK, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol 2016;131:803−20.

[7] Sun L, Zhang C, Yang Z, Wu Y, Wang H, Bao Z, et al. KIF23 is an independent prognostic biomarker in glioma, transcriptionally regulated by TCF-4. Oncotarget 2016;7:24646−55.

[8] Hu H, Mu Q, Bao Z, Chen Y, Liu Y, Chen J, et al. Mutational landscape of secondary glioblastoma guides MET-targeted trial in brain tumor. Cell 2018;175:1665−78 e18.

[9] Chaichana KL, Jusue-Torres I, Navarro-Ramirez R, Raza SM, Pascual-Gallego M, Ibrahim A, et al. Establishing percent resection and residual volume thresholds affecting survival and recurrence for patients with newly diagnosed intracranial glioblastoma. Neuro Oncol 2014;16:113−22.

[10] Aldape K, Brindle KM, Chesler L, Chopra R, Gajjar A, Gilbert MR, et al. Challenges to curing primary brain tumours. Nat Rev Clin Oncol 2019;16:509−20.

[11] Yi GZ, Huang G, Guo M, Zhang X, Wang H, Deng S, et al. Acquired temozolomide resistance in *MGMT*-deficient glioblastoma cells is associated with regulation of DNA repair by DHC2. Brain 2019;142:2352−66.

[12] Frosina G. DNA repair and resistance of gliomas to chemotherapy and radiotherapy. Mol Cancer Res 2009;7:989−99.
[13] Lee JK, Wang J, Sa JK, Ladewig E, Lee HO, Lee IH, et al. Spatiotemporal genomic architecture informs precision oncology in glioblastoma. Nat Genet 2017;49:594−9.
[14] Liu Q, Liu Y, Li W, Wang X, Sawaya R, Lang FF, et al. Genetic, epigenetic, and molecular landscapes of multifocal and multicentric glioblastoma. Acta Neuropathol 2015;130:587−97.
[15] Barthel FP, Johnson KC, Varn FS, Moskalik AD, Tanner G, Kocakavuk E, et al. Longitudinal molecular trajectories of diffuse glioma in adults. Nature 2019;576(7785):112−120.
[16] Wang Q, Hu B, Hu X, Kim H, Squatrito M, Scarpace L, et al. Tumor evolution of glioma-intrinsic gene expression subtypes associates with immunological changes in the microenvironment. Cancer Cell 2017;32:42−56 e6.
[17] Quail DF, Joyce JA. The microenvironmental landscape of brain tumors. Cancer Cell 2017;31:326−41.
[18] Tomczak K, Czerwinska P, Wiznerowicz M. The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol (Pozn) 2015;19:A68−77.
[19] Zhang J, Bajari R, Andric D, Gerthoffert F, Lepsa A, Nahal-Bose H, et al. The International Cancer Genome Consortium Data Portal. Nat Biotechnol 2019;37:367−9.
[20] Zhang J, Baran J, Cros A, Guberman JM, Haider S, Hsu J, et al. International Cancer Genome Consortium Data Portal--a one-stop shop for cancer genomics data. Database (Oxford) 2011;2011:bar026.
[21] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, et al. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2012;2:401−4.
[22] Gao J, Aksoy BA, Dogrusoz U, Dresdner G, Gross B, Sumer SO, et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal 2013;6(269):pl1.
[23] Bowman RL, Wang Q, Carro A, Verhaak RG, Squatrito M. GlioVis data portal for visualization and analysis of brain tumor expression datasets. Neuro Oncol 2017;19:139−41.
[24] Cancer Genome Atlas Research N, Brat DJ, Verhaak RG, Aldape KD, Yung WK, Salama SR, et al. Comprehensive, integrative genomic analysis of diffuse lower-grade gliomas. N Engl J Med 2015;372:2481−98.
[25] Zhao Z, Zhang KN, Chai RC, Wang KY, Huang RY, Li GZ, et al. ADAMTSL4, a secreted glycoprotein, is a novel immune-related biomarker for primary glioblastoma multiforme. Dis Markers 2019;2019:1802620.
[26] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754−60.
[27] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078−9.
[28] Narasimhan V, Danecek P, Scally A, Xue Y, Tyler-Smith C, Durbin R. BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data. Bioinformatics 2016;32:1749−51.
[29] Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 2013;29:15−21.
[30] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 2011;12:323.
[31] Yan W, Zhang W, You G, Zhang J, Han L, Bao Z, et al. Molecular classification of gliomas based on whole genome gene expression: a systematic report of 225 samples from the Chinese Glioma Cooperative Group. Neuro Oncol 2012;14:1432−40.
[32] Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, et al. GSA: Genome Sequence Archive. Genomics Proteomics Bioinformatics 2017;15:14−8.

Figure legends

Figure 1 Schematic of CGGA illustrating the data processing and display approaches.

Figure 2 Main contents of CGGA database and the functionality of WEseq analysis.

A. CGGA contains whole-exome sequencing, mRNA and microRNA expression, and DNA methylation data; clinical data; and several analysis modules. **B.** The web images in the WEseq analysis page to search the oncoprint and prognostic value of target genes. **C.** The mutation profile in all grade gliomas (in the 'WSseq_286' dataset). **D.** Left: The overall survival of glioma patients with *IDH1* mutation and the wild-type gene from all grade gliomas (in the 'WSseq_286' dataset); Middle: The data used to generate the plot; Right: The R code used to generate the plot.

Figure 3 Examples of CGGA RNA-seq analysis.

A. The screenshot of the RNA-seq analysis page to search the distribution, correlated genes and prognostic value of target gene. **B.** The *ADAMTSL4* gene expression distribution in primary gliomas based on the 2016 WHO grading system (in the 'mRNAseq_325' dataset). **C.** The correlation of gene expression between *ADAMTSL4* and *CD274* (in the 'mRNAseq_325' dataset). **D.** The overall survival of glioma patients with low and high expression of *ADAMTSL4* (in the 'mRNAseq_325' dataset).

Tables

Table 1 Clinical and phenotypical characteristics of dataset in CGGA database

CRediT author statement

Zheng Zhao: Methodology, Software, Writing - Original Draft, Visualization. **Ke-Nan Zhang:** Methodology, Formal analysis, Investigation, Writing - Review & Editing. **Qiangwei Wang:** Formal analysis, Investigation, Data Curation, Writing - Review & Editing. **Guanzhang Li:** Investigation, Data Curation. **Fan Zeng:** Investigation, Data Curation. **Ying Zhang:** Data Curation. **Fan Wu:** Resources, Data Curation. **Ruichao Chai:** Resources, Data Curation. **Zheng Wang:** Resources, Data Curation. **Chuanbao Zhang:** Data Curation. **Wei Zhang:** Conceptualization, Validation, Project administration, Funding acquisition. **Zhaoshi Bao:** Conceptualization, Supervision. **Tao Jiang:** Conceptualization, Resources, Supervision, Project administration,

Funding acquisition. All authors read and approved the final manuscript.

Table 1 Clinical and phenotypical characteristics of dataset in CGGA database

	All	Primary LGG	Recurrent LGG	Primary GBM	Recurrent GBM	Secondary GBM
WEseq_286 dataset						
No. of samples –No. (%)	286	126 (44%)	58 (20%)	54 (19%)	48 (17%)	0 (0%)
Age at diagnosis – year						
Mean	42.0 ± 12.3	39.6 ± 10.3	37.3 ± 8.7	50.2 ± 14.7	44.5 ± 13.3	–
Range	10–76	10–69	15–61	19–76	19–69	–
Male sex – No. (%)	168	78 (46%)	35 (21%)	29 (17%)	26 (15%)	–
Therapy						
Radiotherapy only	62	52 (84%)	4 (6%)	4 (6%)	2 (3%)	–
Chemotherapy	13	8 (62%)	2 (15%)	0 (0%)	3 (23%)	–
Chemoradiotherapy	144	49 (34%)	27 (19%)	42 (29%)	26 (18%)	–
No therapy	23	9 (39%)	8 (35%)	4 (17%)	2 (9%)	–
Unknown	44	8 (18%)	17 (39%)	4 (9%)	15 (34%)	–
Survival – month						
Median (95% CI)	51.0 (37.2–98.1)	117.2 (99.4–NA)	28.5 (20.9–76.0)	16.5 (10.2–28.7)	14.7 (8.9–NA)	–
IDH_mut_status						
Mutant	161	88 (55%)	45 (28%)	12 (7%)	16 (10%)	–
Wildtype	125	38 (30%)	13 (10%)	42 (34%)	32 (26%)	–
1p19q_codeletion_status						
Codel	51	28 (55%)	17 (33%)	1 (2%)	5 (10%)	–
Non–codel	139	48 (35%)	33 (24%)	23 (17%)	35 (25%)	–
Unknown	96	50 (52%)	8 (8%)	30 (31%)	8 (8%)	–
RNAseq_1018 dataset						
No. of samples –No. (%)	1018	426 (42%)	199 (20%)	225 (22%)	133 (13%)	30 (3%)

Journal Pre-proofs

Age at diagnosis – year						
Mean	43.2 ± 12.3	40.2 ± 10.8	40.2 ± 9.6	51.0 ± 12.9	45.0 ± 13.2	38.8 ± 11.4
Range	8–79	10–74	15–64	11–79	14–71	8–57
Male sex – No. (%)	601	247 (41%)	115 (19%)	138 (23%)	76 (13%)	21 (3%)
Therapy						
Radiotherapy only	200	128 (64%)	32 (16%)	26 (13%)	10 (5%)	4 (2%)
Chemotherapy	68	30 (44%)	13 (19%)	9 (13%)	11 (16%)	5 (7%)
Chemoradiotherapy	567	204 (36%)	102 (18%)	159 (28%)	85 (15%)	15 (3%)
No therapy	89	41 (46%)	21 (24%)	18 (20%)	5 (6%)	4 (4%)
Unknown	91	23 (25%)	31 (34%)	13 (14%)	22 (24%)	2 (2%)
Survival – month						
Median (95% CI)	35.0 (30.5–39.9)	108.0 (89.9–NA)	33.2 (26.1–39.8)	16.1 (13.7–19.7)	9.6 (8.2–11.0)	8.3 (7.1–14.7)
IDH_mut_status						
Mutant	531	289 (54%)	150 (28%)	35 (7%)	34 (6%)	21 (4%)
Wildtype	435	104 (24%)	40 (9%)	183 (42%)	96 (22%)	9 (2%)
Unknown	52	33 (63%)	9 (17%)	7 (13%)	3 (6%)	0
1p19q_codeletion_status						
Codel	212	137 (65%)	54 (25%)	5 (2%)	11 (5%)	4 (2%)
Non–codel	728	254 (35%)	139 (19%)	192 (26%)	118 (16%)	24 (3%)
Unknown	78	35 (45%)	6 (8%)	28 (36%)	4 (5%)	2 (3%)
mRNA–array_301 dataset						
No. of samples –no. (%)	301	156 (52%)	18 (6%)	108 (36%)	5 (2%)	11 (4%)
Age at diagnosis – year						
Mean	42.4 ± 11.8	39.6 ± 10.7	38.2 ± 11.2	47.3 ± 12.5	45.6 ± 9.6	38.5 ± 8.6
Range	12–70	17–65	24–62	12–70	36–61	27–51
Male sex – No. (%)	180	93 (52%)	8 (4%)	65 (36%)	2 (1%)	9 (5%)
Therapy						
Radiotherapy only	110	74 (67%)	0	33 (30%)	0	3 (3%)

Journal Pre-proofs

Chemotherapy	12	1 (8%)	2 (17%)	4 (33%)	3 (25%)	2 (17%)
Chemoradiotherapy	139	61 (44%)	12 (9%)	60 (43%)	1 (1%)	4 (3%)
No therapy	20	8 (40%)	2 (10%)	6 (30%)	0	2 (10%)
Unknown	20	12 (60%)	2 (10%)	5 (25%)	1 (5%)	0
Survival – month						
Median (95% CI)	38.8 (27.2–53.9)	– (99.8–NA)	39.8 (13.8–NA)	15.4 (13.3–19.0)	10.5 (7.7–NA)	7.2 (6.5–NA)
IDH_mut_status						
Mutant	134	100 (75%)	12 (9%)	14 (10%)	2 (1%)	6 (4%)
Wildtype	165	54 (33%)	6 (4%)	94 (57%)	3 (2%)	5 (3%)
Unknown	2	2 (100%)	0	0	0	0
1p19q_codeletion_status						
Codel	16	14 (88%)	2 (12%)	0	0	0
Non–codel	76	23 (30%)	14 (18%)	27 (36%)	5 (7%)	7 (9%)
Unknown	209	119 (57%)	2 (1%)	81 (39%)	0	4 (2%)
methyl_159 dataset						
No. of samples –No. (%)	159	100 (63%)	8 (5%)	33 (21%)	4 (3%)	6 (4%)
Age at diagnosis – year						
Mean	40.2 ± 12.5	39.5 ± 12.2	35.6 ± 12.0	44.2 ± 14.2	41.5 ± 3.7	33.7 ± 7.4
Range	9–70	17–70	24–57	9–70	38–46	27–46
Male sex – no. (%)	89	58 (65%)	4 (4%)	19 (21%)	3 (3%)	5 (6%)
Therapy						
Radiotherapy only	48	39 (81%)	1 (2%)	8 (17%)	0	0
Chemotherapy	10	0	3 (30%)	1 (10%)	3 (30%)	3 (30%)
Chemoradiotherapy	66	46 (70%)	3 (5%)	16 (24%)	1 (2%)	0
No therapy	12	4 (33%)	1 (8%)	4 (33%)	0	3 (25%)
Unknown	19	11 (58%)	2 (5%)	4 (21%)	3 (16%)	0
Survival – month						
Median (95% CI)	45.8 (36.6–83.9)	107.2 (60.4–NA)	85.0 (43.8–NA)	8.5 (6.4–23.1)	16.0 (5.2–NA)	43.3 (10.6–NA)

Journal Pre-proofs

IDH_mut_status						
Mutant	81	65 (80%)	5 (6%)	5 (6%)	2 (2%)	4 (5%)
Wildtype	64	30 (47%)	3 (5%)	27 (42%)	2 (3%)	2 (3%)
Unknown	14	5 (36%)	0	1 (7%)	0	0
1p19q_codeletion_status						
Codel	7	5 (71%)	2 (29%)	0	0	0
Non–codel	18	7 (39%)	3 (17%)	2 (11%)	2 (11%)	4 (22%)
Unknown	134	88 (66%)	3 (2%)	31 (23%)	2 (1%)	2 (1%)
microRNA–array_198 dataset						
No. of samples –No. (%)	198	99 (50%)	8 (4%)	81 (41%)	4 (2%)	6 (3%)
Age at diagnosis – year						
Mean	41.9 ± 12.5	39.5 ± 12.3	35.6 ± 12.0	46.1 ± 13.1	41.5 ± 3.7	33.7 ± 7.4
Range	12–70	17–70	24–57	12–70	38–46	27–46
Male sex – No. (%)	123	57 (46%)	4 (3%)	54 (44%)	3 (2%)	5 (4%)
Therapy						
Radiotherapy only	57	38 (67%)	1 (2%)	18 (32%)	0	0
Chemotherapy	12	0	3 (25%)	3 (25%)	3 (25%)	3 (25%)
Chemoradiotherapy	99	47 (47%)	3 (3%)	48 (48%)	1 (1%)	0
No therapy	15	4 (27%)	1 (7%)	7 (47%)	0	3 (20%)
Unknown	15	10 (67%)	0	5 (33%)	0	0
Survival – month						
Median (95% CI)	28.4(22.1–43.8)	121.6 (60.4–NA)	85.0 (43.8–NA)	13.7 (12.7–18.8)	16.0 (5.2–NA)	43.3 (10.6–NA)
IDH_mut_status						
Mutant	81	63 (78%)	5 (6%)	7 (9%)	2 (2%)	4 (5%)
Wildtype	106	30 (28%)	3 (3%)	69 (65%)	2 (2%)	2 (2%)
Unknown	11	6 (55%)	0	5 (45%)	0	0
1p19q_codeletion_status						
Codel	7	5 (71%)	2 (29%)	0	0	0

Journal Pre-proofs

Non–codel	19	7 (37%)	3 (16%)	3 (16%)	2 (11%)	4 (21%)
Unknown	172	87 (51%)	3 (2%)	78 (45%)	2 (1%)	2 (1%)

Journal Pre-proofs