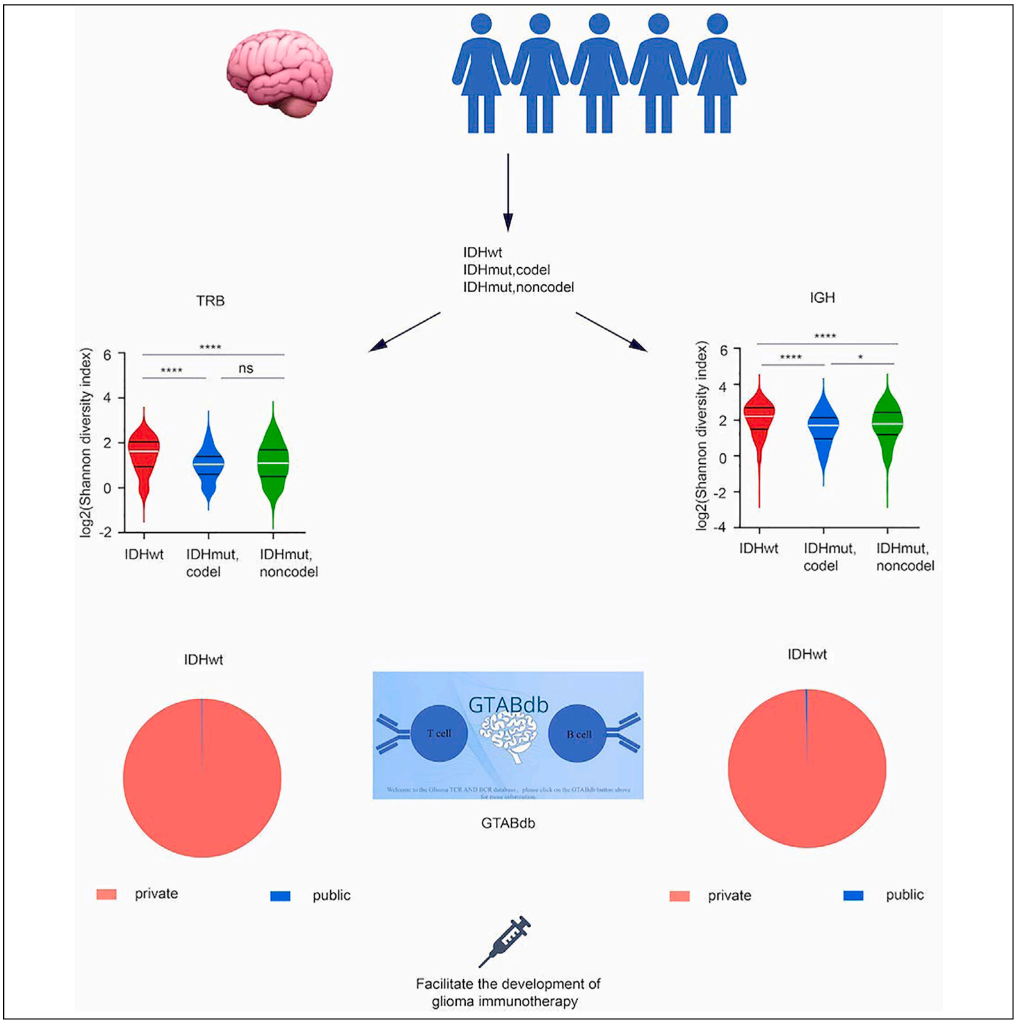
iScience

CellPress
OPEN ACCESS

Article

Comprehensive characterization and database construction of immune repertoire in the largest Chinese glioma cohort



Welcome to the Glioma TCR AND BCR database, please click on the GTABdb button above for more information.

Lu Wang, Zhiyuan Xu, Wei Zhang, Lin Li, Xiao Liu, Jing Zhang

jz2716@126.com

Highlights

Immune repertoire diversity is prominent in IDH wild type glioma

High heterogeneity of immune repertoire within each glioma subtype

Private clonotypes have more hydrophobic residues in CDR3 motifs than public ones

GTABdb, a web-based database designed for characterizing glioma immune repertoire

Wang et al., iScience *27*, 108661
January 19, 2024 © 2023 The Author(s).
https://doi.org/10.1016/j.isci.2023.108661

Article

Comprehensive characterization and database construction of immune repertoire in the largest Chinese glioma cohort

Lu Wang,[1,4] Zhiyuan Xu,[1,4] Wei Zhang,[2,3,4] Lin Li,[1] Xiao Liu,[1] and Jing Zhang[1,5,*]

SUMMARY

Immune receptor repertoire is valuable for developing immunotherapeutic interventions, but remains poorly understood across glioma subtypes including IDH wild type, IDH mutation without 1p/19q codeletion (IDHmut-noncodel) and IDH mutation with 1p/19q codeletion (IDHmut-codel). We assembled over 320,000 TCR/BCR clonotypes from the largest glioma cohort of 913 RNA sequencing samples in the Chinese population, finding that immune repertoire diversity was more prominent in the IDH wild type (the most aggressive glioma). Fewer clonotypes were shared within each glioma subtype, indicating high heterogeneity of the immune repertoire. The TRA-CDR3 was longer in private than in public clonotypes in IDH wild type. CDR3 variable motifs had higher proportions of hydrophobic residues in private than in public clonotypes, suggesting private CDR3 sequences have greater potential for tumor antigen recognition. Finally, we developed GTABdb, a web-based database designed for hosting, exploring, visualizing, and analyzing glioma immune repertoire. Our study will facilitate developing glioma immunotherapy.

INTRODUCTION

Adult diffuse gliomas are the most common brain tumors of the central nervous system, consisting of a heterogeneous group of tumors with poor prognosis and resistance to surgical and chemoradiotherapy regimens.[1–3] According to the newly updated Central Nervous System classification by the World Health Organization, adult diffuse gliomas are classified into three molecular subtypes based on mutation in the isocitrate dehydrogenase 1 and 2 genes (IDH1/2) and the co-deletion states of chromosomal arms 1p and 19q, which includes IDH wild type, IDH mutation without 1p/19q codeletion (IDHmut-noncodel) and IDH mutation with 1p/19q codeletion (IDHmut-codel) gliomas. Management of disease is being redefined in the settings of molecular subtypes of gliomas, which exhibits different histopathology, genetics, prognosis, and therapy responses. But most gliomas recur and lead to limited overall survival. Therefore, patients with gliomas need new treatment modalities.[2,4]

Immunotherapy has revolutionized glioma treatment by utilizing T cells genetically engineered to express tumor recognizing receptors or neoantigen-induced T cell immunity.[5,6] T cell receptor (TCR) is composed of heterodimers of α chains and β chains, which produce highly diverse TCR through V(D)J genes rearrangement. Complementarity determining region 3 (CDR3) is the highly variable region of TCR and is responsible for recognizing tumor neoantigens.[7–9] Studies also show that B cells play an important role in the adaptive immune system, and B cells can produce antibodies and recognize their specific antigens through immunoglobulins, surface antigen receptors.[10–13] Infiltrative B cells are often found present in multiple tumor tissues.[14–16] Characterization of TCR and BCR repertoire has important implications for the development of T- and B-cell based immunotherapy.[17–19]

Recent studies examined the immune repertoire and immunotherapy-related changes through TCR sequencing (TCR-seq) upon a small number of gliomas. BCR sequencing (BCR-seq) in gliomas are few. Profound TCR repertoire dynamics are observed during glioma infiltrating lymphocyte expansion and the expanding capabilities are determined by transcriptional T cell states.[20] The VJ-independent components of tumor-associated repertoires diverge more from their corresponding peripheral repertoires than T cell populations in nonneoplastic brain tissue, particularly for low-grade gliomas.[21] T cell infiltration increased in the tumors of GBM with an early activated and clonally expanded $CD8^+$ T cell cluster whose TCR overlaps with a $CD8^+$ PBMC population observed after anti-PD-1 therapy, although macrophages and monocytes still constitute the majority of infiltrating immune cells.[22] An IDH1(R132H)-specific peptide vaccine (IDH1-vac) are found able to induce

[1]Key Laboratory for Biomechanics and Mechanobiology of Ministry of Education, Beijing Advanced Innovation Centre for Biomedical Engineering, School of Engineering Medicine & School of Biological Science and Medical Engineering, Beihang University, Beijing 100191, China
[2]Department of Molecular Neuropathology, Beijing Neurosurgical Institute, Capital Medical University, Beijing 100070, People's Republic of China
[3]Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, No. 119 South Fourth Ring Road West, Fengtai District, Beijing 100070, People's Republic of China
[4]These authors contributed equally
[5]Lead contact
*Correspondence: jz2716@126.com

https://doi.org/10.1016/j.isci.2023.108661

immune response, but pseudoprogression are also observed being associated with increased vaccine-induced peripheral T cell responses. Tumor-infiltrating CD40LG(+) and CXCL13(+) T helper cell clusters in a patient with pseudoprogression are dominated by a single IDH1(R132H)-reactive T cell receptor.[23] However, these studies were limited to a small number of patients with glioma and the results are yet to be confirmed.

TCR-seq/BCR-seq are of high-cost and infeasible for limited tissue biopsies. In contrast, RNA sequencing (RNA-seq) data also contains expressed TCR and BCR, especially the abundant and clonally expanded receptors. The newly developed TRUST algorithm (version 4)[24] are designed to process the most diverse and critical regions of antigen recognition with good performances and accuracy by assembling longer receptor sequences. Additionally, Chinese Glioma Genome Atlas has generated high-throughput RNA-seq data for the largest Chinese glioma cohort.[25] These together motivated us to examine the immune receptor repertoire and investigated the repertoire difference across IDH wild type, IDHmut-noncodel, and IDHmut-codel gliomas.

To facilitate the characterization of immune repertoire under a variety of clinical settings, we developed GTABdb, a glioma TCR and BCR immune repertoire database. The existing immune repertoire databases such as TCRdb[26] and CDJdb[27] mainly store the receptor sequences without associated clinical information. VisualizIRR[28] integrates immune repertoire from The Cancer Genome Atlas (TCGA) and immunoACCESS, but lack of Chinese population of gliomas. In GTABdb, we processed immune repertoire data from the largest Chinese cohort of gliomas, integrating full clinical information for each sample, which allows users to interactively investigate immune repertoire based on their own interests.

RESULTS

Patient characteristics

In total, 913 patients with glioma included in this study were assigned to three distinct molecular subgroups based on the mutational status of the isocitrate dehydrogenase 1 and 2 genes (IDH1/2), the codeletion status of chromosome arms 1p and 19q (424 for IDH wild type; 177 for IDH mutation with 1p/19q codeletion, thereafter named as IDHmut-codel; and 312 for IDH mutation without 1p/19q codeletion, thereafter referred to as IDHmut-noncodel; Table S1). Of the 913 patients, 374 (41%) and 539 (59%) were female and male, respectively, and the distribution of gender was not significantly different across the three glioma groups ($p = 0.3123$, Chi-square test). The age distribution was significantly different among glioma subgroups ($p < 0.0001$, Kruskal-Wallis rank-sum test; the median age was 49 for IDH wild type, 39 for IDHmut-noncodel, and 41 for IDHmut-codel). IDH wild type gliomas included more WHO IV grade gliomas than IDHmut-noncodel and IDHmut-codel. Kaplan-Meier analysis demonstrated that the overall survival was significantly different (Log rank $p < 0.0001$) (median survival was 407, 1051.5 and 2264.5 days for IDH wild type, IDHmut-noncodel, and IDHmut-codel, respectively) (Table S1 and Figure S1), consistent with the known knowledge that IDH wild type was the most clinically aggressive with the worst survival, followed by IDHmut-noncodel and then IDHmut-codel.

Progressive restriction of T cell receptor immune repertoire correlates with the severity of molecular classes of gliomas

To analyze and compare T cell receptor repertoire diversity across IDH wild type, IDHmut-codel, and IDHmut-noncodel gliomas, we assembled T cell receptors from RNA-seq data of 913 gliomas previously generated by Chinese Glioma Genome Atlas using TRUST algorithm (version 4) with substantial improvements in efficiency and sensitivity over previous version of TRUST.[24] A total of 14,770 clonotypes which were the unique combination of V-gene, D-gene, J-gene and CDR3 amino acid sequence across all TCR chains (*TRA* and *TRB*) were recovered from 913 gliomas.

To study the properties of *TRA* and *TRB* repertoire in distinct molecular classes of gliomas, we examined the frequency of unique V-gene and J-gene pair clonotypes by the graphical representation of repertoire diversity using Voronoi Treemaps. Compared with IDH wild type gliomas, substantial reduction of *TRA* and *TRB* repertoire diversity, associated with clonotypic expansion, was observed in IDHmut-codel gliomas, to a lesser extent, from IDHmut-noncodel gliomas (Figure 1A). Based on clonotypes, we performed quantitative analysis of repertoire diversity and complexity using different commonly-used ecological metrics including the Shannon-Wiener Index, Gini-Simpson index, Chao1, Observed Clonotypes, ACE, and Pielou. IDH wild type gliomas were found to have significantly increased *TRA* and *TRB* repertoire as demonstrated by their higher diversity (Figures 1B—1E) of the clonotypes than IDH mutant gliomas (IDHmut-codel and IDHmut-noncodel) measured by the Shannon-Wiener Index[29] and Gini-Simpson index,[19,29] respectively. Both Chao1 index and ACE index, also indicators for estimating richness using rare clonotypes, confirmed that the richness of clonotypes in *TRA* and *TRB* repertoire of IDHwt gliomas was significantly higher than that of IDH mutant gliomas (Figures 1F–1I). Another richness measure for immune repertoire, the observed clonotypes (the number of *TRA* or *TRB* clonotypes for each sample), also validated that IDH wild type gliomas had more richness than IDH mutant gliomas in *TRA* and *TRB* repertoire (Figures 1J and 1K). Evenness measure, pielou index, indicated that IDH wild type gliomas had less evenness than IDH mutant gliomas in *TRA* and *TRB* repertoire, particularly significantly less than IDHmut-codel gliomas for *TRB* repertoire (Figures 1L and 1M).

Next, we found that there were more *TRA* than *TRB* clonotypes shared among at least two patients (defined as public clonotype) of each glioma subtype (IDH wild type gliomas: 1.21% for *TRA* and 0.23% for *TRB* clonotypes; IDHmut-noncodel: 1.04% for *TRA* and 0.26% for *TRB* clonotypes; IDHmut-codel: 1.01% for *TRA* and 0.25% for *TRB* clonotypes). Most *TRA* and *TRB* clonotypes were unique for each glioma (defined as private clonotype) (IDH wild type gliomas: 98.79% for *TRA* and 99.77% for *TRB* clonotypes; IDHmut-noncodel: 98.96% for *TRA* and 99.74% for *TRB* clonotypes; IDHmut-codel: 98.99% for *TRA* and 99.75% for *TRB* clonotypes) (Figures S2A–S2F). There was a total of 3 *TRA* and 1 *TRB* identical clonotypes shared across the three glioma subtypes, but none of them were able to predict the overall survival of gliomas (Log rank $p > 0.05$) (Table S2). These results revealed that TCR repertoire were of high diversity and individual differences in different glioma molecular subtypes, particularly diversified in IDH wild type gliomas, the one with the worst prognosis.

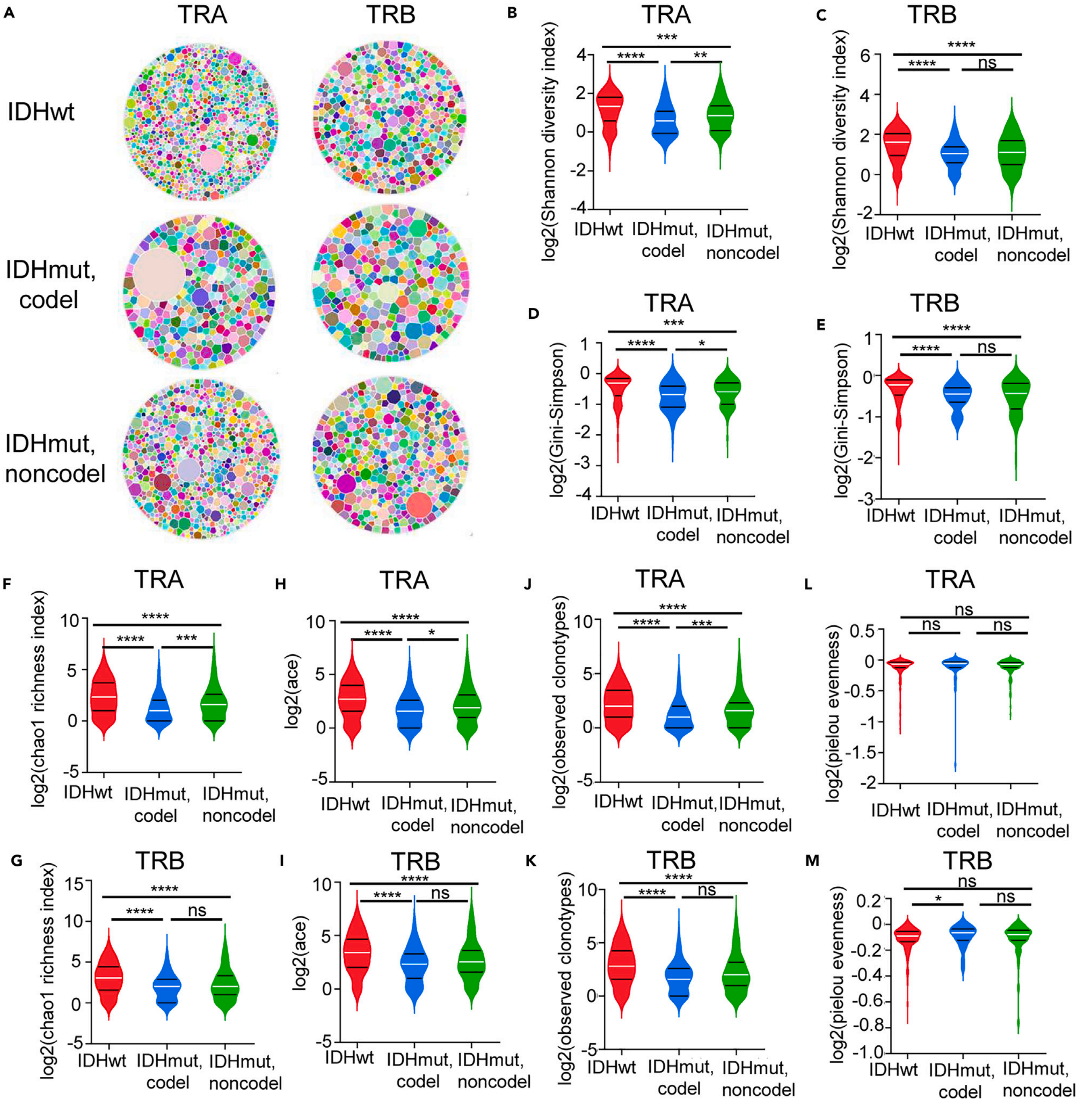


Figure 1. The characteristics of *TRA* and *TRB* repertoire across IDH wild type, IDHmut-noncodel, and IDHmut-codel

(A) Graphical representation of the diversity of *TRA* and *TRB* repertoire from IDH wild type, IDHmut-noncodel, and IDHmut-codel. Each dot represents a unique V-gene and J-gene pair clonotypes, and the size of the dot is the relative frequency of that rearrangement in the entire population.

(B–M) Quantification of the diversity of *TRA* and *TRB* repertoire across IDHwt, IDHmut-noncodel and IDHmut-codel including Shannon diversity index for *TRA* (B) and *TRB* (C) repertoire, Gini-Simpson diversity index for *TRA* (D) and *TRB* (E) repertoire, Chao1 richness index for *TRA* (F) and *TRB* (G), Ace index for *TRA* (H) and *TRB* (I), the number of clonotypes for *TRA* (J) and *TRB* (K), and Pielou evenness index for *TRA* (L) and *TRB* (M). IDHwt is IDH wild type gliomas. IDHmut-noncodel is gliomas with IDH mutation but not 1p/19q codeletion. IDHmut-codel is gliomas with IDH mutation and 1p/19q codeletion. Data are represented as median +/−IQR, where IQR is interquartile range. Significance was determined by Mann-Whitney U test (ns, $p > 0.05$; *, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$; ****, $p \leq 0.0001$).

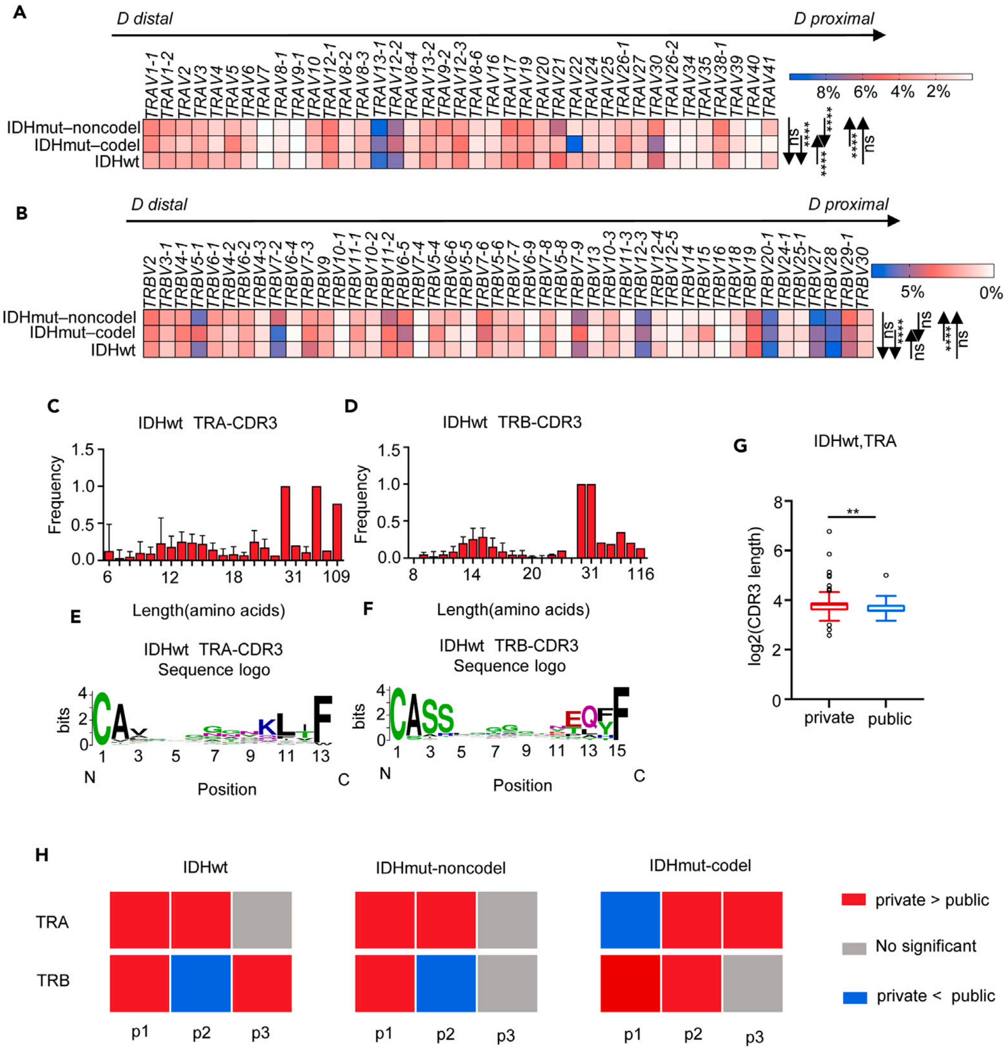


Figure 2. Differential usage of V genes in TRA and TRB repertoire of patients in IDHwt, IDHmut-noncodel and IDHmut-codel

(A and B) Heatmap representing the frequency of V gene usage among unique *TRA* (A) and *TRB* (B) sequences from IDHwt, IDHmut-noncodel and IDHmut-codel. Significance was determined by Chi-square test for goodness of fit.

(C and D) The distribution of the length of unique *TRA*-CDR3 (C) and *TRB*-CDR3 (D) in IDHwt.

(E and F) The amino acid composition of unique *TRA*-CDR3 (E) and *TRB*-CDR3 (F) sequences in IDHwt.

(G) The comparison of the length between private and public TRA-CDR3 sequences. Significance was determined by Mann-Whitney U test.

(H) The comparison of proportion of hydrophobic amino acids in the three middle positions of public and private *TRA*-CDR3 and *TRB*-CDR3 sequences. P1, P2 and P3 are the positions. IDHwt is IDH wild type gliomas. IDHmut-noncodel is gliomas with IDH mutation but not 1p/19q codeletion. IDHmut-codel is gliomas with IDH mutation and 1p/19q codeletion. Data are represented as median +/−IQR. Significance was determined by binomial test and Benjamini–Hochberg correction (ns, $p > 0.05$; *, $p \leq 0.05$; **, $p \leq 0.01$; ***, $p \leq 0.001$; ****, $p \leq 0.0001$).

Skewed usage of V, D, J genes of T cell receptor repertoire in isocitrate dehydrogenase wild type gliomas

To compare the usage of the V-gene, D-gene and J-gene of *TRA* and *TRB* repertoire in different molecular classes of gliomas, we arranged the V(D)J genes in their chromosomal order and performed the chi-square goodness-of-fit test to determine whether the skew usage of the V(D)J genes in the *TRA* and *TRB* repertoire may reflect its topological properties. We noticed that the distribution of V-gene usage in both *TRA* and *TRB* repertoire were significantly different in IDH wild type, IDHmut-noncodel, and IDHmut-codel gliomas (Figures 2A and 2B). J-gene of TRA repertoire exhibited significantly different usage patterns in IDHmut-noncodel from IDH wild type and IDHmut-codel gliomas (Figure S3A). D-gene and J-gene of *TRB* repertoire maintained a similar usage pattern across the three different glioma subtypes (Figure S3B).

Additionally, 12 TRAV, 9 TRAJ genes of *TRA* repertoire and 24 TRBV, 12 TRBJ and 1 TRBD genes of *TRB* repertoire were found having different usage frequencies between IDH wild type and IDH mutant gliomas (IDHmut-noncodel or IDHmut-codel) (Figures S4A–S4H). Altogether, IDH wild type and IDHmut-codel gliomas had different preferences for V-gene and J-gene usage in *TRA* and *TRB* repertoire.

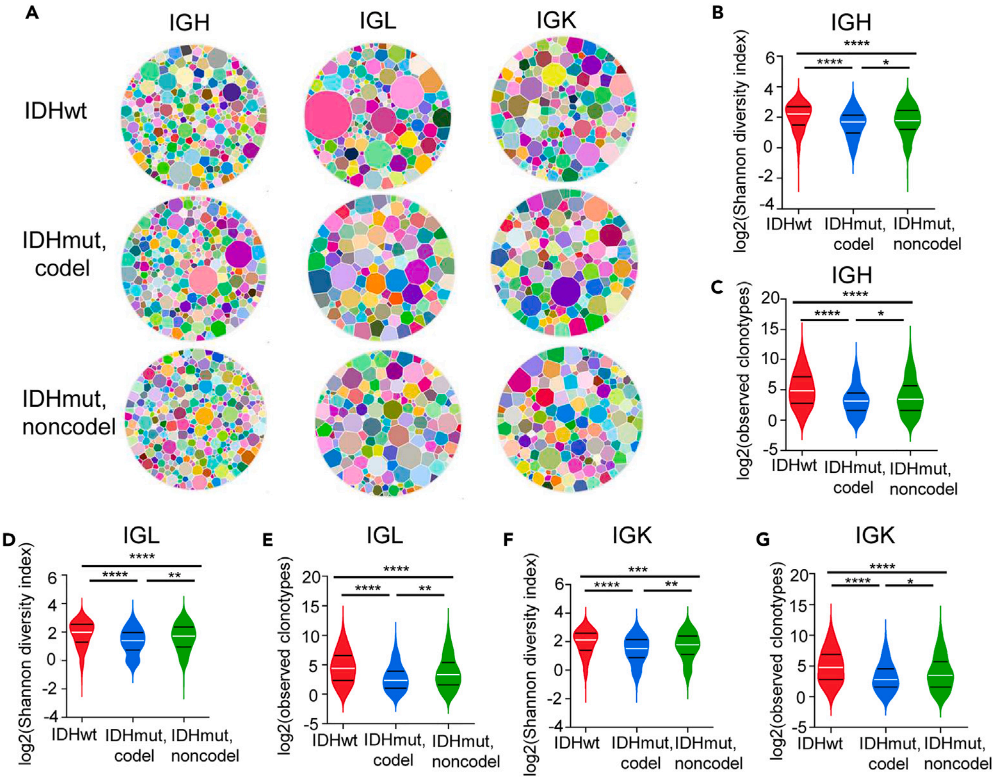


Figure 3. The characteristics of *IGH, IGL* and *IGK* repertoire across IDH wild type, IDHmut-noncodel, and IDHmut-codel

(A) Graphical representation of the diversity of *IGH*, *IGL* and *IGK* repertoire from IDH wild type, IDHmut-noncodel, and IDHmut-codel. Each dot represents a unique V-gene and J-gene pair clonotypes, and the size of the dot is the relative frequency of that rearrangement in the entire population.

(B—G) Quantification of the diversity of *IGH*, *IGL* and *IGK* repertoire across IDHwt, IDHmut-noncodel and IDHmut-codel including Shannon diversity index for *IGH* (B), *IGL* (D), and *IGK* (F), and the number of clonotypes for *IGH* (C), *IGL* (E), and *IGK* (G). IDHwt is IDH wild type gliomas. IDHmut-noncodel is gliomas with IDH mutation but not 1p/19q codeletion. IDHmut-codel is gliomas with IDH mutation and 1p/19q codeletion. Data are represented as median +/—IQR. Significance was determined by Mann-Whitney U test (ns, $p > 0.05$; *, $p \le 0.05$; **, $p \le 0.01$; ***, $p \le 0.001$; ****, $p \le 0.0001$).

The preferences of *TRA-/TRB*-CDR3 length and amino acid composition in isocitrate dehydrogenase wild type gliomas

The distributions of the length of *TRA-/TRB*-CDR3 sequences showed that 11 to 15 amino acid (AA) length of CDR3 sequences having the highest frequency of occurrences in the three glioma subtypes (IDH wild type: 13 AA for *TRA*-CDR3, 15 AA for *TRB*-CDR3; IDHmut-noncodel: 14 AA for *TRA-/TRB*-CDR3; IDHmut-codel: 11 AA for *TRA*-CDR3, 15AA for *TRB*-CDR3) (Figures 2C, 2D, and S5A–S5D). Analysis of the composition of amino acids in the CDR3 sequences with the highest frequencies of occurrences revealed that *TRA-/TRB-* CDR3 sequences were conserved at both ends and variable in the middle segments irrespective of glioma subtypes (Figures 2E, 2F, and S5E–S5H).

β-CDR3 sequence in peripheral blood and tumor infiltrative private T cells was reported significantly longer than the sequence of public T cells.[9,30] Here, we observed that *TRA-/TRB*-CDR3 sequences were longer in private than those in public clonotypes, particularly for *TRA*-CDR3 sequences in IDH wild type gliomas(Figures 2G and S6A–S6E).We then defined the three amino acids in the relative middle of the *TRA-/TRB*-CDR3 sequence as "CDR3 variable motif" as a previous study.[9] In IDH wild type and IDHmut-noncodel gliomas, the first two and the first positions for *TRA-* and *TRB*-CDR3 variable motifs had significantly higher hydrophobic residue proportions in private than in public *TRA* and *TRB* repertoire. The last two positions in *TRA*-variable motifs were of higher proportions of hydrophobic residues in private than in public repertoire in IDHmut-codel (Figures 2H and S6F–S6K). The higher proportions of hydrophobic residues at the three positions of CDR3 variable motifs suggested that private *TRA-/TRB*-CDR3 sequences in gliomas could have greater potential for tumor antigen recognition.

The highly diversified BCR immune repertoire in isocitrate dehydrogenase wild type gliomas

B cell receptor (*IGH*, *IGL*, *IGK*) repertoire are mainly composed of one heavy chain and two light chains. Diversity of the B cell receptor repertoire is key to the protection of an individual's immune system against a variety of potential pathogens.[17,31] Similarly, we also assembled B cell receptors using TRUST algorithm (version 4) and recovered a total of 311,146 clonotypes across BCR chains.

IDH wild type gliomas were found to have significantly increased diversity and richness of *IGH*, *IGL* and *IGK* repertoire compared with IDHmut-noncodel and IDHmut-codel, which were consistently validated by both graphical representation with Voronoi Treemaps (Figure 3A) and different quantitative metrics (Shannon-Wiener Index and Gini-Simpson index for diversity; Chao1 index and ACE index for richness) (Figures 3B–3G and S7A—S7I).The homogeneity of *IGL* and *IGK* other than *IGH* repertoire measured by pielou index was higher in IDHmut-codel than in IDH wild type (p *value* = 0.0023 for IGL, p *value* = 0.0275 for *IGK*, p *value* = 0.8752 for *IGH*) and IDHmut-noncodel gliomas (p *value* = 0.003 for *IGL*, p *value* = 0.0514 for IGK, p *value* = 0.3614 for *IGH*，Mann-Whitney U test) (Figures S7J–S7L)These data demonstrated that IDH wild type gliomas were of highly diversified BCR immune repertoire compared with IDH mutant gliomas.

The usage preferences of V, D, J genes of B cell receptor repertoire in gliomas

The distribution of the usage of V-gene, D-gene and J-genes of *IGH* repertoire (Figures S8A and S8B), and V-gene and J-gene of *IGL* (Figure S8C) and *IGK* repertoire (Figure S8D), exhibited different patterns in IDH wild type compared with IDH mutant gliomas. IDHmut-noncodel gliomas had different usage distribution of V-gene of *IGH* repertoire, V-gene and J-gene of *IGL* (Figures S8A–S8C) and *IGK* (Figure S8D) repertoire from IDHmut-codel gliomas.

In regards to usage frequencies, there were no difference across glioma subtypes for most V(D)J genes of *IGH*, *IGL* and *IGK* repertoire. Only 15 *IGHV* and 5 *IGHD* genes showed significantly different across glioma subtypes with *IGHV4-59*, *IGHD2-8* and *IGHD6-25* having the significantly highest usage frequency in IDHmut-codel gliomas (Figures S9A–S9D). A total of 11 *IGLV* genes had different usage frequencies across glioma subtypes (10 *IGLV* genes between IDH wild type and IDHmut-codel; *IGLV9-49*, *IGLV3-10* and *IGLV3-1* between IDH wild type and IDHmut-noncodel; *IGLV5-45*, *IGLV1-36* and *IGLV2-18* between IDHmut-codel and IDHmut-noncodel) (Figures S9E and S9F). NO *IGLJ* gene usage frequencies were observed different among glioma subtypes (Figure S9G). 13 *IGKV* and 2 *IGKJ* (*IGKJ4* and *IGKJ5*) genes totally exhibited different usage frequencies between IDH wild type and IDHmut-codel, with *IGKV3-15* different between IDH wild type and IDHmut-noncodel gliomas (Figures S9H–S9J).

The preferences of *IGH-/IGL-/IGK*-CDR3 length and amino acid composition

Previous studies have shown that the length of the light chain in the B cell receptor is shorter than that of the heavy chain and the CDR sequences are more conserved.[17,32,33] In glioma, we also observed that *IGH*-CDR3 sequences were on average significantly longer than *IGL*-CDR3 and *IGK*-CDR3 sequences in the B cell receptor repertoire. *IGH*-CDR3 sequences with 16 AA (Figures S10A–S10C), *IGL*-CDR3 with 13 AA (Figures S10D–S10F) and *IGK*-CDR3 with 11 AA (Figures S10G–S10I) were found and 11 AA were found having the highest frequency of occurrences in all the three glioma subtypes. And the two light chains all showed less diversity (Figures S10A–S10I). Analysis of the composition of amino acids in the CDR3 sequences with the highest frequencies of occurrences revealed that *IGH-/IGL-/IGK-* CDR3 sequences were conserved at both ends and variable in the middle segments irrespective of glioma subtypes (Figures S11A–S10I).

Additionally, there was a total of 7 *IGH*, 351 *IGL* and 821 *IGK* identical clonotypes shared across the three glioma subtypes, of which 55 *IGL* clonotypes and 173 *IGK* clonotypes were able to predict the worse overall survival of gliomas (log rank $p < 0.05$) (Table S2). A much higher proportion of private than public clonotypes in B cell receptor repertoire were constantly observed in each glioma subtypes, respectively (Figures S12A–S12I). Differences in the amino acid composition and hydrophobicity of the CDR-H3 regions have potentially important effects on antigen binding. We observed that CDR3 sequences in private *IGH*-clonotypes were significantly longer in IDH wild type gliomas than that in public *IGH*-clonotypes, with no difference between IDHmut-codel and IDHmut-noncodel (Figures S13A–S13C). The length of CDR3 sequences of *IGL* and *IGK* repertoire didn't exhibit significantly different between private and public clonotypes in each glioma subtype (Figures S13D–S13F for *IGL*; Figures S13G–S13I for *IGK*). The proportion of hydrophobic amino acids in at least two positions of the private CDR3 variable motif in the light chain of the B cell receptor repertoire was significantly higher than that in the public CDR3 variable motif (Figures S14A and S14E–S14J). In contrast, the proportion of hydrophobic amino acids in at least two positions of the public CDR3 variable motif in the heavy chain of the B cell receptor repertoire was higher than that in the private CDR3 variable motif (Figures S14A–S14D).

Construction of glioma immune repertoire database for visualization and analysis

We developed GTABdb, a glioma TCR and BCR immune repertoire database specifically designed for hosting, exploring, visualizing and analyzing TCR and BCR immune repertoire of gliomas from the large Chinese population in this study (http://www.oncoimmunobank.cn/glicrdb/database/homepage). GTABdb adopted the Model-View-Controller architecture (Figure 4) to be user-friendly and convenient for parallel development and simplification of updating and integrating new databases. GTABdb stores a variety of clinical information including patient accession ID, age, gender, histology types, WHO grade and overall survival, and full TCR and BCR immune repertoire of gliomas mainly including read count, CDR3 amino acid sequences, CDR3 base sequences and V(D) J gene. GTABdb provides a user-friendly interface to query and download these data. GTABdb also integrates multiple analysis functions including survival curve analysis, CDR3 sequence length analysis of TCR and BCR, frequency analysis of V(D)J rearrangement information, and immune repertoire evaluation metrics analysis.

Home page presents general information about GTABdb, major entry point named as 'GTABdb' to fully access GTABdb and quick links to access 'Clinical information', 'TCR', 'BCR', and 'Online analysis tools'. To help users quickly acquaint themselves with glioma, TCR and BCR, GTABdb provides 'Introduction' page that gives brief descriptions associated with pathogeny and classification systems of glioma, and classification and structures of TCR and BCR. By specifying an item from a drop-down list of clinical information at 'Basic information' page, users can examine the distribution of the selected clinical info across glioma subgroups based on different grouping strategies. GTABdb also supports survival analysis through 'OS' page. By selecting a grouping option from 'Group', users can compare the survival curves across subgroups of glioma and obtain log rank p value generated by Kaplan Meier analysis (Figure 5A). Alternatively, users can also specify multiple groups of patients with glioma, followed by comparing their overall survival and obtaining the significance metrics.

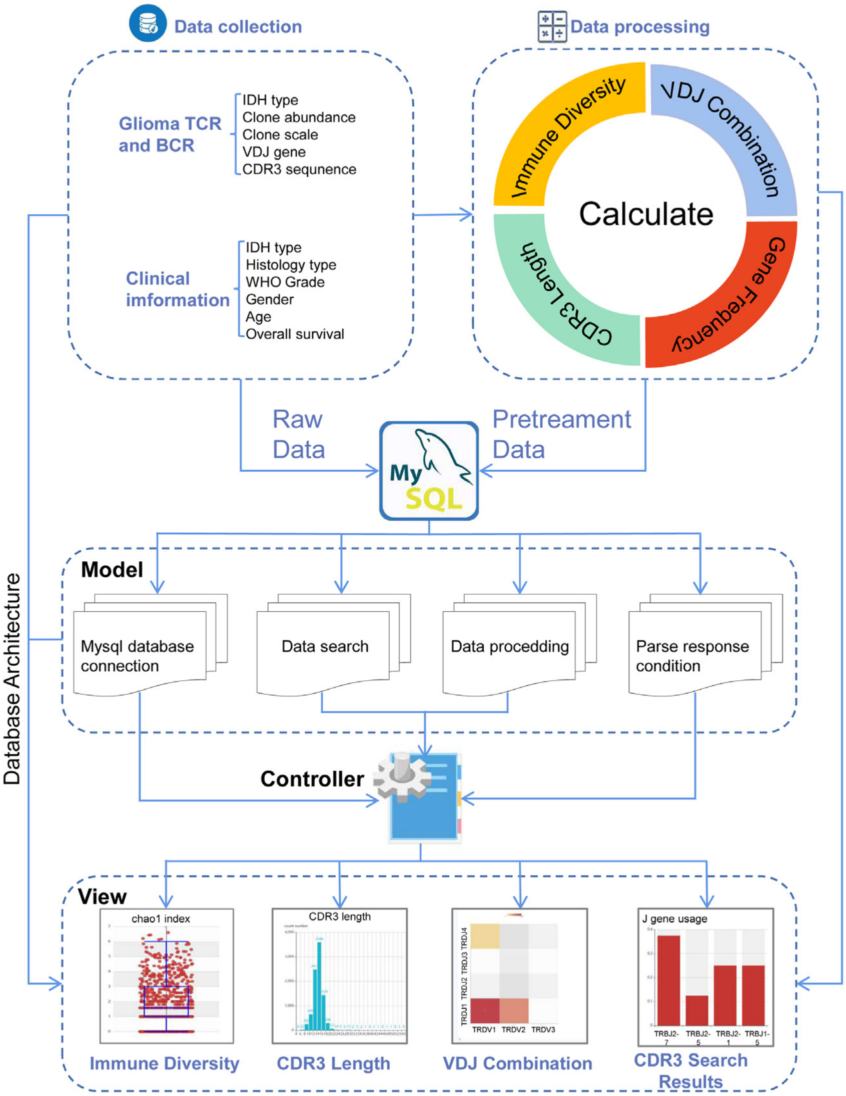
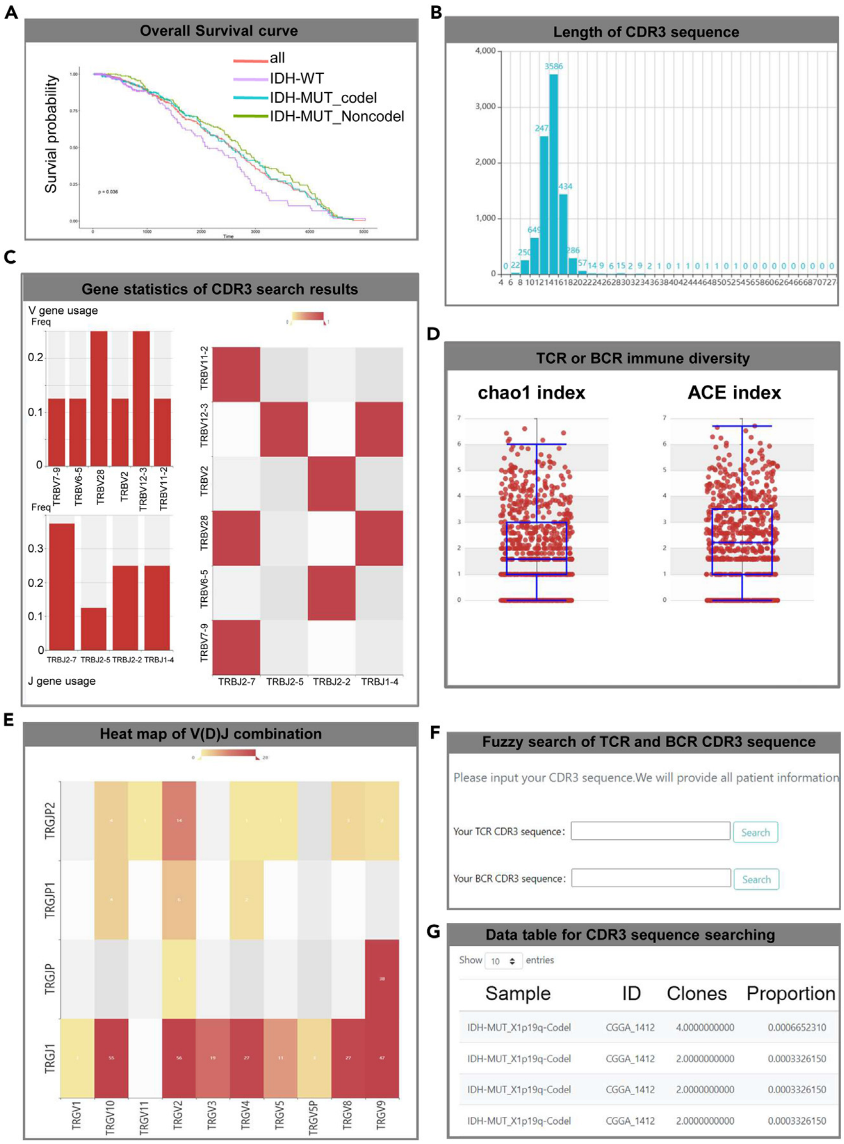


Figure 4. Schematic overview of data collection, data processing and key functionality of GTABdb

GTABdb collects TCR and BCR immune repertoire generated by TRUST4 from 913 RNA-sequencing data of gliomas in CGGA. A variety of methods for quantification of TCR and BCR immune repertoire were calculated for each patient including CDR3 sequence length, immune diversity, VDJ contribution and gene usage frequency. Model-View-Controller architecture was used in GTABdb. The Model component handles data from multiple sources in MySQL databases, calculating the quantitative metrics of immune repertoire, performing statistical analysis and visualization. The View component provides heterogeneous and synchronized views for presenting information and interacting with users, and the front-end template engine of Bootstrap plus HTML and JavaScript provide better visibility and usability of our functionality. The Controller component deals with the application logic, functioning as a mediator between the Model and View components.

GTABdb provides 'Data' where users can fully access to all clinical information and the detailed information about TCR and BCR immune repertoire including Patient accession ID, molecular subtype, Clone, Proportion, complete CDR3 sequence, and the corresponding gene symbols of VDJ genes. To help users investigate the TCR and BCR immune repertoire from patients with glioma, GTABdb provides 'Length' for exploring the length distribution of CDR3 of TCR or BCR (Figure 5B), 'Gene' for probing the frequency of VDJ genes encoding TCR or BCR (Figure 5C), 'Diversity' for analyzing a variety of diversity indexes (Figure 5D), and 'Rearrangement' for examining the combined frequency of VDJ gene rearrangements of CDR3 (Figure 5E). By specifying a subgroup of gliomas based on any established grouping strategy or multiple ones combined, or any patients with glioma they have interests in, users can study the above characteristics of TCR or BCR immune repertoire of the patients with glioma they selected.

To support information search and exploration, GTABdb provides fuzzy search based user-friendly web interface of 'Search' to search for a specific TCR or BCR CDR3 sequence. By inputting CDR3 sequences of TCR or BCR (Figure 5F), users can access to full information associated with this CDR3 sequences including glioma subtype, clones, proportion, and V(D)J genes that contributed to this CDR3 sequences



Please input your CDR3 sequence.We will provide all patient information

Your TCR CDR3 sequence:

Your BCR CDR3 sequence:

Sample	ID	Clones	Proportion
IDH-MUT_X1p19q-Codel	CGGA_1412	4.0000000000	0.0006652310
IDH-MUT_X1p19q-Codel	CGGA_1412	2.0000000000	0.0003326150
IDH-MUT_X1p19q-Codel	CGGA_1412	2.0000000000	0.0003326150
IDH-MUT_X1p19q-Codel	CGGA_1412	2.0000000000	0.0003326150

Figure 5. Functionality at GTABdb showing information and immune repertoire characteristics of gliomas

(A) Kaplan-Meier analysis for comparing overall survival among different selected groups of gliomas by 'OS' page.
(B) The distribution of CDR3 sequences from TCR immune repertoire by 'Length' page.
(C) The usage frequency of VDJ genes encoding BCR by 'Gene' page.
(D) The diversity indexes for selected gliomas by 'Diversity' page.
(E) The combined frequency of VDJ gene rearrangements of CDR3 by 'Rearrangement' page.
(F) The 'Search' page for TCR and BCR CDR3 sequence.
(G) The information associated with a CDR3 sequences by 'Search' page.

(Figure 5G). Additionally, To help advanced users investigate their own TCR-/BCR-immune repertoire, GTABdb provides 'Tools', which lists some commonly used open source or on-line tools including Alignment Analysis, Computational Immunology, and 3D-Structure analysis. GTABdb also has 'Virus epitope', which integrates 316,012 items with matched TCR chain name, CDR3 sequence, HLA-subtypes and viral epitopes collected from IEDB[34] and VDJdb.[35] The help page of GTABdb contains extensive and detailed manual to aid users in understanding the layout and features of the database.

As an example, to explore *TRB* repertoire diversity, we examined Shannon-Wiener Index and Gini-Simpson index, Chao1 index, ACE index, observed clonotypes and pielou index through 'Diversity & Complexity' page at GTABdb for IDH wild type, IDHmut-noncodel, and IDH-mutcodel gliomas, respectively. We discovered that IDH wild type gliomas had significantly higher *TRB* repertoire diversity than IDHmut-noncodel and IDHmut-codel gliomas, which were mutually validated by the different ecological metrics (Figure 1).

DISCUSSION

In this work, we reconstructed more than 320,000 TCR and BCR from the largest glioma cohort in Chinese population including a total of 913 glioma RNA sequencing samples through an improved *de novo* TCR and BCR assembler, TRUST algorithm (vesion 4). We explored the repertoire diversity and complexity across three major glioma molecular subtypes (IDH wild type, IDHmut-noncodel, and IDHmut-codel gliomas) by applying graphical representation and six different ecological metrics upon clonotypes derived from V-gene, D-gene, J-gene and CDR3 amino acid sequence. The diversity and richness of the T and B cell receptor immune repertoire were demonstrated to be the highest in IDH wild type, followed by IDHmut-noncodel, and then IDHmut-codel gliomas, which correlates with the progressive severity of glioma molecular subtypes.

In addition, we found that the three glioma molecular subtypes had strong preferences for the frequency of V(D)J gene usage. The skewed usage of V, D, and J genes may affect CDR3 regions, leading to the varying constrains upon TCR and BCR immune repertoire in different glioma subtypes. We also demonstrated that there were more TCR and BCR clonotypes in IDH wild type compared with IDHmut-noncodel and IDHmut-codel. Moreover, more than 98% TCR and BCR clonotypes were unique in each glioma (private clonotype) with less than 2% shared among at least two patients with glioma (public clonotype). By analyzing the immune repertoire of 913 patients, we found that these data suggest that gliomas have highly large tumor heterogeneity in terms of immune repertoire, and precision medicine based immunotherapeutic interventions would be more suitable for patients with glioma.

Amino acid composition and hydrophobicity characteristics have potentially important effects on both antigen binding and antigen recognition. CDR3 variable motifs were found having higher proportions of hydrophobic residues in private than in public clonotypes. Previous studies[9,30] found β-CDR3 sequence in peripheral blood and tumor infiltrative private T cells significantly longer than the sequence of public T cells. We also observed that *TRB*-CDR3 sequences of private clonotypes was longer that public clones in gliomas, suggesting *TRB*-CDR3 sequences of private clonotypes in the T cell receptor immune repertoire have a greater ability to recognize antigens compared to public clonotypes.

Additionally, to conduct cross-cohort analysis, we used TRUST4 to assemble TCR and BCR from 667 RNA-seq data in TCGA. We found that 232 gliomas (including 158 IDH wild type, 50 IDHmut-noncodel and 24 IDHmut-codel gliomas) and 366 gliomas (including 185 IDH wild type, 121 IDHmut-noncodel and 60 IDHmut-codel gliomas) had different complete TCR and BCR CDR3 sequences successfully assembled, respectively, of which each glioma had varying number of different complete TCR and BCR CDR3 sequences (the median number of TCR CDR3 = 2; the median number of BCR CDR3 sequences = 4). We calculated and compared the diversity, richness and evenness for TCR and BCR repertoire across IDH wild type, IDHmut-noncodel, and IDHmut-codel gliomas. Richness measures (including observed clones, Chao1 and ace) and evenness measure (Pielou) of *TRA* repertoire was significantly higher in IDH wild type than IDHmut-codel or IDHmut-noncodel gliomas (Figure S15). But none of the measures of *TRB* repertoire exhibited significantly different across the three glioma subtypes (Figure S16). Observed clones, Chao1 and Pielou of *IGH* repertoire was significantly higher in IDH wild type than IDHmut-codel or IDHmut-noncodel gliomas (Figure S17). All 6 indexes were higher for IGL repertoire in IDH wild type than IDHmut-codel or IDHmut-noncodel (Figure S18). Gini-Simpson, Shannon diversity index, observed clones, Chao1, Pielou, and ace were significantly higher in IDH wild type than IDHmut-codel or IDHmut-noncodel (Figure S19). In short, there was a non-negligible trend that IDH wild type demonstrated prominent diversity, richness and evenness of TCR and BCR immune repertoire compared with IDHmut-noncodel or IDHmut-codel gliomas, consistent with our observations in the Chinese glioma population.

We integrated the data of the glioma immune repertoire and constructed a glioma immune repertoire database, GTABdb, for researchers to use and study. We also referred to the data visualization and online analysis services provided by the DREAM[36] and the user-friendly open access web interface by NoncoRNA[37] databases when we constructed GTABdb. We have demonstrated how a user could reproduce the immune repertoire analyses in this study and make new observations. The strength of GTABdb lies in the TCR and BCR immune repertoire from the largest glioma cohort in Chinese population, sufficient matched clinical data and a variety of user-friendly repertoire analysis tools. Continued maintenance and expansion of GTABdb by adding new data and cross-cohort analysis would definitely benefit the cancer immunology and immunotherapy research communities.

Limitations of the study

This study presents a comprehensive characterization of the immune repertoire of gliomas in the Chinese population, which represents the largest sample size to date. Additionally, we constructed the immune repertoire database for gliomas. However, the immune repertoire for gliomas from other population remain to be established. Especially, future possible accessibility of high quality controlled-access raw data from international organizations such as GLASS[38] and CPTAC[39] would provide us rich resources for validating and strengthening our findings. Moreover, our findings highlight the characteristics of immune receptor repertoire in different glioma subtypes, and provide rich data and analysis, yet the biological interpretation of the results requires further experimental validation and mechanistic studies to ascertain the potential biological implications. Future research is warranted to elucidate the roles of immune repertoires in tumorigenesis and potential glioma immunotherapy.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108661.

ACKNOWLEDGMENTS

We thank Taicang Qingli Biotechnology Co., Ltd and Ling Wei Pharmaceutical Inc for providing valuable comments and discussion about bioinformatics analysis of this work.

This work was supported by Beijing hospital Authority Clinical medicine Development special funding support, code ZLRK202333 (J.Z.), Fundamental Research Funds for the General Universities (J.Z.), Youth Thousand Scholar Program of China (J.Z.), and Program for Young Talents, Beihang University (J.Z.).

AUTHOR CONTRIBUTIONS

J.Z. designed the experiments. L.L. extracted TCR and BCR from RNA-seq data. L.W. and Z.Y.X. performed immune repertoire analysis. Z.Y.X. and L.W. constructed the database. J.Z., W.Z., L.W., Z.Y.X., and X.L. interpreted the data. J.Z., L.W., X.Z.Y., W.Z., and X.L. wrote the article.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: August 14, 2023
Revised: November 27, 2023
Accepted: December 5, 2023
Published: December 8, 2023

REFERENCES

1. Varn, F.S., Johnson, K.C., Martinek, J., Huse, J.T., Nasrallah, M.P., Wesseling, P., Cooper, L.A.D., Malta, T.M., Wade, T.E., Sabedot, T.S., et al. (2022). Glioma progression is shaped by genetic evolution and microenvironment interactions. Cell *185*, 2184–2199.e16.
2. Berghoff, A.S., Kiesel, B., Widhalm, G., Wilhelm, D., Rajky, O., Kurscheid, S., Kresl, P., Wöhrer, A., Marosi, C., Hegi, M.E., and Preusser, M. (2017). Correlation of immune phenotype with IDH mutation in diffuse glioma. Neuro Oncol. *19*, 1460–1468.
3. Wen, P.Y., Weller, M., Lee, E.Q., Alexander, B.M., Barnholtz-Sloan, J.S., Barthel, F.P., Batchelor, T.T., Bindra, R.S., Chang, S.M., Chiocca, E.A., et al. (2020). Glioblastoma in adults: a Society for Neuro-Oncology (SNO) and European Society of Neuro-Oncology (EANO) consensus review on current management and future directions. Neuro Oncol. *22*, 1073–1113.
4. Louis, D.N., Perry, A., Reifenberger, G., von Deimling, A., Figarella-Branger, D., Cavenee, W.K., Ohgaki, H., Wiestler, O.D., Kleihues, P., and Ellison, D.W. (2016). The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. Acta Neuropathol. *131*, 803–820.
5. Brown, C.E., Alizadeh, D., Starr, R., Weng, L., Wagner, J.R., Naranjo, A., Ostberg, J.R., Blanchard, M.S., Kilpatrick, J., Simpson, J., et al. (2016). Regression of Glioblastoma after Chimeric Antigen Receptor T-Cell Therapy. N. Engl. J. Med. *375*, 2561–2569.
6. Zhang, Y., Mudgal, P., Wang, L., Wu, H., Huang, N., Alexander, P.B., Gao, Z., Ji, N., and Li, Q.J. (2020). T cell receptor repertoire as a prognosis marker for heat shock protein peptide complex-96 vaccine trial against newly diagnosed glioblastoma. OncoImmunology *9*, 1749476.
7. Alt, F.W., Oltz, E.M., Young, F., Gorman, J., Taccioli, G., and Chen, J. (1992). VDJ recombination. Immunol. Today *13*, 306–314.
8. Davis, M.M., and Bjorkman, P.J. (1988). T-cell antigen receptor genes and T-cell recognition. Nature *334*, 395–402.

9. Li, B., Li, T., Pignon, J.C., Wang, B., Wang, J., Shukla, S.A., Dou, R., Chen, Q., Hodi, F.S., Choueiri, T.K., et al. (2016). Landscape of tumor-infiltrating T cell repertoire of human cancers. Nat. Genet. *48*, 725–732.
10. Thorsson, V., Gibbs, D.L., Brown, S.D., Wolf, D., Bortone, D.S., Ou Yang, T.H., Porta-Pardo, E., Gao, G.F., Plaisier, C.L., Eddy, J.A., et al. (2018). The Immune Landscape of Cancer. Immunity *48*, 812–830.e14.
11. Georgiou, G., Ippolito, G.C., Beausang, J., Busse, C.E., Wardemann, H., and Quake, S.R. (2014). The promise and challenge of high-throughput sequencing of the antibody repertoire. Nat. Biotechnol. *32*, 158–168.
12. Nutt, S.L., Hodgkin, P.D., Tarlinton, D.M., and Corcoran, L.M. (2015). The generation of antibody-secreting plasma cells. Nat. Rev. Immunol. *15*, 160–171.
13. Raposo, G., Nijman, H.W., Stoorvogel, W., Liejendekker, R., Harding, C.V., Melief, C.J., and Geuze, H.J. (1996). B lymphocytes secrete antigen-presenting vesicles. J. Exp. Med. *183*, 1161–1172.
14. Nelson, B.H. (2010). CD20+ B cells: the other tumor-infiltrating lymphocytes. J. Immunol. *185*, 4977–4982.
15. Linnebacher, M., and Maletzki, C. (2012). Tumor-infiltrating B cells: The ignored players in tumor immunology. OncoImmunology *1*, 1186–1188.
16. Nielsen, J.S., and Nelson, B.H. (2012). Tumor-infiltrating B cells and T cells: Working together to promote patient survival. OncoImmunology *1*, 1623–1625.
17. Mandric, I., Rotman, J., Yang, H.T., Strauli, N., Montoya, D.J., Van Der Wey, W., Ronas, J.R., Statz, B., Yao, D., Petrova, V., et al. (2020). Profiling immunoglobulin repertoires across multiple human tissues using RNA sequencing. Nat. Commun. *11*, 3126.
18. Hu, X., Zhang, J., Wang, J., Fu, J., Li, T., Zheng, X., Wang, B., Gu, S., Jiang, P., Fan, J., et al. (2019). Landscape of B cell immunity and related immune evasion in human cancers. Nat. Genet. *51*, 560–567.
19. Lee, Y.N., Frugoni, F., Dobbs, K., Tirosh, I., Du, L., Ververs, F.A., Ru, H., Ott de Bruin, L., Adeli, M., Bleesing, J.H., et al. (2016). Characterization of T and B cell repertoire diversity in patients with RAG deficiency. Sci. Immunol. *1*, eaah6109.
20. Lu, K.H.N., Michel, J., Kilian, M., Aslan, K., Qi, H., Kehl, N., Jung, S., Sanghvi, K., Lindner, K., Zhang, X.W., et al. (2022). T cell receptor dynamic and transcriptional determinants of T cell expansion in glioma-infiltrating T cells. Neurooncol. Adv. *4*, vdac140.
21. Sims, J.S., Grinshpun, B., Feng, Y., Ung, T.H., Neira, J.A., Samanamud, J.L., Canoll, P., Shen, Y., Sims, P.A., and Bruce, J.N. (2016). Diversity and divergence of the glioma-infiltrating T-cell receptor repertoire. Proc. Natl. Acad. Sci. USA *113*, E3529–E3537.
22. Lee, A.H., Sun, L., Mochizuki, A.Y., Reynoso, J.G., Orpilla, J., Chow, F., Kienzler, J.C., Everson, R.G., Nathanson, D.A., Bensinger, S.J., et al. (2021). Neoadjuvant PD-1 blockade induces T cell and cDC1 activation but fails to overcome the immunosuppressive tumor associated macrophages in recurrent glioblastoma. Nat. Commun. *12*, 6938.
23. Platten, M., Bunse, L., Wick, A., Bunse, T., Le Cornet, L., Harting, I., Sahm, F., Sanghvi, K., Tan, C.L., Poschke, I., et al. (2021). A vaccine targeting mutant IDH1 in newly diagnosed glioma. Nature *592*, 463–468.
24. Song, L., Cohen, D., Ouyang, Z., Cao, Y., Hu, X., and Liu, X.S. (2021). TRUST4: immune repertoire reconstruction from bulk and single-cell RNA-seq data. Nat. Methods *18*, 627–630.
25. Zhao, Z., Zhang, K.N., Wang, Q., Li, G., Zeng, F., Zhang, Y., Wu, F., Chai, R., Wang, Z., Zhang, C., et al. (2021). Chinese Glioma Genome Atlas (CGGA): A Comprehensive Resource with Functional Genomic Data from Chinese Glioma Patients. Dev. Reprod. Biol. *19*, 1–12.
26. Chen, S.Y., Yue, T., Lei, Q., and Guo, A.Y. (2021). TCRdb: a comprehensive database for T-cell receptor sequences with powerful search function. Nucleic Acids Res. *49*, D468–D474.
27. Shugay, M., Bagaev, D.V., Zvyagin, I.V., Vroomans, R.M., Crawford, J.C., Dolton, G., Komech, E.A., Sycheva, A.L., Koneva, A.E., Egorov, E.S., et al. (2018). VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. Nucleic Acids Res. *46*, D419–D427.
28. Song, L., Ouyang, Z., Cohen, D., Cao, Y., Altreuter, J., Bai, G., Hu, X., Livak, K.J., Li, H., Tang, M., et al. (2022). Comprehensive Characterizations of Immune Receptor Repertoire in Tumors and Cancer Immunotherapy Studies. Cancer Immunol. Res. *10*, 788–799.
29. Ruggiero, E., Nicolay, J.P., Fronza, R., Arens, A., Paruzynski, A., Nowrouzi, A., Ürenden, G., Lulay, C., Schneider, S., Goerdt, S., et al. (2015). High-resolution analysis of the human T-cell receptor repertoire. Nat. Commun. *6*, 8081.
30. Warren, R.L., Freeman, J.D., Zeng, T., Choe, G., Munro, S., Moore, R., Webb, J.R., and Holt, R.A. (2011). Exhaustive T-cell repertoire sequencing of human peripheral blood samples reveals signatures of antigen selection and a directly measured repertoire size of at least 1 million clonotypes. Genome Res. *21*, 790–797.
31. Freeman, J.D., Warren, R.L., Webb, J.R., Nelson, B.H., and Holt, R.A. (2009). Profiling the T-cell receptor beta-chain repertoire by massively parallel sequencing. Genome Res. *19*, 1817–1824.
32. Philibert, P., Stoessel, A., Wang, W., Sibler, A.P., Bec, N., Larroque, C., Saven, J.G., Courtête, J., Weiss, E., and Martineau, P. (2007). A focused antibody library for selecting scFvs expressed at high levels in the cytoplasm. BMC Biotechnol. *7*, 81.
33. Hoi, K.H., and Ippolito, G.C. (2013). Intrinsic bias and public rearrangements in the human immunoglobulin Vlambda light chain repertoire. Gene Immun. *14*, 271–276.
34. Vita, R., Mahajan, S., Overton, J.A., Dhanda, S.K., Martini, S., Cantrell, J.R., Wheeler, D.K., Sette, A., and Peters, B. (2019). The Immune Epitope Database (IEDB): 2018 update. Nucleic Acids Res. *47*, D339–D343.
35. Goncharov, M., Bagaev, D., Shcherbinin, D., Zvyagin, I., Bolotin, D., Thomas, P.G., Minervina, A.A., Pogorelyy, M.V., Ladell, K., McLaren, J.E., et al. (2022). VDJdb in the pandemic era: a compendium of T cell receptors specific for SARS-CoV-2. Nat. Methods *19*, 1017–1019.
36. Li, S., Li, L., Meng, X., Sun, P., Liu, Y., Song, Y., Zhang, S., Jiang, C., Cai, J., and Zhao, Z. (2021). DREAM: a database of experimentally supported protein-coding RNAs and drug associations in human cancer. Mol. Cancer *20*, 148.
37. Li, L., Wu, P., Wang, Z., Meng, X., Zha, C., Li, Z., Qi, T., Zhang, Y., Han, B., Li, S., et al. (2020). NoncoRNA: a database of experimentally supported non-coding RNAs and drug targets in cancer. J. Hematol. Oncol. *13*, 15.
38. GLASS Consortium (2018). Glioma through the looking GLASS: molecular evolution of diffuse gliomas and the Glioma Longitudinal Analysis Consortium. Neuro Oncol. *20*, 873–884.
39. Edwards, N.J., Oberti, M., Thangudu, R.R., Cai, S., McGarvey, P.B., Jacob, S., Madhavan, S., and Ketchum, K.A. (2015). The CPTAC Data Portal: A Resource for Cancer Proteomics Research. J. Proteome Res. *14*, 2707–2713.
40. Oldrini, B., Vaquero-Siguero, N., Mu, Q., Kroon, P., Zhang, Y., Galán-Ganga, M., Bao, Z., Wang, Z., Liu, H., Sa, J.K., et al. (2020). MGMT genomic rearrangements contribute to chemotherapy resistance in gliomas. Nat. Commun. *11*, 3883.
41. Soto, C., Bombardi, R.G., Kozhevnikov, M., Sinkovits, R.S., Chen, E.C., Branchizio, A., Kose, N., Day, S.B., Pilkinton, M., Gujral, M., et al. (2020). High Frequency of Shared Clonotypes in Human T Cell Receptor Repertoires. Cell Rep. *32*, 107882.
42. Nocaj, A., and Brandes, U. (2012). Computing Voronoi Treemaps: faster, simpler and resolution-independent. Comput. Graph. Forum *31*, 855–864.
43. Shannon, C.E. (2001). The mathematical theory of communication. SIGMOBILE Mob. Comput. Commun. Rev. *5*, 3–55.
44. Simpson, E.H. (1949). Measurement of Diversity. Nature *163*, 688.
45. Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. Biometrics *43*, 783–791.
46. Chiu, C.H., Wang, Y.T., Walther, B.A., and Chao, A. (2014). An improved nonparametric lower bound of species richness via a modified good-turing frequency formula. Biometrics *70*, 671–682.
47. O'Hara, R.B. (2005). Species richness estimators: how many species can dance on the head of a pin? J. Anim. Ecol. *74*, 375–386.
48. Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res. *14*, 1188–1190.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Deposited data		
CGGA RNA-seq raw data	National Genomics Data Center	PRJCA001746 PRJCA001747
CGGA clinical information	CGGA portal	http://www.cgga.org.cn/
TCGA RNA-seq raw data	TCGA portal	https://portal.gdc.cancer.gov/
TCGA clinical information	TCGA portal	https://portal.gdc.cancer.gov/
Software and algorithms		
Analysis scripts	This paper	https://doi.org/10.6084/m9.figshare.24708324.v2
TRUST (version4)	Song et al.[24]	https://github.com/liulab-dfci/TRUST4
GraphPad Prism version 9.5.8	GraphPad software	https://www.graphpad.com/
Rstudio 1.3.1093	Posit, USA	https://posit.co/downloads/
R programming language 4.0.3	R Foundation, USA	https://www.r-project.org

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Jing Zhang (jz2716@126.com).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- This paper analyzes existing, publicly available data. These accession numbers for the datasets are listed in the key resources table.
- All original code has been deposited at Figshare and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

All data used in this study were obtained from the public domain.

Ethics statement

Not applicable.

Data collection

In this study, we utilized the RNA-seq raw data of gliomas generated from the previously published Chinese glioma genomic atlas,[25,40] which have been deposited in the National Genomics Data Center (NGDC) with accession numbers PRJCA001746 and PRJCA001747. The clinical information for these samples was obtained from Chinese Glioma Genome Atlas (CGGA) (http://www.cgga.org.cn). TCGA RNA-seq data and corresponding clinical information were downloaded from TCGA portal.

A total of 913 glioma patients from the Chinese population were included, along with 399 glioma patients from the TCGA dataset. The inclusion criteria were as follows.

Exclusion criteria

(1) According to the newly updated Central Nervous System classification by the World Health Organization, samples with missing information on the mutational status of the isocitrate dehydrogenase 1 and 2 genes (IDH1/2) were excluded. For samples with IDH1/2 mutations, those with missing information on the codeletion status of chromosome arms 1p and 19q were also excluded.

(2) Samples containing only incomplete CDR3 sequences were excluded.

The Chinese glioma patients comprised 424 IDH wild type, 177 IDHmut-codel, and 312 IDHmut-noncodel cases, totaling 913 cases analyzed. The TCGA glioma patients included 203 IDH wild type, 64 IDHmut-codel, and 132 IDHmut-noncodel cases, totaling 399 cases analyzed.

METHOD DETAILS

RNA-seq data and TCR and BCR assembly

RNA-seq raw data of gliomas were previously generated by Chinese Glioma Genome Atlas,[25,40] which were deposited at National Genomics Data Center with accession number of PRJCA001746 and PRJCA001747. A total of 913 glioma samples with accurate clinical information about IDH mutation status, 1p/19q co-deletion, age, gender, tumor grade, and overall survival were included in the analysis.

T cell receptors and B cell receptors were assembled from RNA sequencing data using TRUST algorithm (version 4). A total of 14,770 clonotypes across all TCR chains (*TRA*,*TRB*) and a total of 311,146 clonotypes across BCR chains were recovered from 913 gliomas.

Clonotypes definitions

When calculating the diversity, richness, evenness, and clonality of the immune repertoire, we defined the clonotypes as follows: for TRB/IGH, we defined the same V gene, the same D gene, the same J gene, and the same CDR3 amino acid sequence as identical clonotype. For TRA/IGL/IGK, we took the same V gene, the same J gene, and the same CDR3 amino acid sequence as identical clonotype.[41]

Drawing Voronoi Treemaps

We used V gene-J gene pair to draw the tree maps. If there were the same rearrangement of V gene and J gene in the same patient sample, the sum of the read counts of the same V gene-J gene pair was regarded as the read count of the specific V gene-J gene pair clonotype in this patient sample. After processing each patient sample according to the above method, for each specific V gene-J gene pair clonotype, we calculated the mean value of the read count of each V gene-J gene pair clonotype in all patient samples containing the VJ pair clonotype. The formula was as follows:

$$specific\ VJ\ pair\ clonotype\ read\ counts = \frac{\sum_{i=1}^{n} c_i}{n}$$

where n represented the number of patients with a specific V gene-J gene pair clonotype.c_i represented the number of read count in patient sample i that contained a specific V gene -J gene pair clonotype.

Next, we used the data of V gene-J gene pair clonotypes and V gene-J gene pair clonotypes read count to draw Voronoi Treemaps with R package of voronoiTreemap.[42] Each polygon represented a specific V gene-J gene pair, and the size of the polygon represented a specific V gene-J gene pair clonotype of read count.

Diversity indices

To assess the diversity of the immune repertoire (TCR repertoire and BCR repertoire), we used two different diversity evaluation metrics: Shannon–Wiener index[43] and Gini-Simpson index.[44] Both diversity evaluation metrics took into account the richness and evenness of clonotypes. Richness was defined as the number of species of clonotypes, and evenness was defined as the distribution of clonotypes or the relative abundance of clonotypes. The Shannon–Wiener index and Gini-Simpson index were as follows:

$$\text{Shannon-Wienerindex} = -\sum_{i=1}^{s} \frac{n_i}{N} log_2 \frac{n_i}{N}$$

$$\text{Gini-Simpsonindex} = 1 - \sum_{i=1}^{s} \left(\frac{n_i}{N}\right)^2$$

where s was the number of clonotypes, n_i was the number of reads in clonotypes i and N was the sum of all reads in the TCR repertoire or BCR repertoire. The R package used for diversity indices calculation was vegan (version 2.5–7).

Richness indices

Observed Clonotypes (Obc), Chao1 and ACE were used to characterize richness of the immune repertoire.[45–47] Observed clonotypes was the number of species of clonotypes as follows:

$$\text{Obc} = S_{obc}$$

Chao1 was defined as:

$$\mathrm{Chao1} = S_{obc} + \frac{a_1(a_1-1)}{2(a_2+1)}$$

where a_1 was the number of clonotypes with only one read and a_2 was the number of clonotypes with two reads in the TCR repertoire or BCR repertoire. The R package used for richness indices calculation was vegan (version 2.5–7).

ACE was calculated with:

$$\mathrm{ACE} = S_{abund} + \frac{S_{rare}}{C_{ACE}} + \frac{a_1}{C_{ACE}}\gamma^2_{ACE}$$

Where,

S_{abund}: The number of species of clonotypes with more than or equal to 10 reads.

S_{rare}: The number of species of clonotypes with less than 10 reads

a_1 is the number of clonotypes with only one read.

$$C_{ACE} = 1 - \frac{a_1}{N_{rare}} \text{ where } N_{rare} = \sum_{i=1}^{10} ia_i$$

a_i was the number of clonotypes with only i reads

$$\gamma^2_{ACE} = max\left[\frac{S_{rare}}{C_{ACE}} \frac{\sum_{i=1}^{10} i(i-1)a_i}{N_{rare}(N_{rare}-1)} - 1, 0\right]$$

Evenness indices

Evenness was calculated with the Pielou as:

$$Pielou = \frac{-\sum_{i=1}^{s} \frac{n_i}{N} log_2 \frac{n_i}{N}}{log_2(S)}$$

where s was the number of clonotypes, n_i was the number of reads in clonotypes i and N was the sum of all reads in the TCR repertoire or BCR repertoire of a given sample. The R package used for evenness indices calculation was vegan (version 2.5–7).

Gene usage

In the TCR immune repertoire and BCR immune repertoire, we first obtained the chromosomes of V, D, and J genes in the immune repertoire from the IMGT database (https://www.imgt.org/IMGTrepertoire/LocusGenes/). The order of arrangement and functional genes were retained. The V gene, D gene, and J gene we finally got were the coding genes arranged in order on the chromosome for subsequent analysis.

We calculated the gene usage at the abundance of clonotypes，which was to use the corresponding read count to weight genes. We first divided the patients into three groups: IDH wild type, IDHmut-noncodel and IDHmut-codel. Taking the V gene of the IDH wild type patient as an example, we regarded the use frequency of a specific V gene in IDH wild type as the average of the use frequency of this specific V gene in all samples in IDH wild type. The specific calculation formula is as follows:

$$V_m(specific\ V\ gene\ usage\ in\ IDH\ wild\ type) = \frac{\sum_{i=1}^{n} r_i}{n}$$

Among them, the number of patient samples in IDH wild type group was n, the read count of the specific V gene of the n_{th} patient sample was r_n, and m was the frequency of use of the m_{th} V gene.

The calculation formula of D gene and J gene was the same as that of V gene. The calculation formula of the gene frequency of IDHmut-codel and IDHmut-noncodel was the same as that of IDH wild type.

For goodness-of-fit tests, we calculated the expected distribution of genes in IDH wild type, IDHmut-codel and IDHmut-noncodel using the following formulas.

Took the V gene in IDH wild type as an example: :

$$IDH\ wild\ type\ expect\ probabilities = \frac{V_i}{\sum_{i=1}^{m} V_i}$$

m was the number of V genes in IDH wild type.

We calculated the gene usage at the abundance of clonotypes , which was to use the corresponding read count to weight genes. Took the calculation of the V gene as an example: in a patient sample, there was a total of n V genes, and the usage calculation formula of the nth V gene was as follows:

$$specific\ V\ gene\ usage = \frac{g_i}{\sum_{i=1}^{n} g_i}$$

Where g_i was the read count of the i_{th} V gene.

For statistical significance, we retained genes that were present in 3 and more samples.

The R package used for gene usage calculation was immunarch.

CDR3 length distributions and conservative analysis

The distribution of CDR3 lengths for each patient was determined by the unique clonotype. First, we calculated the length of CDR3 amino acid sequences in each patient's unique clonotype. We then calculated the frequency of the length distribution of CDR3 amino acid sequences in each clonotype based on the abundance of clonotypes. Finally, the distribution frequencies of the identical CDR3 amino acid sequences lengths were summed. The graph for conservative analysis is created using WebLogo(version 2.8.2).[48]

Percentage of public clonotypes and private clonotypes between samples

We used the following rules to calculate the percentage of public clonotypes and private clonotypes between samples. Took IDH wild type as an example, in the IDH wild type, we defined the identical clonotype that two or more patients had as public clonotypes, and each patient's unique clonotype was defined as private clonotypes. The proportion of public clonotypes was defined as the number of public clonotypes divided by the total number of public clonotypes and private clonotypes. The proportion of private clonotypes was defined as the number of private clonotypes divided by the total number of public clonotypes and private clonotypes.

The calculation of length of public clonotypes and private clonotypes

To determine the length distribution of CDR3 amino acid sequences in public clonotypes and private clonotypes, we extracted CDR3 amino acid sequences in public clonotypes and private clonotypes, respectively, and calculated the lengths, and then filtered out the repetitive CDR3 amino acid sequences.

Proportion of hydrophobic amino acids of public clonotypes and private clonotypes

To determine the ratio of hydrophobic amino acids in the middle of the CDR3 amino acid sequences in the private clonotypes and the public clonotypes, we first extracted the CDR3 sequence of the private clonotypes and the public clonotypes and its length, and divided the length into odd and even numbers. For CDR3 amino acid sequence of odd lengths, the amino acid positions we extracted were calculated using the following formula:

$$position1 = \frac{odd\ length}{2} - 0.5$$

$$position2 = \frac{odd\ length}{2} - 0.5 + 1$$

$$position3 = \frac{odd\ length}{2} - 0.5 + 2$$

For CDR3 amino acid sequences of even length, the calculation formula was as follows:

$$position1 = \frac{even\ length}{2}$$

$$position2 = \frac{even\ length}{2} + 1$$

$$position3 = \frac{even\ length}{2} + 2$$

The R language pseudocode for the ratio of CDR3 amino acid sequences to the hydrophobic amino acids in the three middle positions was (took the position1 as an example):

$$proportion = \frac{sum(position1\ \%in\%\ c(''M'', ''F'', ''W'', ''V'', ''L'', ''I'', ''A'', ''P'', ''G''))}{length(position1)}$$

We used the binomial test to compare the difference in the ratio of hydrophobic amino acids in the public and private clonotypes at the same position, with the ratio of hydrophobic amino acids in the public clonotypes as the expected probability.

Construction of the immune repertoire database

GTABdb was implemented through mysql, HTML, JavaScript and PHP, and was deployed on Apache 2 (http://httpd.apache.org/) on Ubuntu Linux. MySQL (http://www.mysql.org) is a data management system for storing and querying data. The response conditions in the interactive website connect with MySQL through PHP, query the original data of interest and return it to the web, and then process the data through JavaScript and visualize the final results in HTML web. The structure of the whole website follows the model-view-controller architecture, each component is independent, supports parallel development, and has good scalability. The model component realizes the main functions of the website, including data search, processing response conditions from the website and data from mysql, which is the core function of the system. View component is the main user interface component, which provides a view to display website content and interact with users. The view component uses the front-end template of bootstrap to ensure the overall unity of the website interface. The controller component is the intermediary connecting the model component and the view component, which is responsible for transmitting front-end and back-end data and processing application logic.

In the survival curve analysis function section, the survival curve diagram was generated by using R package suvival and surviviner. In order to connect the database website and R, R package rserve was used. The search conditions at the front end of the web page are passed to the model component through the controller component, and then execute the SQL search command to get the original data. Then establish a connection between rserve and R, execute the R script to generate the survival curve analysis diagram and return to the front end of the web page. The implementation of other functions uses similar methods, such as obtaining search conditions from the front end of the web page, executing SQL search commands, processing the obtained data and visualization. JavaScript plug-in ecarts (https://echarts.apache.org/zh/index.html) was used for data drawing and visualization, including heatmap, histogram, box whisker diagram, etc. Tables in all functions use DataTables (http://datatables.club) DataTables is a jQuery table plug-in. It is a highly flexible tool that can add advanced interactive functions to any HTML table, realize the server-side display of a large amount of data, and speed up the response speed of the table. We also provide access links to online analysis tools, and visitors can freely use these tools to analyze data. GTABdb is available free of charge at the following website: http://www.oncoimmunobank.cn/glicrdb/database/homepage.

QUANTIFICATION AND STATISTICAL ANALYSIS

Chi-square test was used to describe the gender and grade differences of glioma patients with IDH wild type, IDHmut-codel and IDHmut-noncodel. Kruskal-Wallis tests was used to compare the differences of ages among the three glioma molecular classifications. Kaplan-Meier curve were plotted by GraphPad Prism 9.5.8, where p value was determined by log-rank test. Mann-Whitney U test was used to compare the Shannon-Wiener Index, Gini-Simpson index, Chao1, Observed Clonotypes, ACE, and Pielou differences between the two in IDH wild type gliomas, IDHmut-codel gliomas and IDHmut-noncodel gliomas.

Besides, Mann-Whitney U test was applied for the comparison of the CDR3 sequences in public clonotypes and private clonotypes. Binomial test and Benjamini–Hochberg correction was adopted for the comparison of proportion of hydrophobic amino acids between private and public of the middle of CDR3 sequences. For gene usage analysis, Chi-square test for goodness of fit was performed to study the frequency of V(D)J gene usage in immune repertoire of patients in IDHwt, IDHmut-noncodel and IDHmut-codel. Mann-Whitney U test and Benjamini–Hochberg correction were applied to compare the relative frequency of usage of V(D)J gene usage from IDHwt, IDHmut-noncodel and IDHmut-codel. Above statistical analysis were performed using GraphPad Prism 9.5.8 and R(version 4.0.3).