



OPEN

DATA DESCRIPTOR

# WES data from 286 diffuse gliomas under the 2021 WHO Classification of Tumors of the Central Nervous System

Zheng Zhao<sup>1,9</sup>, Ke-Nan Zhang<sup>1,9</sup>, Zhiyan Sun<sup>1,9</sup>, Changlin Yang<sup>1,9</sup>, Qiangwei Wang<sup>1,2</sup>, Guanzhang Li<sup>1</sup>, Zhiliang Wang<sup>3</sup>, Fan Zeng<sup>1</sup>, Ruichao Chai<sup>1</sup>, Zenghui Qian<sup>3</sup>, Zheng Wang<sup>3</sup>, Yanwei Liu<sup>4</sup>, Wenping Ma<sup>1,5</sup>✉, Fan Wu<sup>1</sup>✉ & Tao Jiang<sup>1,3,6,7,8</sup>✉

Diffuse gliomas (DGs) are the most common and lethal primary neoplasms in the central nervous system. The latest 2021 World Health Organization (WHO) Classification of Tumors of the Central Nervous System (CNS) was published in 2021, immensely changing the approach to diagnosis and decision making. As a part of the Chinese Glioma Genome Atlas (CGGA) project, our aim was to provide genomic profiling of gliomas in a Chinese cohort. Two hundred eighty six gliomas with different grades were collected over the last decade. Using the Illumina HiSeq platform, over 75.8 million high-quality 150 bp paired-end reads were generated per sample, yielding a total of 43.4 billion reads. We also collected each patient's clinical and pathological information and used it to annotate their genetic data. All patients were diagnosed and classified by neuro-pathologist under the 2021 WHO classification. This dataset provides an important reference for researchers and will significantly advance our understanding of gliomas.

## Background & Summary

Diffuse gliomas (DGs) are the most common and lethal type of primary neoplasm in the central nervous system and are the leading cause of cancer death in adolescents and young adults (AYAs)<sup>1,2</sup>. In part due to the disadvantages in historical classification, standard treatment provided limited benefit<sup>3</sup>. Patients survival ranged from 6 months to decades<sup>4</sup>. Neuropathologists have been devoted to classifying diffuse gliomas more precisely to predict survival and guide treatment strategies for decades and have gradually confirmed the significance and rationality of incorporating molecular characteristics into the classification strategy<sup>5,6</sup>.

The cancer genome atlas (TCGA) and the Rembrandt Project (REpository for Molecular BRAin Neoplasia DaTa) contain thousands of glioma specimens and genomic data<sup>7-9</sup>, but the included patients had incomplete molecular characterization. Asian patients were also severely underrepresented. In 2005, the Chinese Glioma Genome Atlas (CGGA) project began to enroll patients, collect tissue specimens, conduct multi-omics sequencing, and finally developed an online portal in 2020<sup>10</sup>.

Here, as a part of the CGGA project, we provide a whole-exome sequencing (WES) dataset with molecular biomarker information classified under the 2021 WHO Classification of Tumors of the Central Nervous

<sup>1</sup>Beijing Neurosurgical Institute, Capital Medical University, 100070, Beijing, China. <sup>2</sup>Department of Neurosurgery, The Second Affiliated Hospital of Zhejiang University School of Medicine, Hangzhou, 310009, China. <sup>3</sup>Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, 100070, Beijing, China. <sup>4</sup>Department of Radiotherapy, Beijing Tiantan Hospital, Capital Medical University, 100070, Beijing, China. <sup>5</sup>Department of Neurosurgery, Beijing Children's Hospital, Capital Medical University, National Center for Children's Health, 100045, Beijing, China. <sup>6</sup>Center of Brain Tumor, Beijing Institute for Brain Disorders, 100069, Beijing, China. <sup>7</sup>China National Clinical Research Center for Neurological Diseases, 100070, Beijing, China. <sup>8</sup>Research Unit of Accurate Diagnosis, Treatment, and Translational Medicine of Brain Tumors, Chinese Academy of Medical Sciences, 100070, Beijing, China. <sup>9</sup>These authors contributed equally: Zheng Zhao, Ke-Nan Zhang, Zhiyan Sun, Changlin Yang. ✉e-mail: [mawenping@bjmu.edu.cn](mailto:mawenping@bjmu.edu.cn); [wufan0510284@163.com](mailto:wufan0510284@163.com); [taojiang1964@163.com](mailto:taojiang1964@163.com)

System<sup>11</sup>, which depicts the genomic landscape of DGs under the new classification. Encouragingly, this dataset is the largest public WES dataset from the Chinese DG cohort.

In this work, the WES libraries were sequenced using the Illumina HiSeq platform. Approximately 43.4 billion 150-bp paired-end reads were generated, with an average of over 75.8 million sequence reads per sample. For each sample, we first aligned the raw reads to the reference human genome. Then, we developed a computational pipeline to identify glioma-associated somatic mutations to catalogue the genomic mutational profiles of the cohort of 286 DG specimens. The data are expected to have many utilities, ranging from depicting the molecular characteristics of subtypes, exploring novel biomarkers, identifying prognostic signatures, and analyzing treatment-related variations. Furthermore, the data represent the largest number of diffuse glioma samples so far under the 2021 classification. These data are therefore a significant addition to global DG genome sequence databases and can be used with the new classification.

## Methods

**Specimen collection.** A total of 286 glioma tissues and paired peripheral blood specimens were collected from Beijing Tiantan Hospital, Beijing Puren Hospital and Sanbo Hospital in Beijing. All patients were diagnosed with diffuse glioma by consensus, according to multiple pathological reviews by independent board-certified neuropathologists and further graded under the 2021 WHO classification.

All research was approved by the Tiantan Hospital Institutional Review Board (IRB) and performed under IRB KY2013-017-01. Written informed consent was obtained from all patients in accordance with the requirements of Beijing Tiantan Hospital Ethics Committee, and the principles of the Helsinki Declaration were carefully followed.

The specimens were frozen in liquid nitrogen within 5 min of resection. Follow-up information for each patient was also collected, including general information, survival status, clinical therapy, neuropathological classification and the requisite molecular information (Supplementary Table 1).

**Whole-exome sequencing.** Genomic DNA from tumor tissue and matched blood specimens was extracted and confirmed to have high integrity by 1% agarose gel electrophoresis. The DNA concentration was measured by a Qubit® DNA Assay Kit in Qubit® 2.0 Fluorometer (Invitrogen, USA). A total amount of 0.6 µg genomic DNA per sample was used as input material for the DNA sample preparation. Sequencing libraries were generated using an Agilent SureSelect Human All Exon V6 kit (Agilent Technologies, CA, USA) following the manufacturer's recommendations and index codes were added to each sample. The clustering of the index-coded samples was performed on a cBot Cluster Generation System using a HiSeq PE Cluster Kit (Illumina, USA) according to the manufacturer's instructions. After cluster generation, the DNA libraries were sequenced on an Illumina HiSeq platform and 150 bp paired-end reads were generated.

**Mapping and mutation calling.** Whole-exome sequencing data were mapped to the hg19 genome by applying BWA software (version 0.7.12-r1039, bwa mem)<sup>12</sup> with default parameters. We used SAMtools (version 1.2)<sup>13</sup> (<http://broadinstitute.github.io/picard/>) to sort the reads by coordinates and applied Picard (version 2.0.1, Broad Institute) to mark duplicates for further analysis. An Empirical Bayesian-based tool - SAVI2 was applied to somatic mutations calling (including SNVs and short insertion/deletions) as previously described<sup>14,15</sup>. In this pipeline, SAMtools mpileup and bcftools were used to find variants, then the preliminary variants was further filtered out if the following criteria were met: (1) insufficient sequencing depth; (2) positions with only low-quality reads; (3) positions biased towards either strand. In particular, mutations were selected if the mutation allele frequency in the tumors was significantly higher than that in normal controls. Additionally, we used the CNVkit<sup>16</sup> software to detect copy number changes. The entire flow of the processing data is illustrated in Fig. 1.

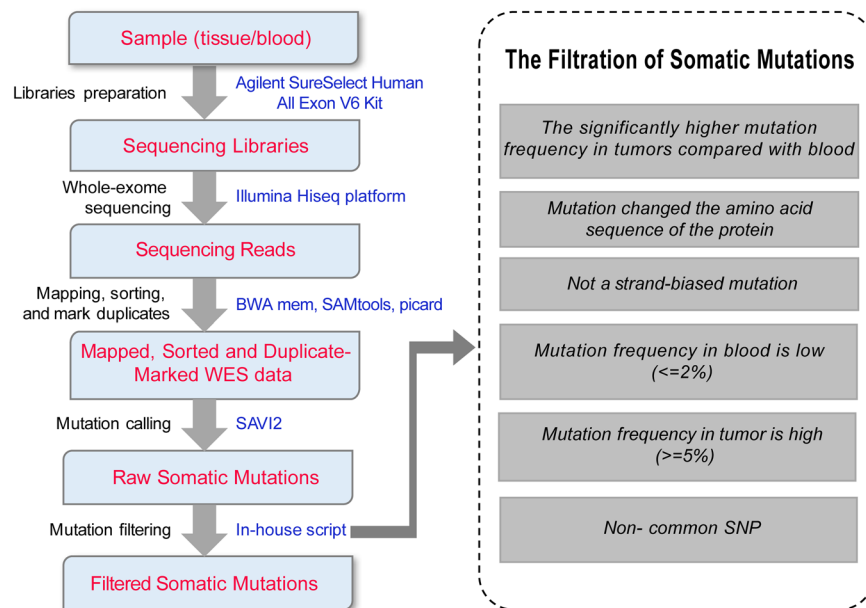
**Molecular pathological character supplementation.** The *IDH1* and *IDH2* (*IDH*) mutation status was collected from pathological records and examined by pyrosequencing or immunohistochemistry with anti-IDH1 R132H antibody. *TERT* promoter (*TERTp*) mutation information was collected from pathological records examined by pyrosequencing. Due to the early collection of patients, some patients lacked *TERTp* mutation information.

*EGFR* amplification, *CDKN2A/B* homozygous deletion, chromosomal 1p/19q codeletion, chromosomal 7 amplification and chromosomal 10 deletion were calculated by CNVkit<sup>16</sup> and manually checked in Integrative Genome Viewer by two independent molecular neuropathologist.

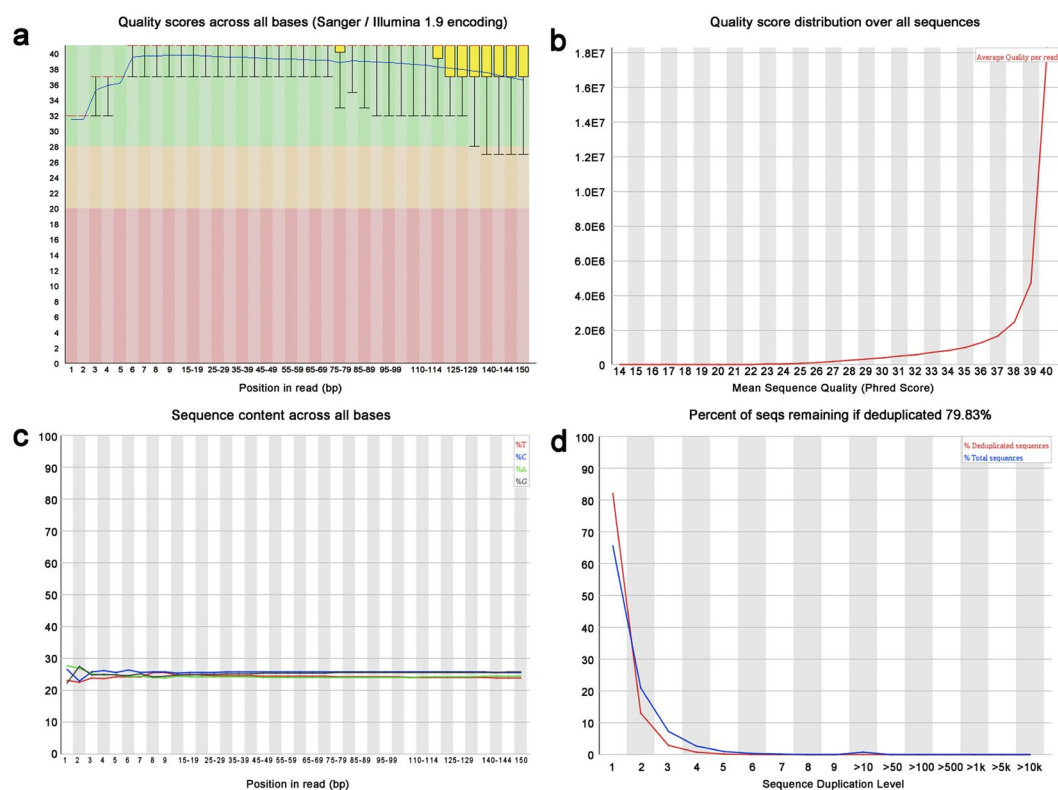
**Relevance to TCGA and Rembrandt cohort.** The Cancer Genome Atlas (TCGA) includes 916 cases of whole exome sequencing data from 401 GBMs and 515 lower-grade gliomas (LGGs), and the Rembrandt cohort includes 263 cases of SNP array data from gliomas. However, Asian patients accounted for less than 5 percent. Moreover, due to the lack of subsequent updates, the patients included could not be classified by the new classification and could not be used in the novel research system.

For our dataset, 286 Chinese glioma patients with whole exome sequencing data were included, filling the gap between these two cohorts. Our dataset could be used as an independent validation dataset for comparative analysis of TCGA, calling more focal copy number changes and covering more genetic mutations than the Rembrandt cohort.

More importantly, relying on the CGGA project, our dataset has included many newly reported molecular pathological biomarkers, and the patients can be classified by the 2021 WHO classification, providing crucial materials for global DG researchers.



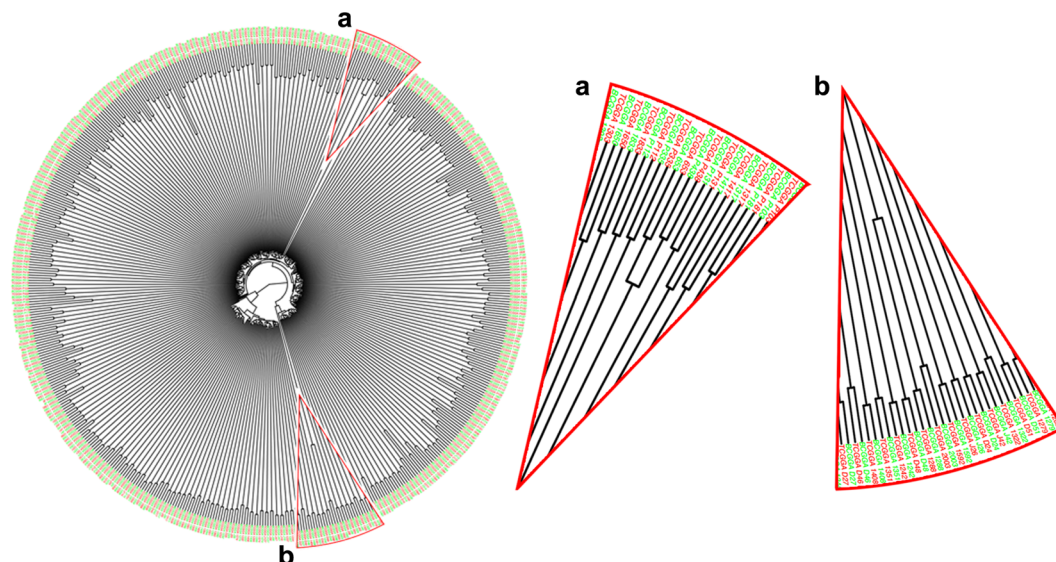
**Fig. 1** Data processing scheme. Data processing scheme from specimen preparation to mutation calling.



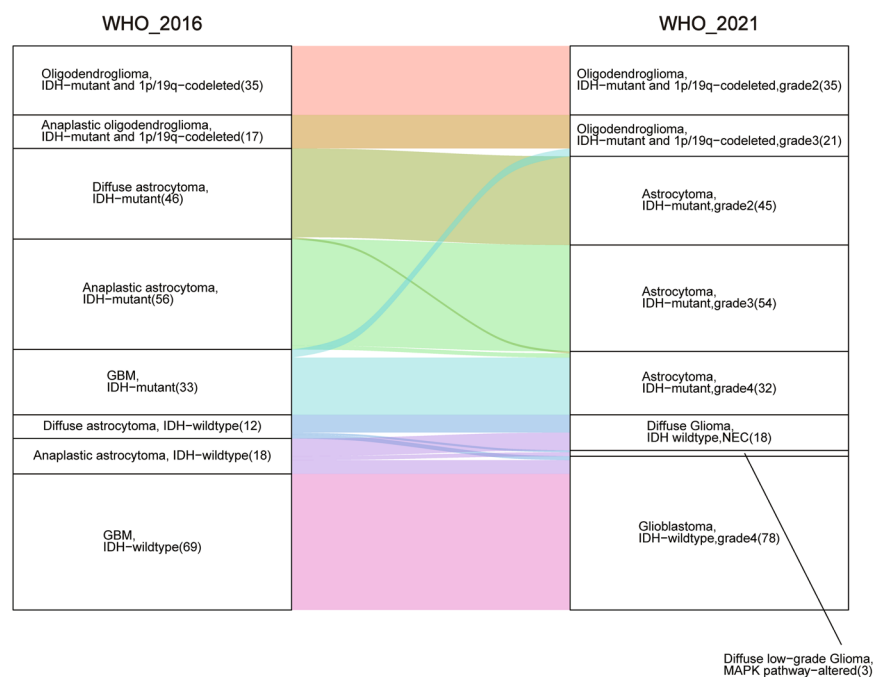
**Fig. 2** Quality control of the sequencing data (CGGA\_653, forward reads). (a) Sequence quality scores across all bases, (b) quality score distribution over all sequences, (c) sequence content across all bases, and (d) sequence duplication levels.

### Data Records

For these 286 cases, the whole-exome sequencing (WES) data of 572 tissue and blood specimens in paired FASTQ files produced by the Illumina HiSeq platform have been deposited in the National Genomics Data Center (NGDC) under accession number HRA000071<sup>17</sup>. The clinical characteristics are summarized in Supplementary Table 1, including histology, WHO grade, critical molecular genetic information, sex, age at diagnosis, overall survival and current status, chemoradiotherapy, etc. Detailed clinical and molecular pathologic

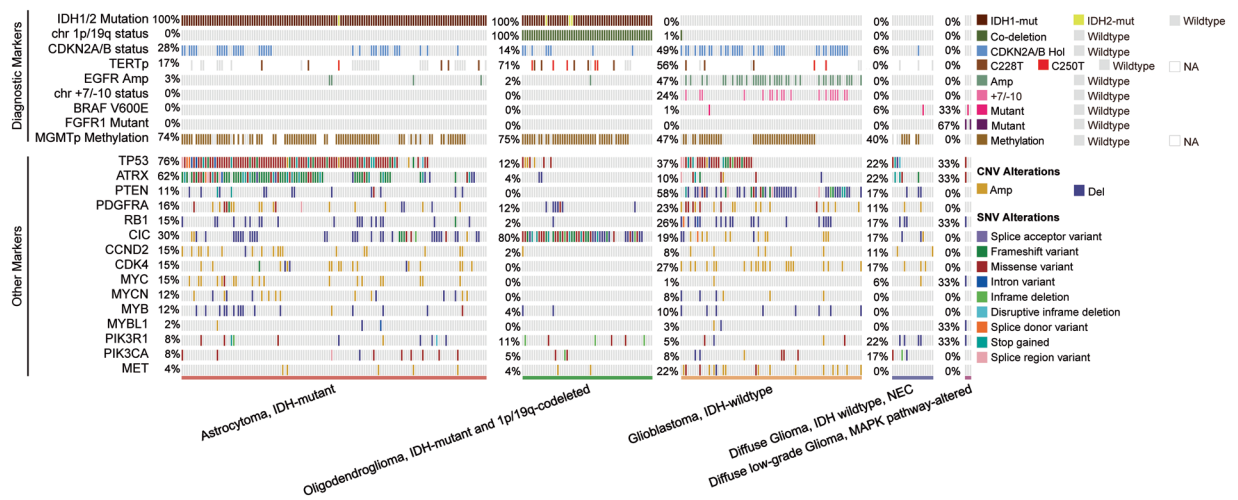


**Fig. 3** Fingerprint analysis of WES samples. Dendrogram of hierarchical clustering of 286 tumors and 286 normal samples based on Pearson correlation coefficients of SNP allele fractions. Case\_ID and the tissue specimen are indicated (blood: green; tumor: red). The enlarged local views show the primary and recurrent samples from the same patient, a) CGGA\_653 and CGGA\_P438, b) CGGA\_1288 and CGGA\_2003, respectively. The correct matching of each of the 286 tumor-blood DNA pairs was determined.



**Fig. 4** Sankey diagram of patients classified under the WHO 2021 classification. The Sankey diagram showed the changes in the WHO classification of the included patients.

information is also deposited at figshare<sup>18</sup>. All the called SNVs per sample made by SAVI2 were deposited at figshare<sup>19</sup>. To classify the diffuse gliomas, IDH-wildtype NEC, IlluminaInfinium Methylation EPICBeadChip (Illumina, USA) was used to test DNA methylation information. The raw methylation data of the five samples was deposited in the GEO Expression Omnibus under the accession number of GSE216383<sup>20</sup>.



**Fig. 5** Oncoprint of collected diffuse gliomas. Characteristic mutations and copy number variations commonly occurred in specific classes.

## Technical Validation

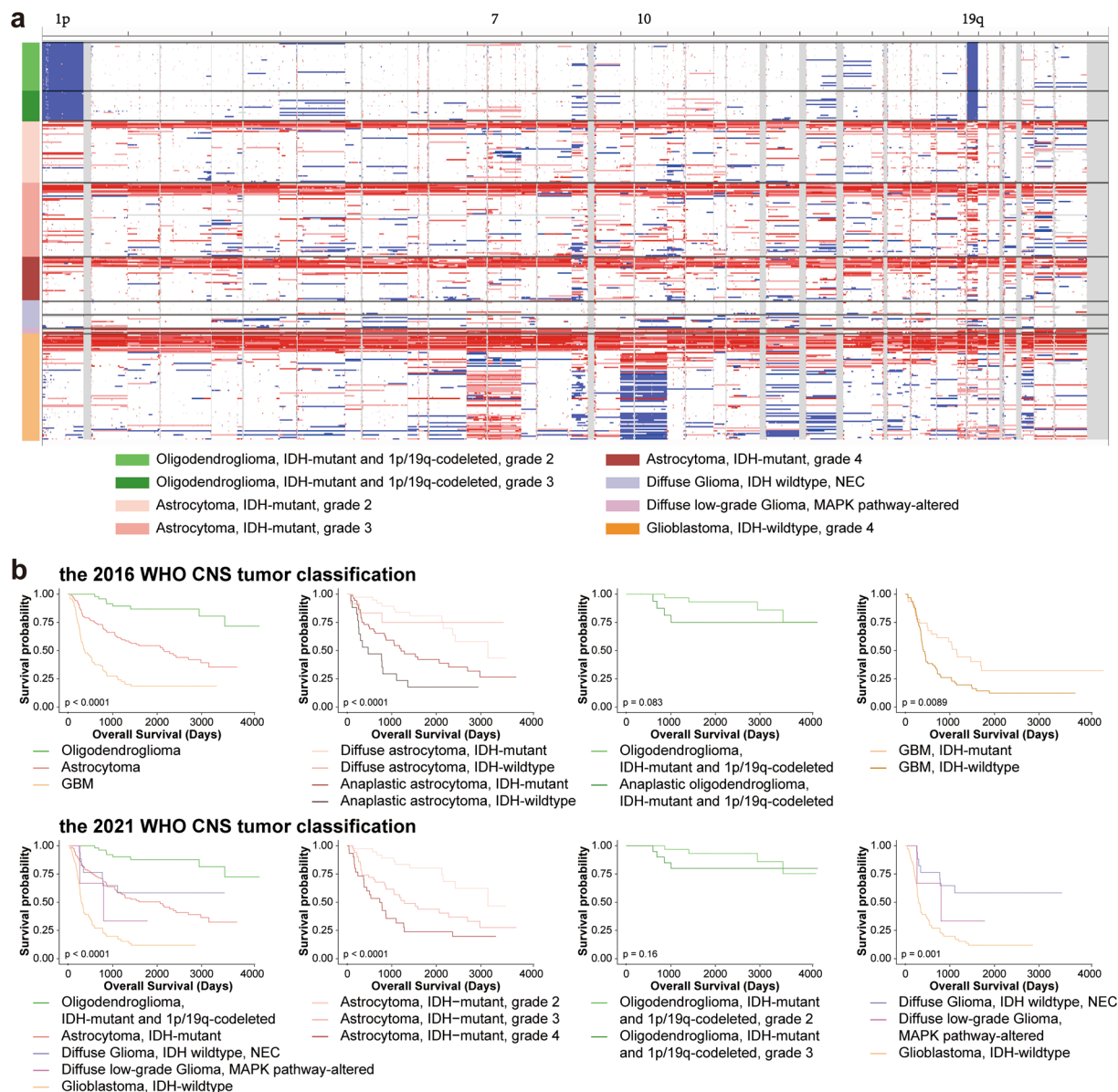
**Quality validation – sequencing data.** We used FASTQC (version 0.10.1) to analyze the data quality via several measures in the FASTQ files, including a) sequence quality scores across all bases, b) quality score distribution over all sequences, c) sequence content across all bases, and d) sequence duplication levels. We selected sample CGGA\_653 as a representative sample. A representative summary plot is provided in Fig. 2. Quality scores per base above 25 are considered base calls indicating high sequencing quality (Fig. 2a). The quality score distribution over all sequences was analyzed to check if a subset of sequences had universally poor quality. As a result, the average quality for most sequences was high, with scores over 30, which indicated that a significant proportion of the sequences in each run had overall high quality (Fig. 2b). Per Base Sequence Content plots out the proportion of each base position for which each of the four normal DNA bases was called. As expected, there was little difference between the different bases of each sequence run, suggesting that the library did not produce a biased sequence composition (Fig. 2c). By examining the sequence bias during polymerase chain reaction (PCR) amplification, we found that less than 5% of sequences were shown over 5 times, demonstrating low sequence duplication in the WES data (Fig. 2d).

**Fingerprint analysis of the WES samples.** Whole exome sequencing (WES) was performed on 286 glioma samples and matched blood DNA. SAVI2 was used to call germline and somatic single nucleotide variants (SNVs) and small insertions and deletions (indels) as previously described<sup>14,15</sup>. All the called SNVs per sample were deposited at Figshare<sup>19</sup>. To confirm the tumor and blood specimens from the same individual, we established a dendrogram of hierarchical clustering of tumor and blood based on known common SNP sites. Here, a common SNP is one that has at least one 1000 Genomes population with a minor allele of frequency  $\geq 1\%$  and for which 2 or more founders contribute to that minor allele frequency. As a result, the clustering subtree of tissue and blood samples for each individual confirmed the matching of normal blood and glioma tissue for each paired specimen (Fig. 3). In particular, since CGGA\_653 (primary oligodendroglioma, CNS WHO grade 2) and CGGA\_P438 (recurrent oligodendroglioma, CNS WHO grade 3) are primary and recurrent specimens from the same patient, their tumor specimens and corresponding blood specimens are clustered under a subtree. A similar phenomenon was observed between CGGA\_1288 (primary astrocytoma, IDH-mutant, CNS WHO grade 4) and CGGA\_2003 (recurrent astrocytoma, IDH-mutant, CNS WHO grade 4). As expected, the tumor and blood specimens from each patient were clustered together and could be used for downstream analysis.

## Usage Notes

WES has been widely used in studying genetic variations of cancers for decades. It is a powerful method to systematically depict the characteristics of tumors and discover novel diagnostic, prognostic and therapeutic biomarkers.

One major advantage of the data is that detailed molecular pathological characteristics have been provided and patients can be classified under the 2021 WHO Classification of Tumors of the Central Nervous System. We summarize the changes of subtype classification of patients in Fig. 4. For instance, the landscape of genetic variations of the newly classified diffuse gliomas is shown in Fig. 5. None of the “Diffuse glioma, IDH-wildtype (NEC)” cases had a TERT promoter mutation (coverage at C228T and C250T), EGFR amplification, or chromosome 7/10 alteration. Consistent with previous studies, astrocytic drivers showed frequent mutations in astrocytoma, e.g., *TP53* and *ATRX*. Correspondingly, loss of function of *PTEN*, *CDKN2A/B* deletion and *EGFR* amplification occurred commonly in glioblastomas. Additionally, the CNV plot and survival curve of different tumor types was shown in Fig. 6.



**Fig. 6** CNV plot and survival curve of different tumor types. **(a)** CNV plot of patients in different molecular subtypes. **(b)** Survival curve of patients in different molecular subtypes.

We have to acknowledge the limitation of incomplete TERTp mutation information of all patients. And relying on the CGGA project, we will continue to collect diffuse glioma samples, extend the sequencing data and supplement the molecular pathological characteristics, keeping them up to date. Additionally, we will continue updating the clinical information regularly. This clinical and molecular pathological information will help promote the further development of classification standards.

### Code availability

All data analyses were completed on the Linux system using standard bioinformatic tools. The codes of the main procedures are described below.

#### (1) Sequencing data mapping

```
bwa mem -t 16 <ref_dir>/hg19.fa <sample_blood>_1.fq.gz <sample_blood>_2.fq.gz | samtools view -@ 16 -Shb -o <sample_blood>.bam
samtools sort -@ 16 <sample_blood>.bam <sample_blood>.sorted
picard MarkDuplicates INPUT = <sample_blood>.sorted.bam OUTPUT = <sample_blood>.sorted.dedup.bam METRICS_FILE = <sample_blood>.metrics.txt
picard BuildBamIndex I = <sample_blood>.sorted.dedup.bam O = <sample_blood>.sorted.dedup.bam.bai
```

#### (2) Coordinate sorting and duplicate marking

```

bwa mem -t 16 <ref_dir>/hg19.fa<sample_tissue>_1.fq.gz<sample_tissue>_2.fq.gz | samtools view -@ 16
-Shb -o<sample_tissue>.bam
samtools sort -@ 16<sample_tissue>.bam<sample_tissue>.sorted
picard MarkDuplicates INPUT = <sample_tissue>.sorted.bam OUTPUT = <sample_tissue>.sorted.dedup.
bam METRICS_FILE = <sample_tissue>.metrics.txt
picard BuildBamIndex I = <sample_tissue>.sorted.dedup.bam O = <sample_tissue>.sorted.dedup.bam.bai
(3) Identification of somatic mutations
<savi_dir>/savi.py --memory 16 --superverbose
--bams <sample_blood>.sorted.dedup.bam, <sample_tissue>.sorted.dedup.bam --names NORMAL,TUMOR
--ref <ref_dir>/hg19.fa --outputdir <sample_dir>

```

Received: 21 April 2022; Accepted: 2 November 2022;

Published online: 11 November 2022

## References

1. Miller, K. D. *et al.* Cancer statistics for adolescents and young adults, 2020. *Ca-Cancer J Clin* **70**, 443–459, <https://doi.org/10.3322/caac.21637> (2020).
2. Jiang, T. *et al.* Clinical practice guidelines for the management of adult diffuse gliomas. *Cancer Lett* **499**, 60–72, <https://doi.org/10.1016/j.canlet.2020.10.050> (2021).
3. Gritsch, S., Batchelor, T. T. & Gonzalez Castro, L. N. Diagnostic, therapeutic, and prognostic implications of the 2021 World Health Organization classification of tumors of the central nervous system. *Cancer* **128**, 47–58, <https://doi.org/10.1002/cncr.33918> (2022).
4. Ostrom, Q. T., Cioffi, G., Waite, K., Kruchko, C. & Barnholtz-Sloan, J. S. CBTRUS Statistical Report: Primary Brain and Other Central Nervous System Tumors Diagnosed in the United States in 2014–2018. *Neuro Oncol* **23**, iii1–iii105, <https://doi.org/10.1093/neuonc/noab200> (2021).
5. Louis, D. N. *et al.* The 2021 WHO Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol* **23**, 1231–1251, <https://doi.org/10.1093/neuonc/noab106> (2021).
6. Brat, D. J. *et al.* Molecular Biomarker Testing for the Diagnosis of Diffuse Gliomas. *Arch Pathol Lab Med* **146**, 547–574, <https://doi.org/10.5858/arpa.2021-0295-CP> (2022).
7. Brennan, C. W. *et al.* The somatic genomic landscape of glioblastoma. *Cell* **155**, 462–477, <https://doi.org/10.1016/j.cell.2013.09.034> (2013).
8. Cancer Genome Atlas Research, N. *et al.* Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med* **372**, 2481–2498, <https://doi.org/10.1056/NEJMoa1402121> (2015).
9. Gusev, Y. *et al.* The REMBRANDT study, a large collection of genomic data from brain cancer patients. *Sci Data* **5**, 180158, <https://doi.org/10.1038/sdata.2018.158> (2018).
10. Zhao, Z. *et al.* Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients. *Genomics Proteomics Bioinformatics* **19**, 1–12, <https://doi.org/10.1016/j.gpb.2020.10.005> (2021).
11. Varlet, P., Nielsen, G., Righi, A., Tanaka, S. & Tirabosco, R. (WHO: Geneva, Switzerland, 2021).
12. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, <https://doi.org/10.1093/bioinformatics/btp324> (2009).
13. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079, <https://doi.org/10.1093/bioinformatics/btp352> (2009).
14. Wang, J. *et al.* Clonal evolution of glioblastoma under therapy. *Nature genetics* **48**, 768–776, <https://doi.org/10.1038/ng.3590> (2016).
15. Hu, H. *et al.* Mutational Landscape of Secondary Glioblastoma Guides MET-Targeted Trial in Brain Tumor. *Cell* **175**, 1665–1678, <https://doi.org/10.1016/j.cell.2018.09.038> (2018).
16. Talevich, E., Shain, A. H., Botton, T. & Bastian, B. C. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol* **12**, e1004873, <https://doi.org/10.1371/journal.pcbi.1004873> (2016).
17. *Genome Sequence Archive PRJCA006139* <https://ngdc.cncb.ac.cn/bioproject/browse/PRJCA006139> (2021).
18. Zhao, Z. Z. *et al.* WES data from 286 diffuse gliomas under the 2021 WHO Classification of Tumors of the Central Nervous System. *figshare* <https://doi.org/10.6084/m9.figshare.21388383.v1> (2022).
19. Zhao, Z. Z. *et al.* WES data from 286 diffuse gliomas under the 2021 WHO Classification of Tumors of the Central Nervous System. *figshare* <https://doi.org/10.6084/m9.figshare.21394908.v1> (2022).
20. *Gene Expression Omnibus GSE216383*. <https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE216383> (2022).

## Acknowledgements

All authors would like to thank all of the patients enrolled. This work was supported by the National Natural Science Foundation of China (NSFC) funds (Nos. 81761168038, 81903078 and 82002647) and the Brain Tumor Precision Diagnosis and Treatment and Translational Medicine Innovation Unit of Chinese Academy of Medical Sciences (2019-I2M-5-021).

## Author contributions

T.J., F.W. and W.M. conceived the research; Z.Z., K.Z., Z.S. and C.Y. designed the research; Q.W., G.L., Zhiliang Wang, F.Z., Z.Q., Zheng Wang and Y.L. collected clinical specimens and information; Z.Z. and K.Z. performed the data processing; Z.S. and R.C. performed the WES data analysis; K.Z. and Z.S. reviewed the process of data processing and analysis; Z.Z., K.Z. and Z.S. wrote the manuscript; all authors reviewed and approved the manuscript.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41597-022-01823-3>.

**Correspondence** and requests for materials should be addressed to W.M., F.W. or T.J.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022