Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data for Chinese Glioma Patients

Zheng Zhao[1,#,a], Ke'nan Zhang[1,#,b], Qiangwei Wang[1,#,c], Guanzhang Li[1,d], Fan Zeng[1,e], Ying Zhang[1,f], Fan Wu[1,g], Ruichao Chai[1,h], Zheng Wang[2,i], Chuanbao Zhang[2,j], Wei Zhang[2,k], Zhaoshi Bao[2,*,l], Tao Jiang[1,2,3,4,*,m]

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

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

[3]Center of Brain Tumour, Beijing Institute for Brain Disorders, Beijing 100069, China

[4]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*

[a]ORCID: 0000-0001-8945-9632

[b]ORCID: 0000-0001-7270-569X

[c]ORCID: 0000-0002-7308-049X

[d]ORCID: 0000-0002-0353-5751

[e]ORCID: 0000-0001-5351-2155

[f]ORCID: 0000-0002-7613-6188

[g]ORCID: 0000-0001-9256-0176

[h]ORCID: 0000-0003-3451-8871

[i]ORCID: 0000-0003-1687-6990

[j]ORCID: 0000-0003-2615-4190

[k]ORCID: 0000-0001-7800-3189

[l]ORCID: 0000-0003-4922-4470

[m]ORCID: 0000-0002-7008-6351

Total word counts: 3,030

Total references: 31

Total figures: 1

Total tables: 1

Total supplementary figures: 0

Total supplementary tables: 0

Total supplementary files: 0

Abstract:

Gliomas are the most common and malignant intracranial tumours in adults. Recent studies have shown that functional genomics greatly aids in the understanding of the pathophysiology and therapy of glioma. However, comprehensive genomic data and analysis platforms are relatively limited. In this study, we developed the Chinese Glioma Genome Atlas (CGGA, http://www.cgga.org.cn), a user-friendly data portal for storage and interactive exploration of multi-dimensional functional genomic data that includes nearly 2,000 primary and recurrent glioma samples from Chinese cohorts. CGGA currently provides access to whole-exome sequencing (286 samples), messenger RNA sequencing (1,018 samples) and microarray (301 samples), DNA methylation microarray (159 samples), and microRNA microarray (198 samples) data, as well as detailed clinical data (e.g., WHO grade, histological type, critical molecular genetic information, age, sex, chemoradiotherapy status and survival data). In addition, we developed an analysis tool to allow users to browse mutational, mRNA/microRNA expression, and DNA methylation profiles and perform survival and correlation analyses of specific glioma subtypes. CGGA greatly reduces the barriers between complex functional genomic data and glioma researchers who seek rapid, intuitive, and high-quality access to data resources and enables researchers to use these immeasurable data sources for biological research and clinical application. Importantly, the free provision of data will allow researchers to quickly generate and provide data to the research community.

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

Introduction

Gliomas are the most frequent malignant tumours of the adult brain. According to a multi-centre cross-sectional study on brain tumours in China, the prevalence of primary brain tumours in all populations is approximately 22.52 per 100,000 persons, with gliomas accounting for 31.1% of the population aged 20–59 years [1-3]. According to the histopathological classification of the 2016 World Health Organization (WHO) grading system, glioma is classified from grade II to grade IV by both histological characteristics and several new molecular pathological features, such as *IDH* mutation status and chromosome 1p/19q co-deletion status [4]. Despite advances in current treatment standards, the survival rate of patients with glioma has not changed in decades, especially for aggressive gliomas (with a poor median survival time of only 12 to 14 months) [5, 6]. In addition, most lower-grade gliomas (grade II and III, LGG) will progress to glioblastoma (grade IV, GBM) in less than 10 years [4, 7, 8]. At present, the reasons for glioma recurrence or malignant progression may be as follows: 1) infiltrative tumour cells cannot be completely removed by neurosurgical resection [9, 10]; 2) retained tumour cells cannot be effectively suppressed by limited postoperative treatment options [3, 11, 12]; 3) multiple lesions may develop [13, 14]; 4) cell cloning is rapid under chemotherapy and/or radiotherapy [7, 15]; 5) the adaptive tumour microenvironment permits tumour cells [16, 17]; and 6) limited data resources lead to limited research. Therefore, it is essential to collect clinical specimens and generate genomic data for the glioma research community.

Recent high-throughput technologies have enabled extensive characterization of 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, including 516 LGGs and 617 GBMs before Oct. 18, 2019) [18] and the International Cancer Genome Consortium (ICGC, excluding TCGA samples, including 80 adult GBMs and 50 paediatric GBMs before April. 3, 2019) [19, 20], have generated an unparalleled amount of functional genomic data. These projects have begun to transform our understanding of cancer and even lead to improvements in our ability to diagnose, treat, and prevent human cancers. Importantly, they have provided an opportunity to make and validate important discoveries for cancer genomic researchers around the globe. However, the data resources generated by these projects are often not easy to access directly, analyse or visualize, especially for researchers with no

bioinformatics skills, thus preventing the translation of functional genomics results into novel findings of biological significance for drug development and clinical treatment. Although several webservers, such as cBioportal [21, 22] and GlioVis [23], have been built to integrate analysed multi-dimensional glioma data, they have ignored the presence of cancer heterogeneity in gliomas, which cannot be examined in specific subtypes and is rarely found in recurrent glioma samples.

Here, we introduce the CGGA (Chinese Glioma Genome Atlas, http://www.cgga.org.cn) database, which is an open-access and easy-to-use platform for interactive exploration of multi-dimensional functional genomic datasets for nearly 2,000 primary and recurrent glioma samples from Chinese cohorts. CGGA currently contains whole-exome sequencing (286 samples), messenger RNA (mRNA) sequencing (1,018 samples), microarray (301 samples), DNA methylation microarray (159 samples), microRNA microarray (198 samples) and comprehensive clinical data. We also developed an analysis module to allow users to browse the mutational landscape profile, mRNA/microRNA expression profile and DNA methylation profile as well as to perform survival and correlation analyses for specific glioma subtypes. We believe that this website will greatly reduce the barriers between complex functional genomic data and glioma researchers who seek rapid, intuitive, and high-quality access to data resources.

Results

Database content and usage

The CGGA database was designed to store functional genomic data and to allow interactive exploration of multi-dimensional datasets from primary and recurrent gliomas in Chinese cohorts; it is available at http://www.cgga.org.cn/. Currently, CGGA contains whole-exome sequencing data (286 samples), messenger RNA sequencing data (total: 1,018 samples, batch 1 with 693 samples and batch 2 with 325 samples), microarray data (301 samples), DNA methylation microarray data (159 samples), and microRNA microarray data (198 samples) for glioma. The database also contains detailed clinical data (including WHO grade and histological type, critical molecular genetic information, age, sex, chemoradiotherapy status and survival data). Detailed statistical information for each dataset is provided in **Table 1**. We organized the web interface of CGGA according to the three main functional features: (i) Home, (ii) Analyse, and (iii) Download. In the following context, we provide an example for using CGGA.

The home page

On the 'Home' page, CGGA provides a statistical table for a glioma dataset, including the dataset name, data type, number of samples in each subgroup, clinical data and analysis purposes. For instance, we performed messenger RNA sequencing on 1,018 glioma samples included in two datasets (693 samples in batch 1 and 325 samples for 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 GBMs, 24 recurrent GBMs and 30 secondary GBMs in batch 2). To the best of our knowledge, CGGA is the first database to store the functional genomic data for both LGG and GBM recurrent gliomas. In addition, users can obtain a visualized result for the analysis of each dataset for a specific glioma subtype by clicking on a hyperlink on the 'Home' page. The 'Download' and 'Help' pages can also be accessed directly from the 'Home' page.

Overall analyses and results

To facilitate analysis of the CGGA data by researchers, we developed four online modules in the 'Analyse' tab, including 'WEseq data', 'mRNA data', 'methylation data', and 'microRNA data', to analyse whole-exome, mRNA expression, DNA methylation and microRNA expression data, respectively (**Figure 1A**). A key feature of CGGA is that it is easy to use. In the context below, we demonstrate the use of the 'Analyse' tab in CGGA.

On the 'WEseq data' page, users are allowed to visualize the mutational profile of a gene set of interest and survival analysis of a specific gene of interest in a specific glioma subtype. In the 'Oncoprint' section, users are guided to a) input a gene set of interest ('*IDH1 TP53 ATRX'* for example), and b) select a dataset of interest ('All' for example). Based on user input, this tool automatically generates visualized results. In this result, each case or patient is represented as columns, each gene is displayed as rows, and a colour map on the bottom is used to depict specific clinical information (**Figure 1B**). This 'Oncoprint' can be very useful for visualizing the mutational profile for a gene set of interest in a specific glioma subtype and for intuitively validating trends such as mutational frequency and mutual exclusivity or co-occurrence for a gene pair. In the above example, mutations in the *IDH1* (47%), *TP53* (46%) and *ATRX* (30%) genes were the most common mutations in all gliomas. In the 'Survival' section, users are allowed to a) input a specific gene of interest ('*IDH1'* for example), and b) select a dataset of interest ('Primary LGG*'* for example) to investigate the association of the mutation with severe functional consequences. Consistent with

previous studies [24], primary LGG cases with *IDH1* mutations have a better overall survival than do cases with *IDH1* wild-type tumours ($p < 0.0001$, **Figure 1C,** left). These analysis results from the 'WEseq data' section can be exported as a PDF file. For the sake of reproducibility, we provide the analysis data (**Figure 1C,** middle) and R code (**Figure 1C,** right), which allow users to reproduce the figure to be able to modify or adapt each figure according to each researcher's demands.

On the 'mRNA data' page, users are allowed to perform gene expression distribution, correlation and survival analyses for a specific gene of interest in a specific glioma subtype. Three mRNA datasets are available for users, including two batch 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 ('mRNAseq_325' for example) and inputting a gene name of interest ('*ADAMTSL4'* for example). The results show the gene expression pattern in each glioma subtype classified by clinical information. 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 (**Figure 1D**, left). Moreover, a critical 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 is allowed to validate the co-expression pattern by selecting a dataset ('mRNAseq_325' for example) and inputting a gene pair ('*ADAMTSL4*' and '*CD274*' for example). As a result, the co-expression patterns in each glioma subtype will be displayed with the results of Pearson's test and the p value (**Figure 1D**, middle). In the 'Survival' section, users can perform survival analysis based on gene expression by selecting a dataset ('mRNAseq_325' for example) and inputting a gene of interest ('*ADAMTSL4*' for example). All primary glioma patients with low *ADAMTSL4* expression showed better overall survival than did those with high *ADAMTSL4* expression ($p < 0.0001$, **Figure 1D**, right). The above results from the 'mRNA data' section are consistent with our previous study [25]. Similar to the 'mRNA data' page, users can also display the methylation/microRNA distribution and perform correlation and survival analyses on the 'methylation data' page and the 'microRNA data' page, respectively.

Data acquisition

All the data sets in CGGA can be downloaded on the 'Download' page by both the community and researchers. Each data type is saved at the gene and/or probe level and is then combined with the available clinical data, including basic clinical information, survival and therapy information.

Perspectives and concluding remarks

The current version of the CGGA is the first release of our database, and it incorporates multi-dimensional functional genomic glioma data, including whole-exome sequencing, mRNA and microRNA expression, and DNA methylation data for nearly 2,000 samples from Chinese cohorts. Considering the importance of these data for glioma research, CGGA is publicly available. To the best of our knowledge, CGGA is the first database to store the functional genomic data for both recurrent LGGs and GBMs. In addition, CGGA provides several tools that allow users to analyse these datasets, including mutational profile, distribution pattern, correlation and survival analysis tools. These tools will be useful for users to generate or validate findings of novel biological significance.

We anticipate several future directions for our CGGA database. First, through the Beijing Neurosurgical Institute, Beijing Tiantan Hospital and Chinese Glioma Cooperative Group (CGCG) Research Network, we will continue to collect glioma samples and perform multiple 'Omics' sequencing/microarray analyses, and we will continue to update this database regularly in the future. Second, we also plan to add image-genomic data that match the 'Omics' data in CGGA. Third, we will develop more advanced features, including data for other 'Omics' analyses, search functions for clinical information on a patient of interest, and further extensions for the data analysis tools. In summary, CGGA facilitates access to functional genomic data for Chinese cohorts for the entire glioma community. It provides an easy-to-use, user-friendly interface for obtaining integrated data sets, performing intuitive visualized analysis, and downloading these datasets. CGGA greatly reduces the barriers between complex functional genomic data and glioma researchers, which empowers researchers to use functional genomic data into important biological insights and potential clinical applications.

Materials and methods

Clinical specimen collection

Glioma tissues, 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. 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. According to the central pathology reviews of independent committee certified neuropathologists, all the subjects were consistently diagnosed with glioma and further classified according to the 2007/2016 WHO classification system. All patients provided written informed consent. 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 tumours and the matched blood samples was extracted, and high integrity was confirmed by 1% agarose gel electrophoresis. The DNA was subsequently fragmented and quality-controlled, and paired-end libraries were prepared. Agilent SureSelect kit v5.4 was used for target capture. Sequencing was performed using the Illumina HiSeq 4000 platform with a paired-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. SAMtools (version 1.2) [27] and Picard (version 2.0.1, Broad Institute) were then used to sort the reads by coordinates and mark duplicates. Statistics such as sequencing depth and coverage were calculated based on the resulting BAM files. SAVI2 was used to identify somatic mutations (including single-nucleotide variations and short insertions/deletions) as previously described [7, 8]. Briefly, in this pipeline, SAMtools mpileup and bcftools (version 0.1.19) [28] were employed to perform variant calling, and the preliminary variant list was filtered to remove positions with no sufficient sequencing depth, positions with only low-quality reads, and positions biased toward either strand. Somatic mutations were identified and evaluated by an empirical Bayesian method. In particular, mutations with a significantly higher mutation allele frequency in tumours than in normal controls were selected.

Data processing for mRNA sequencing data

Prior to library preparation, total RNA was isolated using RNeasy Mini Kit (Qiagen) according to the manufacturer's instructions. A pestle and QIAshredder (Qiagen) were used to disrupt and homogenize frozen tissue. The RNA intensity was checked using a 2,100 Bioanalyzer (Agilent Technologies), 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 TruSeq RNA library preparation kit (Illumina) in accordance with the low-throughput protocol, except that SuperScript III reverse transcriptase (Invitrogen) was used to synthesize first-strand cDNA. After PCR enrichment and purification of adapter-ligated fragments, the concentration of DNA with adapters was determined by quantitative PCR (Applied Biosystems 7,500) using primers QP1 5'-AATGATACGGCGACCACCGA-3' and QP2 5'-CAAGCAGAAGACGGCATACGAGA-3'. The length of the DNA fragment was measured using a 2,100 Bioanalyzer, with median insert sizes of 200 nucleotides. The RNA-seq libraries were sequenced using the Illumina HiSeq 2,000, 2,500 or 4,000 Sequencing System. The libraries were prepared using the paired-end strategy with read lengths 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 processed by using STAR (version v2.5.2b) [29] and RSEM (version 1.2.31) software [30]. Briefly, 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 an FPKM (fragments per kilobase transcriptome per million fragments) matrix. We defined a gene as expressed only if its expression level was greater than 0 in half of the samples. Finally, we retained only expressed genes in the mRNA expression profile.

Data processing for mRNA microarray data

A rapid haematoxylin & eosin stain for frozen sections was performed on each sample to assess the tumour cell proportion before RNA extraction. RNA was extracted from only samples with >80% tumour cells. Total RNA was extracted from frozen tumour tissue with the mirVana miRNA Isolation Kit (Ambion), as described previously [31]. A NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies) was used to evaluate the quality and concentration of extracted total RNA and an Agilent 2100 Bioanalyzer (Agilent) to assess the integrity. The qualified RNA was collected for further processing. cDNA and biotinylated cRNA were synthesized and hybridized to 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 (Version 9.1). GeneSpring GX11.0 was applied to calculate the probe intensity.

Data processing for methylation microarray data

A haematoxylin and eosin-stained frozen section was prepared for assessment of the percentage of tumour cells before RNA extraction. Only samples with greater than 80% tumour cells were selected. Genomic DNA was isolated from frozen tumour tissues using the QIAamp DNA Mini Kit (Qiagen) according to the manufacturer's protocol. The DNA concentration and quality were assessed using a NanoDrop ND-1000 spectrophotometer (NanoDrop Technologies, Houston, TX). The microarray analysis was performed using Illumina Infinium HumanMethylation27 Bead-Chips (Illumina Inc.), which contains 27,578 highly informative CpG sites covering more than 14,000 human RefSeq genes. This allows researchers to investigate all 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 Wellcome Trust Centre for Human Genetics Genomics Lab, Oxford, UK. The array results were examined with the BeadStudio software (Illumina).

Data processing for microRNA microarray data

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

Implementation

In CGGA, all data are organized with MySQL 14.14 based on relational schema, which will be supported by 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 visualization. The 'Analyse' module was realized with Perl and R scripts. The CGGA website has been fully tested in Google Chrome and Safari browsers.

Data Availability

All data for this article can be found online at http://www.cgga.org.cn.

Authors' contributions

TJ, ZB, WZ, and ZZ conceived and supervised this study. ZZ, KZ and QW designed the research. ZZ, GL, FZ, YZ, FW, RC, ZW, CZ performed data analysis. ZZ developed CGGA web server. TJ, ZB, and ZZ wrote the manuscript. All authors read and approved the final manuscript.

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).

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] 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.
[5] 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.
[6] 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.
[7] Wang J, Cazzato E, Ladewig E, Frattini V, Rosenbloom DI, Zairis S, et al. Clonal evolution of glioblastoma under therapy. Nat Genet 2016;48:768-76.
[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. Molecular Cancer Research 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.
[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: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.

Figures

Figure 1. An overview of the CGGA database.

A. The CGGA contains whole-exome sequencing, mRNA and microRNA expression, and DNA methylation data, clinical data, and several analysis modules; B. The mutation profile in all gliomas (in the 'WSseq_286' dataset); C. left: the overall survival of glioma patients with IDH1 mutation and the wild-type gene from primary LGGs (in the 'WSseq_286' dataset); middle: the data was used to generate the plot; right: the R code was used to generate the plot; D. left: the ADAMTSL4 gene expression distribution in primary gliomas based on 2016 WHO grading system (in the 'mRNAseq_325' dataset); middle: the gene expression correlation between ADAMTSL4 and CD274 genes (using 'mRNAseq_325' dataset); right: the overall survival of glioma patients with low and high ADAMTSL4 gene expression (in the 'mRNAseq_325' dataset).

Table 1. Clinical and Phenotypical Characteristics of Data Set 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 – yr.						
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%)	-
Cheomoradiotherapy	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-)	28.5 (20.9-76.0)	16.5 (10.2-28.7)	14.7 (8.9-)	-
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. (%)	1,018	426 (42%)	199 (20%)	225 (22%)	133 (13%)	30 (3%)
Age at diagnosis – yr.						
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

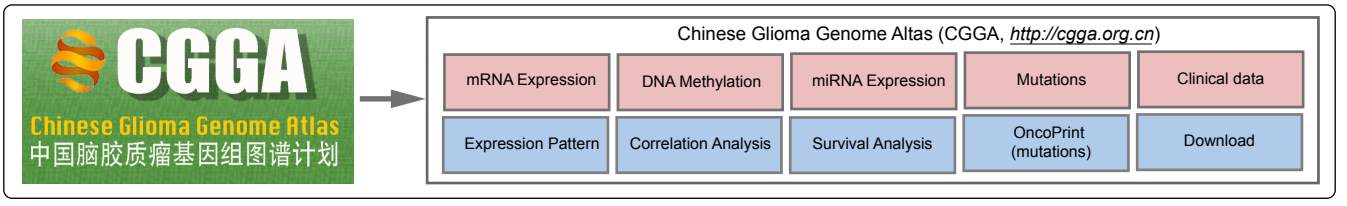
Male sex – no. (%)	601	247 (41%)	115 (19%)	138 (23%)	76 (13%)	21 (3%)
Therapy						
Radiotherapy only	200	128 (64%)	32 (16%)	26 (13%)	10 (5%)	4 (2%)
Chemotherapy	68	30 (44%)	13 (19%)	9 (13%)	11 (16%)	5 (7%)
Cheomoradiotherapy	567	204 (36%)	102 (18%)	159 (28%)	85 (15%)	15 (3%)
No therapy	89	41 (46%)	21 (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-)	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 – yr.						
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%)
Chemotherapy	12	1 (8%)	2 (17%)	4 (33%)	3 (25%)	2 (17%)
Cheomoradiotherapy	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-)	39.8 (13.8-)	15.4 (13.3-19.0)	10.5 (7.7-)	7.2 (6.5-)
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 – yr.						
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%)
Cheomoradiotherapy	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-)	85.0 (43.8-)	8.5 (6.4-23.1)	16.0 (5.2-)	43.3 (10.6-)
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						

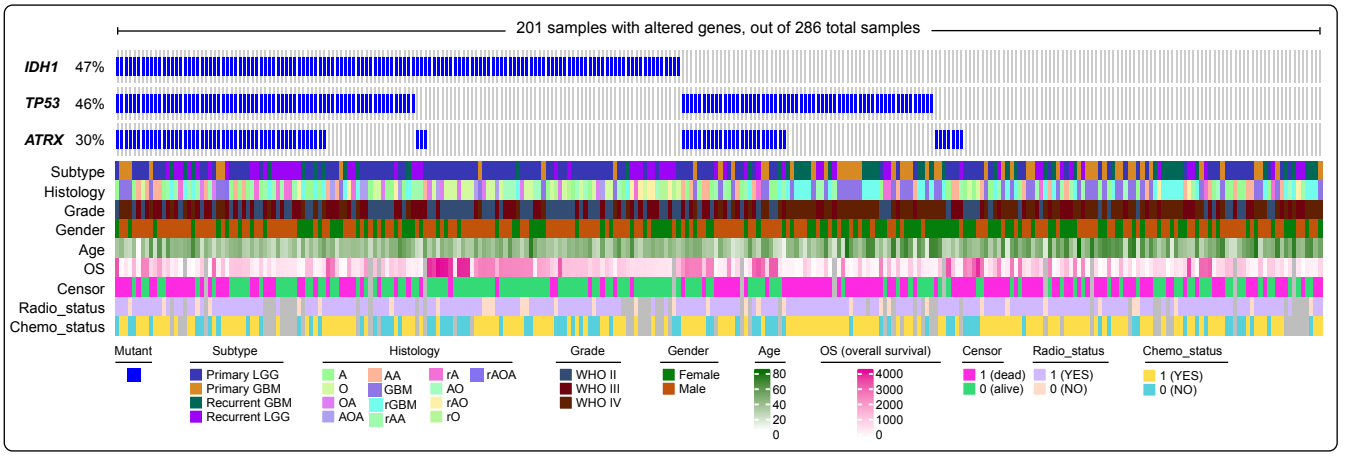
Codel	7	5 (71%)	2 (29%)	0	0	0
Non-codel	18	7 (39%)	3 (17%)	2 (11%)	2 (11%)	4 (22%)
Unknown	134	88 (66%)	3 (2%)	31 (23%)	2 (1%)	2 (1%)
microRNA-array_198 dataset						
No. of samples –no. (%)	198	99 (50%)	8 (4%)	81 (41%)	4 (2%)	6 (3%)
Age at diagnosis – yr.						
Mean	41.9±12.5	39.5±12.3	35.6±12.0	46.1±13.1	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%)
Cheomoradiotherapy	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-)	85.0 (43.8-)	13.7 (12.7-18.8)	16.0 (5.2-)	43.3 (10.6-)
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
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%)

1

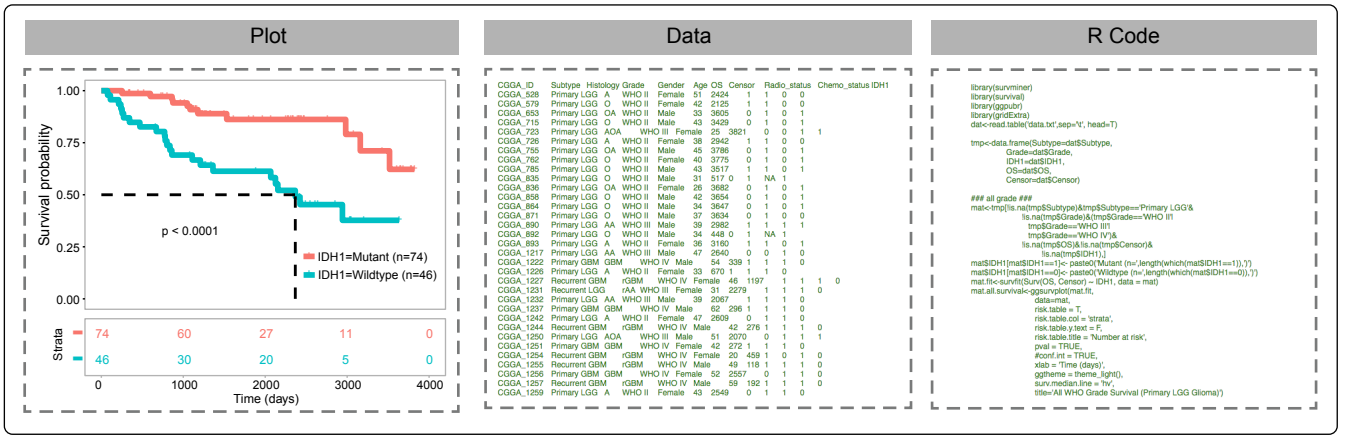
A CGGA Database Content



B OncoPrint of Key Gene Alternations in All Grade Gliomas



C IDH1 Mutation Survival Analysis



D ADAMTSL4 Expression Analysis