MYC amplification at diagnosis drives therapy-induced hypermutation of recurrent glioma

Jiguang Wang (✉ jgwang@ust.hk)
The Hong Kong University of Science and Technology https://orcid.org/0000-0002-6923-4097

Quanhua Mu
The Hong Kong University of Science and Technology

Ruichao Chai
Beijing Neurosurgical Institute

Hanjie Liu
Beijing Neurosurgical Institute

Yingxi Yang
Division of Life Science and Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong

Zheng Zhao
Beijing Neurosurgical Institute

Ming Hong Lui
Division of Life Science and Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong

Zhaoshi Bao
Beijing Neurosurgical Institute

Dong Song
hong kong university of science and technology https://orcid.org/0000-0002-5213-5121

Biaobin Jiang
Division of Life Science and Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong

Jason Sa
Korea University https://orcid.org/0000-0002-3251-5004

Hee Jin Cho
Samsung Medical Center

Yuzhou Chang
Beijing Neurosurgical Institute

Kaitlin Hao Yi Chan
HKUST https://orcid.org/0000-0001-7363-2082

Danson Shek Chun Loi
HKUST

Sindy Sing Ting Tam
HKUST
Aden Ka Yin Chan
CUHK
Angela Wu
Division of Life Science and Department of Chemical and Biological Engineering, Hong Kong University of Science and Technology, Hong Kong https://orcid.org/0000-0002-3531-4830
Wai San Poon
The Chinese University of Hong Kong
H.K. Ng
The Chinese University of Hong Kong
Danny Chan
Prince of Wales Hospital
Antonio Iavarone
Columbia University https://orcid.org/0000-0002-0683-4634
Do-Hyun Nam
Samsung Medical Center
Tao Jiang
Beijing Neurosurgical Institute https://orcid.org/0000-0002-7008-6351

Biological Sciences - Article

Keywords: Clonal Evolution, Adult Diffuse Gliomas, Matched Initial-recurrent Tumor Pairs, Grade Progression, Loss-of-function Mutations, Cancer Evolution

Posted Date: January 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-138020/v1

MYC amplification at diagnosis drives therapy-induced hypermutation of recurrent glioma

Quanhua Mu[1,#], Ruichao Chai[2,#], Hanjie Liu[2], Yingxi Yang[1], Zheng Zhao[2], Ming Hong Lui[1], Zhaoshi Bao[1,2], Dong Song[1], Biaobin Jiang[1], Jason K. Sa[4,5], Hee Jin Cho[4], Yuzhou Chang[2], Kaitlin Hao Yi Chan[3], Danson Shek Chun Loi[3], Sindy Sing Ting Tam[3], Aden Ka Yin Chan[7], Angela Ruohao Wu[3], Wai Sang Poon[6], Ho Keung Ng[7], Danny Tat Ming Chan[6], Antonio Iavarone[8,*], Do-Hyun Nam[4,9,10,11,*], Tao Jiang[2,11,*], Jiguang Wang[1,11,*]

[1] Division of Life Science, Department of Chemical and Biological Engineering, Center of Systems Biology and Human Health, State Key Laboratory of Molecular Neuroscience, The Hong Kong University of Science and Technology, Hong Kong SAR, China.

[2]Beijing Neurosurgical Institute, Capital Medical University, Beijing, China.

[3]Division of Life Science, Department of Chemical and Biological Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China.

[4]Institute for Refractory Cancer Research, Samsung Medical Center, Seoul, South Korea.

[5]Department of Biomedical Sciences, Korea University College of Medicine, Seoul, Korea.

[6]CUHK Otto Wong Brain Tumour Centre, Department of Surgery, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong SAR, China.

[7]Department of Anatomical and Cellular Pathology, Prince of Wales Hospital, The Chinese University of Hong Kong, Hong Kong SAR, China.

[8]Institute for Cancer Genetics, Columbia University, New York, New York, USA.

[9]Department of Neurosurgery, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, Korea.

[10]Department of Health Science & Technology, Samsung Advanced Institute for Health Sciences & Technology, Sungkyunkwan University School of Medicine, Seoul, Korea.

[11]Chinese Glioma Genome Atlas (CGGA) and Asian Glioma Genome Atlas (AGGA) Research Networks.

# These authors contributed equally

* Emails:
jgwang@ust.hk; ai2102@cumc.columbia.edu; nsnam@skku.edu; taojiang1964@163.com.

Abstract

Clonal evolution drives cancer progression and therapeutic resistance[1,2]. Recent longitudinal analyses revealed divergent clonal dynamics in adult diffuse gliomas[3–11]. However, the early genomic and epigenomic factors that steer post-treatment molecular trajectories remain unknown. To track evolutionary predictors, we analyzed sequencing and clinical data of matched initial-recurrent tumor pairs from 511 adult diffuse glioma patients. Using machine learning we developed methods capable of predicting grade progression and hypermutation from tumor characteristics at diagnosis. Strikingly, *MYC* copy number gain in initial tumors emerged as a key factor predicting development of hypermutation under temozolomide (TMZ) treatment. The driving role of MYC in TMZ-associated hypermutagenesis has been experimentally validated in a model of TMZ-induced hypermutator using both patient-derived gliomaspheres and established glioma cell lines. Subsequent studies showed that c-Myc binding to open chromatin and transcriptionally active regions increases the vulnerability of genomic regions to TMZ-induced mutagenesis. Consequently, MYC target genes, including the key mismatch repair genes, develop loss-of-function mutations, thus triggering the hypermutation process. This study reveals MYC as an early predictor of cancer evolution and provides a machine learning platform for predicting cancer dynamics to improve patient management.

Main

Malignant diffuse gliomas (WHO grade II to IV) constitute the most common primary brain tumors in adults[12]. Despite surgery combined with radiotherapy plus alkylator chemotherapy, aggressive gliomas inevitably recur. At relapse, a proportion of lower-grade gliomas (LGG) remain grade II-III but a remarkable number of cases progress into grade IV glioblastoma (GBM), leading to worse prognosis[3]. Molecular mechanisms that drive this process remain elusive.

To characterize the evolutionary landscape of glioma under therapy, enormous efforts were devoted to sequence cancer genomes at multiple time points[3–11]. The hypermutation (HM) phenomenon was detected in temozolomide-treated samples in several independent cohorts, highlighting the role of stress-induced mutagenesis in shaping adaptive glioma evolution[13,14]. Our recent study revealed a mutually exclusive pattern between HM and an alternative TMZ-resistant mechanism driven by *MGMT*[15]. Moreover, group efforts of Glioma Longitudinal AnalySiS (GLASS) consortium reconstructed molecular trajectories of 222 individual cases, demonstrating a heterogeneous and largely stochastic pattern of glioma evolution after treatment[8]. Yet, it is still unknown how the evolutionary routes were determined and whether the early predictors of cancer evolution exist.

Here we analyzed longitudinal sequencing data from a large number of diffuse glioma cases that cover all three molecular subtypes—IDH wildtype (IDHwt), IDH-mutant without chromosome 1p/19q co-deletion (IDHmut-noncodel), and IDH-mutant with chromosome 1p/19q co-deletion (IDHmut-codel). Leveraging information theory and

machine learning, we aim to quantify clonal stochasticity and track early molecular predictors that drive cancer evolution.

Longitudinal sequencing reveals mutational dynamics in glioma

We assembled 511 matched initial and recurrent diffuse glioma pairs, among which 95 out of 219 LGG cases progressed to GBM at recurrence. Out of these samples, 276 were profiled by whole-genome sequencing (WGS), 989 by whole exome sequencing (WES), 394 by RNA sequencing, and 201 by clinical panel sequencing (**Extended Data Fig. 1a-b**, **Extended Data Table 1-2**). Using calibrated computational pipelines, we portrayed the dynamic landscape of somatic mutations in key driver genes (**Fig. 1a**, **Extended Data Fig. 1c-d**). Consistent with previous reports[3,8,9,16], we found that 17.0% (62 out of 365) of TMZ-treated cases in our cohort gained hypermutation at recurrence. Excluding hypermutated samples, elevated mutational burden at recurrence was observed in each of the three molecular subtypes of glioma with *P* values of 8.0×10^{-6}, 2.9×10^{-6} and 2.0×10^{-3} for IDHwt, IDHmut-noncodel, and IDHmut-codel (paired *t*-test), respectively (**Extended Data Fig. 1e**). Ternary plot showed that unlike IDHwt gliomas, IDH-mutant tumors had a remarkable number of key alterations enriched in recurrent-specific area (**Extended Data Fig. 2a-c**). In particular, 82 out of 128 IDHmut-noncodel and 13 out of 59 IDHmut-codel gliomas developed recurrence-specific alterations in canonical GBM drivers such as *CDKN2A* deletion and/or *CDK4* amplification and deletions or mutations in *PTEN* (**Extended Data Fig. 2b-c**). Interestingly, six recurrent IDHwt gliomas lost their initial alteration in *EGFR*vIII and nine recurrent IDHwt gliomas lost the copy number amplification in *MDM4*. Moreover, development of HM at recurrence was observed in all three subtypes, while one

IDHwt and two IDHmut-noncodel cases developed MGMT translocation. These results collectively implied that diffuse glioma recurred and developed chemoresistance in response to current therapeutic intervention via divergent molecular mechanisms.

We next systematically compared genetic alterations of initial and recurrent tumors in all patients and found that 70.2% of cases (269 out of 383 patients sequenced by WES/WGS) lost >5 coding mutations (single nucleotide variants and small indels) while 81.9% (313 out of 383) cases gained >5 new mutations (**Fig. 1b**). For each molecular subtype, we observed significantly larger number of recurrence-specific compared to initial-specific mutations ($P = 8.4\times10^{-7}$, 6.3×10^{-12} and 1.5×10^{-4} for IDHwt, IDHmut-noncodel, and IDHmut-codel, respectively) (**Fig. 1b**). To quantify the changeability of each glioma feature during evolution, we defined its longitudinal divergence by binarizing the change of its mutational status over time and calculating the information entropy (**Fig. 1c**). As expected, no longitudinal change was observed for mutational status of *IDH1* or co-deletion of 1p/19q and hence these features contained zero entropy, implying that these alterations either achieved fixation due to their overwhelming survival advantage or occurred extremely early during glioma evolution (**Fig. 1d**, **Extended Data Fig. 2d-f**). In contrast, the pathologic grade, somatic hypermutation (HM), together with genomic deletions in *CDKN2A*, *RB1* and *PTEN* contained the highest entropy, implying these features changed frequently over time (**Fig. 1d**). We reason that the change of mutational status of a glioma feature was not only determined by stochasticity, but also related to potential predisposition factors observed early. Therefore, we calculated mutual information between each feature's longitudinal divergence and data acquired from initial

tumor to quantify predictability of the corresponding feature (Methods). Interestingly, pathological grading, and HM are two features with the highest predictability, implying these two events might be deployed earlier (**Fig. 1d**, **Extended Data Fig. 2g-h**).

Genomic characteristics of initial glioma informs grade progression

We then investigated the longitudinal change of pathologic grade in our cohort. Not surprisingly, we found that patients who progressed to GBM showed significantly worse survival compared to those who recurred as LGG, regardless of their molecular subtypes (**Fig. 2a**). The above information theory analysis suggested that the risk of glioma grade progression was related to early genetic and clinical factors. We therefore developed a machine learning model to predict grade progression (LGG to GBM) from genomic and clinical characteristics collected at diagnosis.

To develop the machine learning model, we compiled data of 116 LGG cases as the training cohort (**Fig. 2b**, **Extended Data Table 1**). Particularly, the training cohort included 53 LGG-GBM longitudinal pairs together with 63 LGG-LGG pairs. Overall, three classification models were trained using different sets of features: Model 1 used the initial grade as the single predictor, Model 2 used the initial grade, age, gender, and treatment as predictors, while Model 3 included the initial grade, age, gender, treatment, and genomic alterations as predictors. The performance of the models was measured by the area under the receiver operating characteristic curve (AUC). Based on five-fold cross validation we found that the model using integrated features including patient clinical information, applied treatment plus genomic alterations achieved the highest AUC (0.87, 95% CI 0.81-

0.94, DeLong test, **Fig. 2c**). The SHAP (Shapley additive explanations) score, a metric to quantify the importance of each feature[17,18], was then calculated and used to prioritize the relevant features. On the top of the list is IDH status, where IDH mutation contributed negatively to grade progression, and IDH wildtype are more prone to progress. Following IDH mutations are 17p copy neutral loss of heterozygosity (CNLOH), 1p/19q codel, alkylator treatment and *CDKN2A* deletion (**Fig. 2d**). Other relevant genetic factors include amplification/gain of *MET*, chr7p and *MYC*, deletions of *TP53* and *RB1*, as well as *TERT* promoter mutations. Notably, *MYC*, a master regulator of stemness and proliferation, has been recently revealed to regulate glioma progression[19].

To test the model performance, 50 independent LGG cases from GLASS were used as the hold-out testing cohort. In this cohort, 21 LGGs progressed to GBM at recurrence and the other 29 did not. Our model achieved an AUC of 0.79 (95% confidence interval 0.67-0.92, DeLong test; **Fig. 2e**). We then applied the model on 496 TCGA primary LGGs to assess whether the predicted progression risk to GBM at recurrence might be associated with patient survival outcome. As a result, remarkable worse overall survival (OS, $P = 1.34 \times 10^{-8}$, log-rank test) and progression-free survival (PFS, $P = 1.15 \times 10^{-6}$, log-rank test) outcomes were observed for cases predicted to progress to GBM, compared to those predicted to recur as LGG (**Fig. 2f**). More importantly, when glioma patients were evaluated in each of the three molecular groups, poorer PFS and OS were consistently associated with predicted progression risk (**Extended Data Fig. 3a-b**). These findings demonstrated that our machine learning model was able to integrate clinical information

and molecular features in order to evaluate the risk of glioma progression and early predict patient prognosis.

Features of initial tumors predispose hypermutation at recurrence

We adopted CELLO toolkit[20] to infer HM from multi-platform sequencing data and identified 68 HM samples with average mutation load 157.7 mutations per Mb and average HM score 1.39 (**Extended Data Fig. 4a**). In our dataset, we demonstrated that HM was associated with glioma progression and impacts post-progression survival (**Extended Data Fig. 4b-c**). Consistent with previous studies[16,21,22], most (16 out of 18 samples with available MGMT methylation status, 88.9%) of HM samples were MGMT methylated, and 59 of 65 (90.8%) HM cases gained loss-of-function mutations in the mismatch repair (MMR) pathway genes such as *MSH6* and *MSH2* (**Extended Data Fig. 4d**. We then compared the transcriptional profile of TMZ-treated HM versus NHM recurrent gliomas. As expected, *MGMT* expression is significantly lower in HM samples compared to the alkylator-treated non-hypermutated glioma, while genes such as cyclin dependent kinase 2 (*CDK2*) and polo-like kinase 1 (*PLK1*) had significantly higher expression in HM (**Extended Data Fig. 4e**). The elevated *PLK1* level in HM glioma is compatible with results of a previous study that applied PLK1 inhibitors in MMR-deficient glioma cell lines[23]. Subsequent gene set enrichment analysis revealed MYC targets V2 as the most significantly up-regulated gene set (normalized enrichment score (NES) = 3.32, $P = 5.2 \times 10^{-4}$, **Extended Data Fig. 4f**) in hypermutated recurrent glioma. Other most upregulated pathways included DNA repair, E2F targets, G2M checkpoint, and MYC targets V1, implying HM samples had more active transcription and proliferation[24].

To identify the molecular features at diagnosis that drive the occurrence of HM at recurrence, we compared the transcriptome of initial gliomas that recurred as hypermutated (pre-HM, n=21) versus the initial gliomas that recurred as non-HM tumors (pre-NHM, n = 143). Mimicking the recurrent tumors, we found significant upregulation of MYC targets, G2M checkpoint, E2F targets and mitotic spindle pathways in pre-HM samples (Figure S5a-b). MYC pathway activation is known to be associated with cell proliferation. Indeed, the commonly used cell proliferation marker, MKI67, is significantly up regulated genes in pre-HM gliomas ($P = 1.54 \times 10^{-3}$, Wilcoxon's rank-sum test, **Extended Data Fig. 5c**). Immunohistochemistry staining of 67 gliomas showed that gliomas with high Ki-67 staining have significantly higher probability to develop HM after TMZ treatment ($P = 0.04$, Fisher's exact test, **Extended Data Fig. 5d**). *MGMT* was down-regulated in pre-HM samples, which is consistent with the observation that MGMT-methylated gliomas, when treated with TMZ, were prone to develop hypermutation ($P = 0.021$, Fisher's exact test, **Extended Data Fig. 5e**). Strikingly, comparing genomic alterations of pre-HM and pre-non-HM gliomas we found that genomic gain of *MYC* was the most enriched feature in the pre-HM tumors (fold change = 5.31, $P = 3.23 \times 10^{-5}$, Fisher's exact test), well in line with the upregulation of MYC pathways (**Fig. 3a-b**). Specifically, about 46.9% of *MYC*-gained gliomas developed HM after TMZ treatment, while the proportion in gliomas without *MYC* gain was only 12.7% (**Fig. 3c**). Following MYC gain, other genomic alterations enriched in initial tumors of hypermutators include *ATRX* mutation and *RB1* deletion (**Fig. 3a**). Loss-of-function mutations in *ATRX* have been shown to impair DNA repair, leading to genetically unstable tumors that rapidly accumulate oncogenic mutations [25]. RB1 encodes

Rb protein and plays an essential role as a cell-cycle regulator acting at the G1/S cell cycle checkpoint [26]. To make use of the comprehensive information for hypermutation prediction, machine learning models were trained using genomic features of initial glioma plus alkylator treatment information and achieved an Area Under Precision-Recall Curve (AUPRC) of 0.88. When the expression of *MGMT* and *MKI67* were included, the AUPRC further improved to 0.95 (**Fig. 3d**). SHAP analysis showed that alkylator treatment was the most predictive feature in the model, followed by *MYC* gain, *ATRX*, *FUBP1* and *IDH* mutations, and *EGFR* amplification (**Fig. 3e**).

To interrogate the association between MYC and HM in glioma, we grew a MYC-amplified glioma cell line U251 and a MYC-wild-type glioma cell line U87 in medium containing escalating concentrations of TMZ (**Extended Data Fig. 5f-g**), until they developed elevated TMZ resistance (TR; **Extended Data Fig. 5h**). During the whole process, U251 maintained its high copy number and high expression of *MYC* gene, while U87 had relatively low MYC expression (**Extended Data Fig. 5i**). Upon resistance, U87TR showed elevated *MGMT* expression (**Extended Data Fig. 5j**). On the contrary, U251TR maintained low level of *MGMT* expression but gained thousands of mutations. Over 90% of U251TR mutations were C>T (G>A) single nucleotide variants, including an *MSH6* T1219I hotspot mutation with allele frequency 38%, demonstrating the emergence of the TMZ-induced hypermutation *in vitro* (**Extended Data Fig. 5i-j**). Considering U87 and U251 might not perfectly represent glioma patients, we further carried out the TMZ inducing experiment in the newly derived gliomasphere (T2-4) sample from a patient with high MYC copy number and expression (**Fig. 3f**). This sample developed dramatic TMZ

resistance after the induction experiment (**Fig. 3g**). Interestingly, T2-4TR developed hypermutation with a stop codon mutation *MLH1* W597* (AF = 23%) and a splicing-donor mutation (MLH1 c.453+1G>A, AF = 32%, **Fig. 3h**), while maintaining a low level of *MGMT* expression and a high level of MYC copy number and expression. Notably, the splicing-donor mutation in *MLH1* caused an aberrant splicing which included an additional nine basepairs in the spliced RNA (**Fig. 3i**). We next tracked the development of hypermutator phenotype during T2-4 cell line passaging, and found the percent spliced in (PSI) of the aberrant isoform emerged at passage six and increased over time (**Fig. 3i**). Whole-exome sequencing of different passages demonstrated that hypermutation occurred almost simultaneously with loss-of-function mutations in MLH1 at the sixth-seventh passage of cell culture (**Fig. 3i-j**), confirming the underlying association between a burst of somatic mutation and the MMR deficiency in glioma cells.

Hypermutation occurs frequently in actively transcribed genomic regions with MYC binding

To explore the mechanism of how MYC activation promotes somatic hypermutation, we first characterized the genomic distribution of hypermutation sites in relation to epigenetic regulatory elements. The human genome was annotated using chromosomal binding peaks from chromatin immunoprecipitation sequencing (ChIP-seq) data (**Extended Data Fig. Table 3**, **Methods**), and then the mutational density of the TMZ-associated hypermutations (HM) and non-TMZ-associated conventional somatic mutations (NHM) were calculated and normalized for comparison. Consistent with previous observations that conventional somatic mutations in the cancer genomes were not randomly distributed but enriched in

transcriptionally inactive regions[27–31], we found remarkably higher density of NHM mutations around the H3K9me3 modification sites which are often associated with heterochromatin (**Fig. 4a**, upper panel). In contrast, the density of hypermutations in these regions were close to random. Conversely, while NHM mutations were significantly lower than expected in the open chromatin regions marked by DNase hypersensitivity, HM had remarkably higher mutation (**Fig. 4a**, lower panel). More importantly, the HM mutation was significantly enriched in the periphery of H3K27ac and H3K4me3 sites that respectively mark active enhancers and promoters, while NHM was of lower density (**Fig. 4b**, upper panel). In regions surrounding binding sites of RNA polymerase II (RPII), a key enzyme in the transcriptional machinery, the density of HM mutations was almost two-fold of NHM mutations (**Fig. 4b**, lower left panel). It has been shown that in normal cells, the actively transcribed genes recruit mismatch repair complex to protect the gene body from mutation by H3K36me3[32,33]. We investigated the mutation densities around H3K36me3 modification sites and noted that the mutation density of both HM and NHM mutations were lower than random control, but the density of HM was much higher than NHM (**Fig. 4b**, lower right panel). Altogether, these results demonstrated that in TMZ-associated hypermutators, mutations were enriched to the proximity of active transcription sites.

To explore the role of MYC gain in hypermutagenesis, we compared MYC gene expression between samples with and without MYC gain and found significantly higher MYC expression in gliomas with MYC gain ($P = 1.4\times10^{-7}$, Wilcoxon rank-sum test; **Extended Data Fig. 6a**). In tumor cells with elevated MYC expression, c-Myc forms a heterodimer with MAX and binds to the promoters and enhancers of active genes, causing global

transcription amplification[34]. To quantify the global transcriptional activity, we used fourteen housekeeping genes as reference and counted the number of high-expression genes (**Methods**). As expected, more high expression genes were observed in MYC-gain-positive gliomas ($P = 1.0 \times 10^{-3}$, Wilcoxon rank-sum test, **Fig. 4c**), indicating globally amplified transcription in these samples. Comparison of the expression level of MYC direct targets also showed consistent results ($P = 1.3 \times 10^{-4}$, Wilcoxon rank-sum test, **Extended Data Fig. 6b**).

We next divided the open chromatin regions and active transcription sites based on whether c-Myc also binds to the site, and then compared the density of HM mutations in these two types of regions (**Fig. 4d**). The density of HM mutations was higher in open chromatin with c-Myc binding compared to those without c-Myc binding. Similarly, when checking H3K27ac modification sites with and without c-Myc binding, we found the mutation density of HM to be much higher in sites with c-Myc binding than that in sites without c-Myc binding (**Fig. 4d**). The same phenomenon was observed for H3K4me3 modification sites, RPII binding sites and H3K36me3 modification sites with and without c-Myc binding (**Fig. 4d**).

c-Myc binds MMR genes and increases probability of hypermutation

The eight MMR genes, namely *PMS1, PMS2, MSH2, MSH3, MSH5, MSH6, MLH1* and *MLH3*, all had H3K4me3 modification, c-Myc/MAX and RPII binding, especially in the periphery of the promoter regions (**Fig. 4e**). In U87MG cell line, we experimentally induced MYC overexpression and confirmed that MYC overexpression upregulated the

253 expression of MSH6 and MSH2, the two most frequently altered MMR genes in HM (**Fig.**
254 **4f**, **Extended Data Fig. 6c**). Chi-square test with Yates' correction revealed significant
255 enrichment of MMR genes in c-Myc binding targets ($P = 0.0002$, **Fig. 4g**), a group which
256 was highly enriched for temozolomide-driven mutations. Therefore, we propose that MYC
257 over activation increased the density of TMZ-induced mutations in the genomic regions of
258 MMR genes. With an elevated risk of developing loss-of-function mutations in MMR
259 genes, MYC-amplified gliomas hereby tend to develop HM in response to TMZ treatment.
260 Mechanistically, the double strand DNA is opened during transcription, and MAX and c-
261 Myc binding at the active promoters prolonged the time window of opening status. Under
262 TMZ treatment, the drug molecules attack the single stranded DNA, causing more damages
263 in the transcription active regions. When such damages occurred in MMR genes and failed
264 for repair, mutations will accumulate in other open sites and finally lead to hypermutation
265 (**Fig. 4h**).

Discussion

The high rate of recurrence represents the major obstacle in improving survival of adult diffuse glioma patients. Here, we have assembled and analyzed a mostly original cohort of 511 cases with longitudinal matched samples of gliomas. This work highlighted the longitudinal trajectory of evolution for each of the three main subtypes of gliomas (IDH mutant codel, IDH mutant non-codel, IDH wild type). Through the development of novel computational approaches, we found that grade progression and gain of hypermutation are largely predictable from the molecular features of the initial glioma. Following training on clinical and genomic features, our machine learning model robustly distinguished LGGs that tend to progress to GBM from those that recur as LGG. Among the features with highest importance, *CDKN2A* deletion, a key cell-cycle related alteration has been correlated with aggressive disease and tumor enhancing.[8,35] Another important feature that emerged from the model was *TP53* copy-neutral loss of heterozygosity. Despite the importance and prevalence of TP53 inactivation in glioma, little is known about how TP53 loss of function alterations contribute to glioma malignancy. Other important contributing factors in the model include alkylator treatment and MYC gain, both of which are associated with hypermutation. The relevance of these alterations as drivers of glioma aggressiveness is underscored by the significant association between hypermutation and grade progression.

The most important finding that emerged from our work is the discovery of the gain-of-function genomic alterations of MYC as drivers of the HM state at recurrence. Several reports including our own suggested that therapy-associated hypermutations in recurrent

glioma are driven by MMR deficiency[9,13,16]. In our cohort, MMR deficiency was found in 86% of hypermutated gliomas at recurrence. Indeed, whereas MMR deficiency is the direct causal factor leading to hypermutation, the absence of mutations in MMR genes in untreated tumors suggested that other, upstream alterations might be necessary to set in motion the machinery leading to the hypermutated state. The key early event we uncovered, was gain of MYC, an event significantly detected in untreated glioma primed to recur as hypermutated tumors. The role of MYC is especially significant in the context of the IDH mutant non-codel subtype, which is also the glioma group with the highest likelihood to undergo transformation to the hypermutated state at recurrence. Through the integration of exome and transcriptomic analyses, we found that gain of MYC results in marked activation in downstream pathways associated with proliferation and deregulated cell cycle progression, which are the primary biological functions of MYC in mammalian cells. The role of MYC as unexpected driver of the hypermutation state has been experimentally validated in human glioma cells. Furthermore, by integrating genomic and epigenetic data, we confirmed our previous hypothesis that mutagenic mechanisms related to TMZ treatment and subsequent MMR alteration act more efficiently in highly expressed regions of open chromatin[16], and further showed that c-Myc binding augmented the chance of MMR mutations. Interestingly, highly expressed regions are protected by transcription-coupled repair (TCR)[36], and hypermutated tumors exhibited an enrichment of somatic mutations in TCR genes in the hypermutated glioma samples. However, whether the TCR pathway is functional in these samples will have to be established by future studies.

Collectively, this study painted the evolution routes of three molecular types of glioma, trained machine learning models to predict grade progression based on clinical and genomic features of initial glioma and demonstrated MYC as a potential predictor of hypermutation after treatment. In addition, we have developed an interactive, publicly available web resource (**Extended Data Fig. 7**; https://wanglab.shinyapps.io/cello/) to explore the longitudinal glioma dataset and make predictions of developing treatment-induced hypermutation and grade progression.

Acknowledgements

This work was supported by NSFC Excellent Young Scientists Fund (Hong Kong and Macau) (No. 31922088), RGC grants (N_HKUST606/17, 26102719), ITC grant (ITCPD/17-9), National Natural Science Foundation of China (NSFC)/Research Grants Council (RGC) Joint Research Scheme (No. 81761168038), National Natural Science Foundation of China (No. 81903078, No. 81972816), National Key Research and Development Project of China (No. 2019YFE0109400), Hong Kong Epigenomics Project (LKCCFL18SC01-E), and a grant from the Korea Health Technology R&D through the Korea Health Industry Development Institute funded by the Ministry of Health & Welfare, Republic of Korea (HI14C3418). The authors would like to thank all contributors to the Chinese Glioma Genome Atlas. In addition, the authors acknowledge data generators of the published datasets, especially TCGA research network, the GLASS Consortium, Gregory M. Kiez and Mehmet Kutman Foundation, and the Yale University Department of Neurosurgery for providing access to the raw and/or processed sequencing data.

Author Contributions

J.W., T.J., D.H.N., and A.I. conceptualized and supervised the project. Q.M. performed the computational and statistical analyses; R.C., H.L. and Y.C. carried out the cell line and PDC experiments; Y.Y. and Q.M. developed the CELLO webserver; Z.Z., Z.B., and T.J. contributed to sample preparation, genomic sequencing, and clinical data collection of the CGGA cohort; M.H.L., D.S. and B.J. contributed to some of the bioinformatics analyses; J.K.S., H.J.C., and D.H.N. contributed to sample preparation, genomic sequencing and clinical data collection of the SMC cohort; A.C., W.S.P., H.K.N., and D.M.C. collected the

CUHK samples; K.C., D.L, S.T, and A.W. prepared libraries for samples of the CUHK cohort; J.W. and Q.M. wrote the manuscript, which was then revised and proof-read by all authors.

Competing Interests

The authors declare no potential conflict of interests.

References

1. Greaves, M. & Maley, C. C. Clonal evolution in cancer. *Nature* **481**, 306–313 (2012).
2. McGranahan, N. & Swanton, C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. *Cell* (2017). doi:10.1016/j.cell.2017.01.018
3. Hu, H. *et al.* Mutational Landscape of Secondary Glioblastoma Guides MET-Targeted Trial in Brain Tumor. *Cell* **175**, 1665-1678.e18 (2018).
4. Lee, J.-K. *et al.* Spatiotemporal genomic architecture informs precision oncology in glioblastoma. *Nat. Genet.* **49**, 594–599 (2017).
5. Kim, J. *et al.* Spatiotemporal Evolution of the Primary Glioblastoma Genome. *Cancer Cell* **28**, 318–328 (2015).
6. Mazor, T. *et al.* DNA Methylation and Somatic Mutations Converge on the Cell Cycle and Define Similar Evolutionary Histories in Brain Tumors. *Cancer Cell* **28**, 307–317 (2015).
7. Suzuki, H. *et al.* Mutational landscape and clonal architecture in grade II and III gliomas. *Nat. Genet.* **47**, 458–468 (2015).
8. Barthel, F. P. *et al.* Longitudinal molecular trajectories of diffuse glioma in adults. *Nature* **576**, 112–120 (2019).
9. Johnson, B. E. *et al.* Mutational Analysis Reveals the Origin and Therapy-Driven Evolution of Recurrent Glioma. *Science (80-.).* **343**, 189–193 (2014).
10. Bai, H. *et al.* Integrated genomic characterization of IDH1-mutant glioma malignant progression. *Nat. Genet.* **48**, 59–66 (2016).
11. Wang, Q. *et al.* Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **32**, 42-56.e6 (2017).
12. Schwartzbaum, J. A., Fisher, J. L., Aldape, K. D. & Wrensch, M. Epidemiology and molecular pathology of glioma. *Nature Clinical Practice Neurology* **2**, 494–503 (2006).
13. Touat, M. *et al.* Mechanisms and therapeutic implications of hypermutation in gliomas. *Nature* **580**, 517–523 (2020).
14. Cipponi, A. *et al.* MTOR signaling orchestrates stress-induced mutagenesis, facilitating adaptive evolution in cancer. *Science* **368**, 1127–1131 (2020).
15. Oldrini, B. *et al.* MGMT genomic rearrangements contribute to chemotherapy resistance in gliomas. *Nat. Commun.* **11**, 1–10 (2020).
16. Wang, J. *et al.* Clonal evolution of glioblastoma under therapy. *Nat. Genet.* **48**, 768–776 (2016).
17. Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nat. Mach. Intell.* **2**, 56–67 (2020).
18. Lundberg, S. M. & Lee, S.-I. A Unified Approach to Interpreting Model Predictions. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) **30**, 4765–4774 (Curran Associates, Inc., 2017).
19. Bai, H. *et al.* Integrated genomic characterization of IDH1-mutant glioma malignant progression. *Nat. Genet.* **48**, 59–66 (2016).
20. Jiang, B., Song, D., Mu, Q. & Wang, J. CELLO: a longitudinal data analysis toolbox untangling cancer evolution. *Quant. Biol.* **8**, 256–266 (2020).

21. van Thuijl, H. F. *et al.* Evolution of DNA repair defects during malignant progression of low-grade gliomas after temozolomide treatment. *Acta Neuropathol.* **129**, 597–607 (2015).
22. Mathur, R. *et al.* MGMT promoter methylation level in newly diagnosed low-grade glioma is a predictor of hypermutation at recurrence. *Neuro. Oncol.* (2020). doi:10.1093/neuonc/noaa059
23. Higuchi, F. *et al.* PLK1 inhibition targets MYC-activated malignant glioma cells irrespective of mismatch repair deficiency–mediated acquired resistance to temozolomide. *Mol. Cancer Ther.* **17**, 2551–2563 (2018).
24. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Syst.* **1**, 417–425 (2015).
25. Koschmann, C. *et al.* ATRX loss promotes tumor growth and impairs nonhomologous end joining DNA repair in glioma. *Sci. Transl. Med.* **8**, 328ra28-328ra28 (2016).
26. Bertoli, C., Skotheim, J. M. & De Bruin, R. A. M. Control of cell cycle transcription during G1 and S phases. *Nature Reviews Molecular Cell Biology* **14**, 518–528 (2013).
27. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
28. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
29. Gonzalez-Perez, A., Sabarinathan, R. & Lopez-Bigas, N. Local Determinants of the Mutational Landscape of the Human Genome. *Cell* **177**, 101–114 (2019).
30. Makova, K. D. & Hardison, R. C. The effects of chromatin organization on variation in mutation rates in the genome. *Nature Reviews Genetics* **16**, 213–223 (2015).
31. Akdemir, K. C. *et al.* Somatic mutation distributions in cancer genomes vary with three-dimensional chromatin structure. *Nat. Genet.* **52**, 1178–1188 (2020).
32. Huang, Y., Gu, L. & Li, G. M. H3K36me3-mediated mismatch repair preferentially protects actively transcribed genes from mutation. *J. Biol. Chem.* **293**, 7811–7823 (2018).
33. Li, F. *et al.* The histone mark H3K36me3 regulates human DNA mismatch repair through its interaction with MutSα. *Cell* **153**, 590–600 (2013).
34. Lin, C. Y. *et al.* Transcriptional amplification in tumor cells with elevated c-Myc. *Cell* **151**, 56–67 (2012).
35. Jonsson, P. *et al.* Genomic correlates of disease progression and treatment response in prospectively characterized gliomas. *Clin. Cancer Res.* **25**, 5537–5547 (2019).
36. Hanawalt, P. C. & Spivak, G. Transcription-coupled DNA repair: Two decades of progress and surprises. *Nature Reviews Molecular Cell Biology* **9**, 958–970 (2008).
37. Kim, H. *et al.* Whole-genome and multisector exome sequencing of primary and post-treatment glioblastoma reveals patterns of tumor evolution. *Genome Res.* **25**, 316–327 (2015).
38. Ceccarelli, M. *et al.* Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell* **164**, 550–563 (2016).
39. Zhao, J. *et al.* Immune and genomic correlates of response to anti-PD-1

immunotherapy in glioblastoma. *Nat. Med.* **25**, 462–469 (2019).
40. Hu, H. *et al.* Mutational Landscape of Secondary Glioblastoma Guides MET-Targeted Trial in Brain Tumor. *Cell* **175**, 1665-1678.e18 (2018).
41. Lee, J.-K. *et al.* Pharmacogenomic landscape of patient-derived tumor cells informs precision oncology therapy. *Nat. Genet.* **50**, 1399–1411 (2018).
42. Trifonov, V., Pasqualucci, L., Tiacci, E., Falini, B. & Rabadan, R. SAVI: A statistical algorithm for variant frequency identification. *BMC Syst. Biol.* **7**, S2 (2013).
43. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6**, 80–92 (2012).
44. Sherry, S. T. *et al.* dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.* **29**, 308–11 (2001).
45. Tate, J. G. *et al.* COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res.* **47**, D941–D947 (2019).
46. 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 (2016).
47. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
48. Haas, B. J. *et al.* Accuracy assessment of fusion transcript detection via read-mapping and de novo fusion transcript assembly-based methods. *Genome Biol.* **20**, 213 (2019).
49. Wang, J. *et al.* Tumor evolutionary directed graphs and the history of chronic lymphocytic leukemia. *Elife* **3**, (2014).
50. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016). doi:10.1145/2939672.2939785

Figure and legend

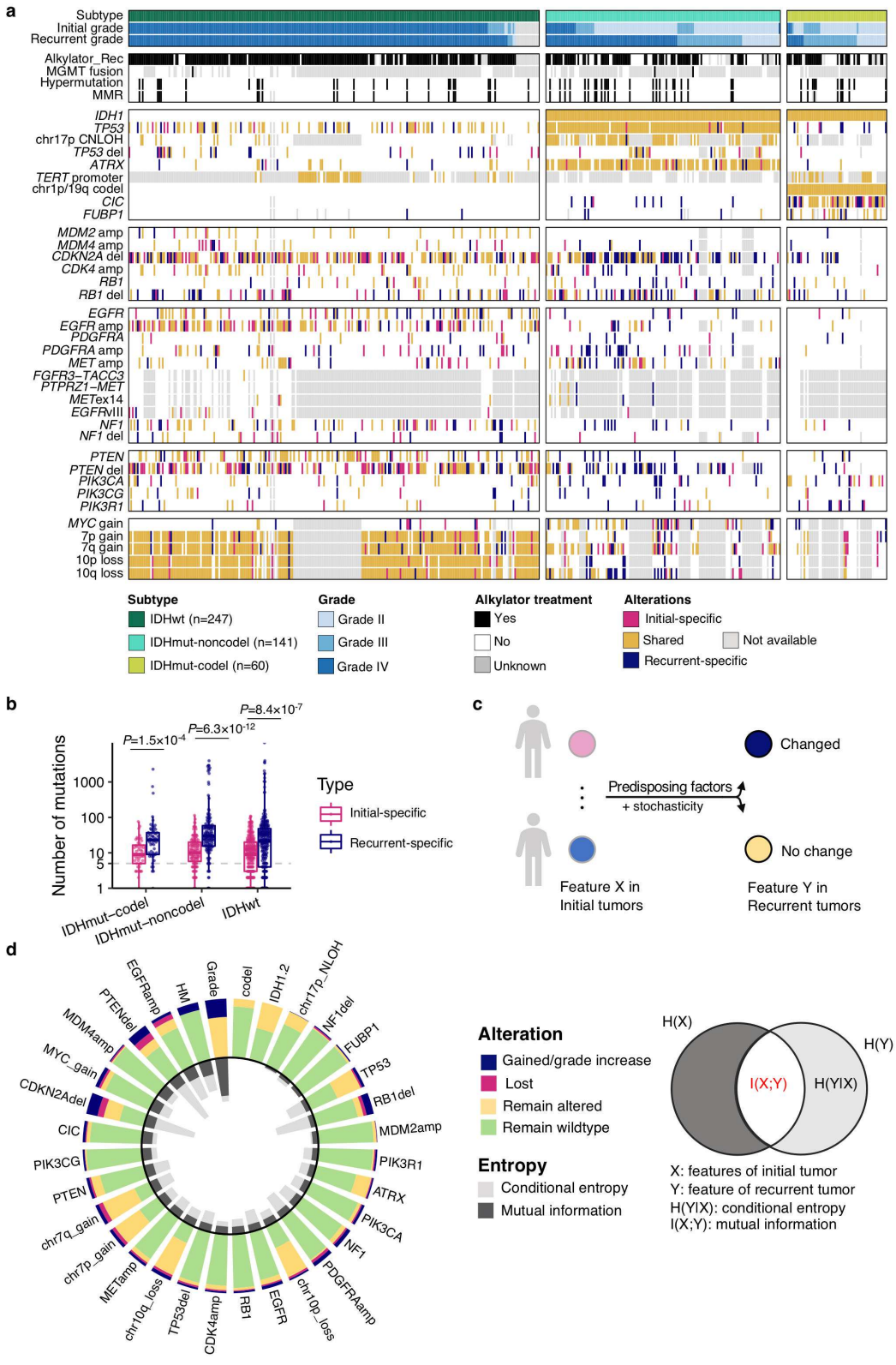
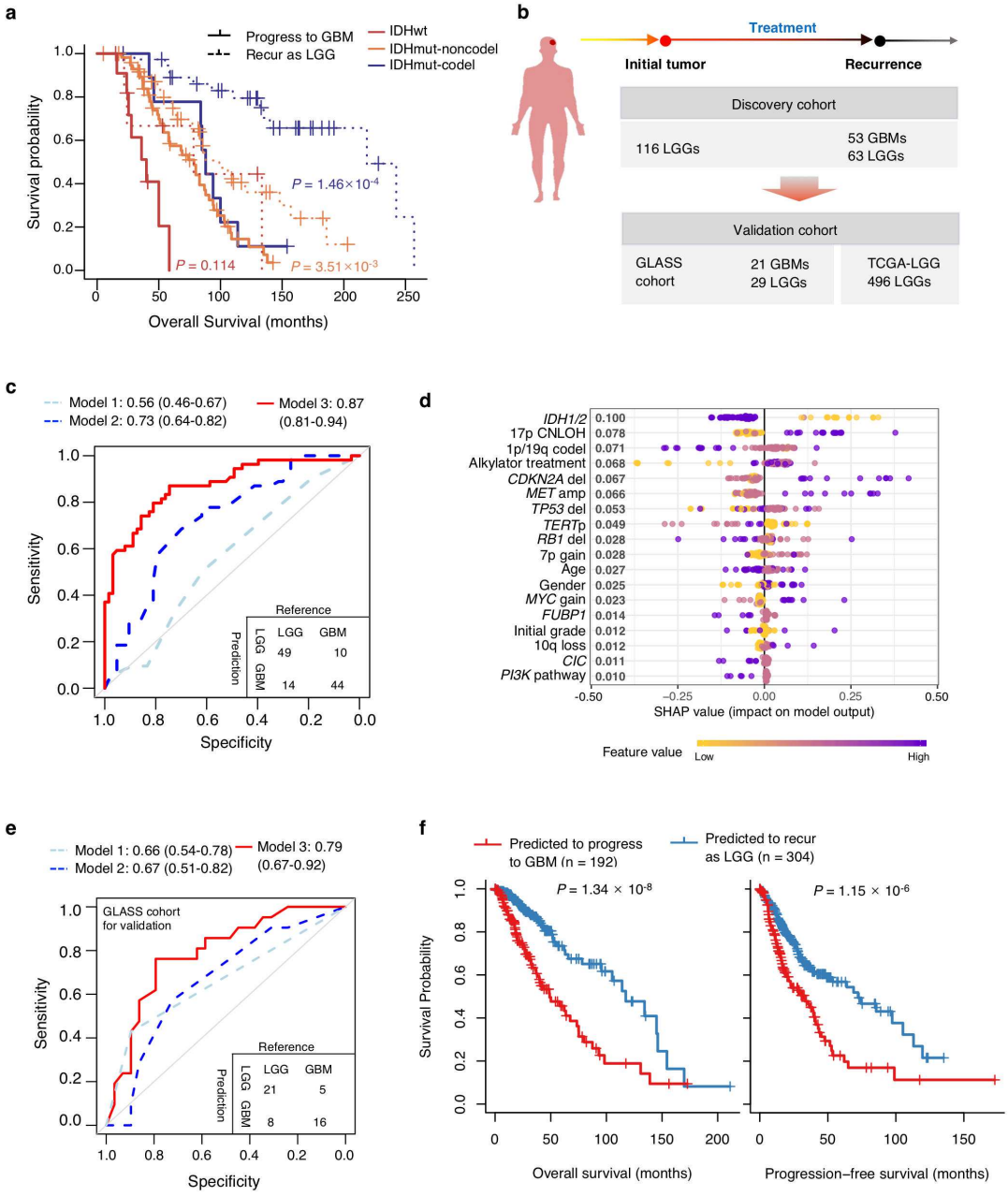


Figure 1. Identification of predictable events from pan-glioma evolutionary landscape. **(a)** Longitudinal molecular landscape of glioma patients with paired DNA sequencing data. Each column represents a patient which were stratified by the molecular subtype, and each row represents a clinical feature or molecular variant. The color of each variant indicates whether it is specific to initial, recurrence or shared. MMR: mismatch repair; CNLOH: copy-neutral loss of heterozygosity. **(b)** Comparison of the number of initial- and recurrent-specific somatic mutations during evolution in each glioma subtype. The horizontal dashed line represents five mutations. The *P* values were calculated by Wilcoxon's rank-sum test. **(c)** Proposed model explaining the changes of the clinical and molecular features over time. Predisposing factors in the initial tumor, together with stochasticity, determine whether the status of the feature will change at recurrence. **(d)** Alteration status, entropy and mutual information of clinical and genomic features. Stacked bars outside the circle summarizes the frequency of a clinical feature or molecular variant to be lost, gain, remaining wildtype or remaining altered. Inside the circle shows the overlapped bar plots of mutual information (black) and conditional entropy (grey), where high proportion of overlap indicates better predictability. The features are ranked by mutual information.



487

488 **Figure 2. Genomic characteristics of initial glioma informs grade progression after**

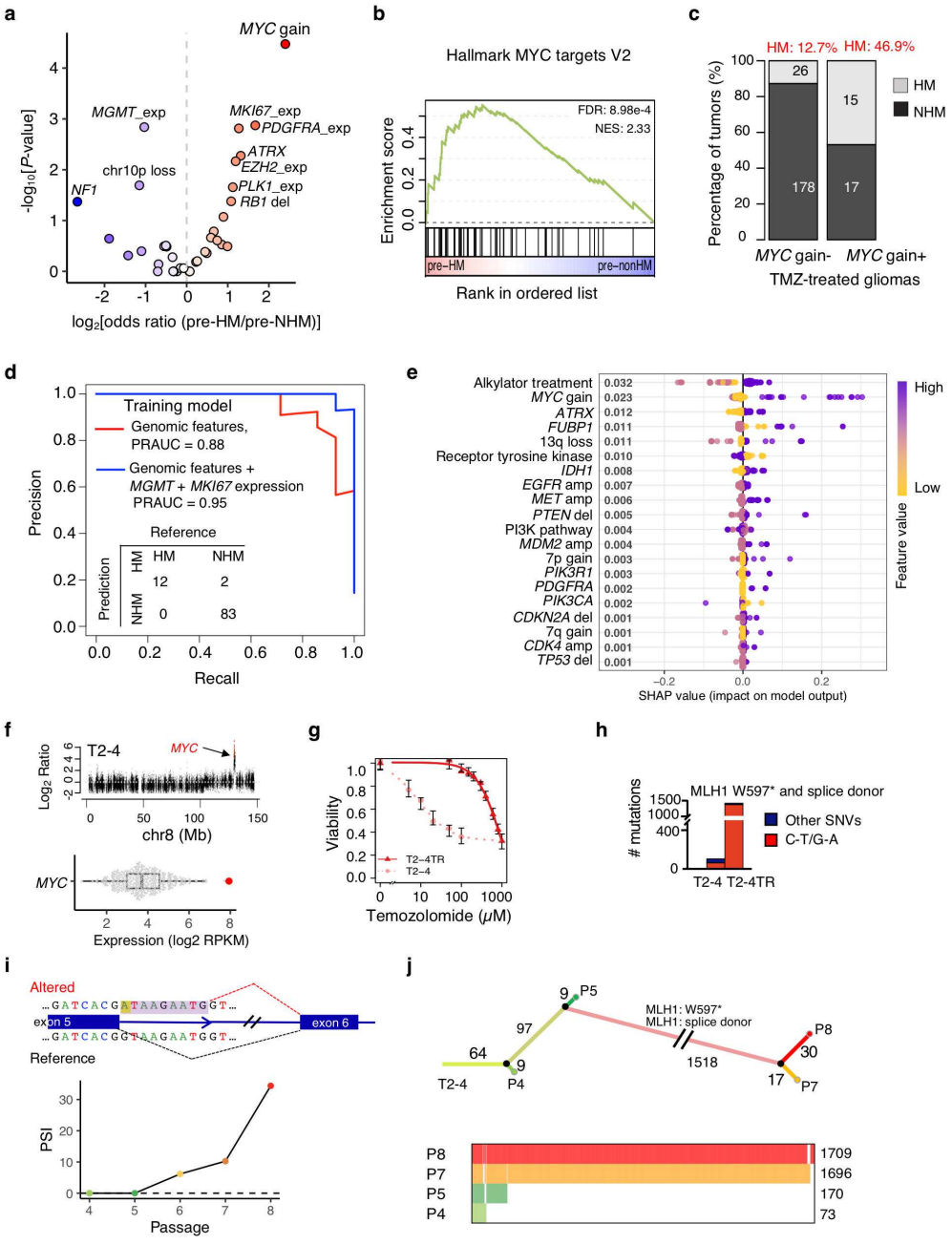
489 **treatment.**

490 **(a)** Overall survival of low-grade gliomas that progressed to glioblastoma (GBM) and those

491 that recurred as low-grade glioma (LGG). The colors represent glioma subtypes, while the

492 line type (solid or dashed) show whether the patient progressed to GBM at recurrence. The

493 *P* values were calculated by log-rank test. **(b)** Design of the machine learning tasks. The
494 discovery cohort, including 117 LGGs, was used to train machine learning models to
495 predict whether the initial LGG patients would progress to GBM at recurrence. The models
496 were then applied to two independent cohorts for validation. **(c)** Receiving operation
497 characteristic (ROC) curves showing the cross-validation performance of the models
498 trained using different combinations of features. Model 1: the initial grade as the single
499 predictor; Model 2: the initial grade, age, gender, and treatment as predictors; Model 3: the
500 initial grade, age, gender, treatment, and genomic alterations as predictors. **(d)** Shapley
501 additive explanation (SHAP) scores of features in the prediction model of grade
502 progression. The color of the points represents value of each feature. **(e)** ROCs showing
503 the three models' prediction performance in the GLASS cohort. A confusion matrix is
504 shown at the right bottom. **(f)** Overall survival (left panel) and progression-free survival
505 (right panel) of LGGs from TCGA-LGG cohort, stratified by their predicted risk of
506 progressing to GBM. Patients that were in this longitudinal cohort were excluded. The *P*
507 values were calculated by log-rank test.



508

509 **Figure 3. Features of initial tumors predispose TMZ-associated hypermutation at**

510 **recurrence**.

511 **(a)** Comparison of genomic and transcriptomic features of the TMZ-treated initial gliomas

512 preceding hypermutators (pre-HM) and initial tumors preceding non-hypermutators (pre-

513 NHM). **(b)** Gene set enrichment analysis. Gene expression profile of pre-HM gliomas were

compared to pre-NHM gliomas. FDR: false-discovery rate; NES: normalized enrichment score. **(c)** The frequency of developing hypermutation in TMZ treated gliomas with or without *MYC* copy number gain. **(d)** Precision-recall curve of HM prediction the model trained using genomic features (red curve) and genomic features plus expression of *MGMT* and *MKI67* (blue curve). **(e)** SHAP value of each feature in the model trained with genomic features plus expression of *MGMT* and *MKI67* in (c). **(f)** The DNA copy number of *MYC* in T2-4 (upper panel), and the expression level (lower panel) of T2-4 (the red dot) and other RNA sequenced gliomas in this study (grey dots). **(g)** Response curves of T2-4 and the induced T2-4TR cell line to TMZ. **(h)** Number of somatic mutations in T2-4 and T2-4 TR. Dark red represents C to T or G to A mutations, while dark blue represents other somatic mutations. **(i)** Aberrant splicing related to the MLH1 splicing donor mutation in T2-4. The upper panel shows the aberrant splicing which included nine additional base pairs in the spliced mRNA. The lower panel shows the percentage of spliced-in (PSI) reads of the aberrant isoform in different passages of T2-4. **(j)** Mutations in T2-4 at different time points during the TMZ-inducing experiment. The upper panel shows the phylogenetic tree constructed from the mutations in passages 4, 5, 7 and 8 as compared to the untreated T2-4 PDC, and the lower panel shows a heatmap of the mutations.

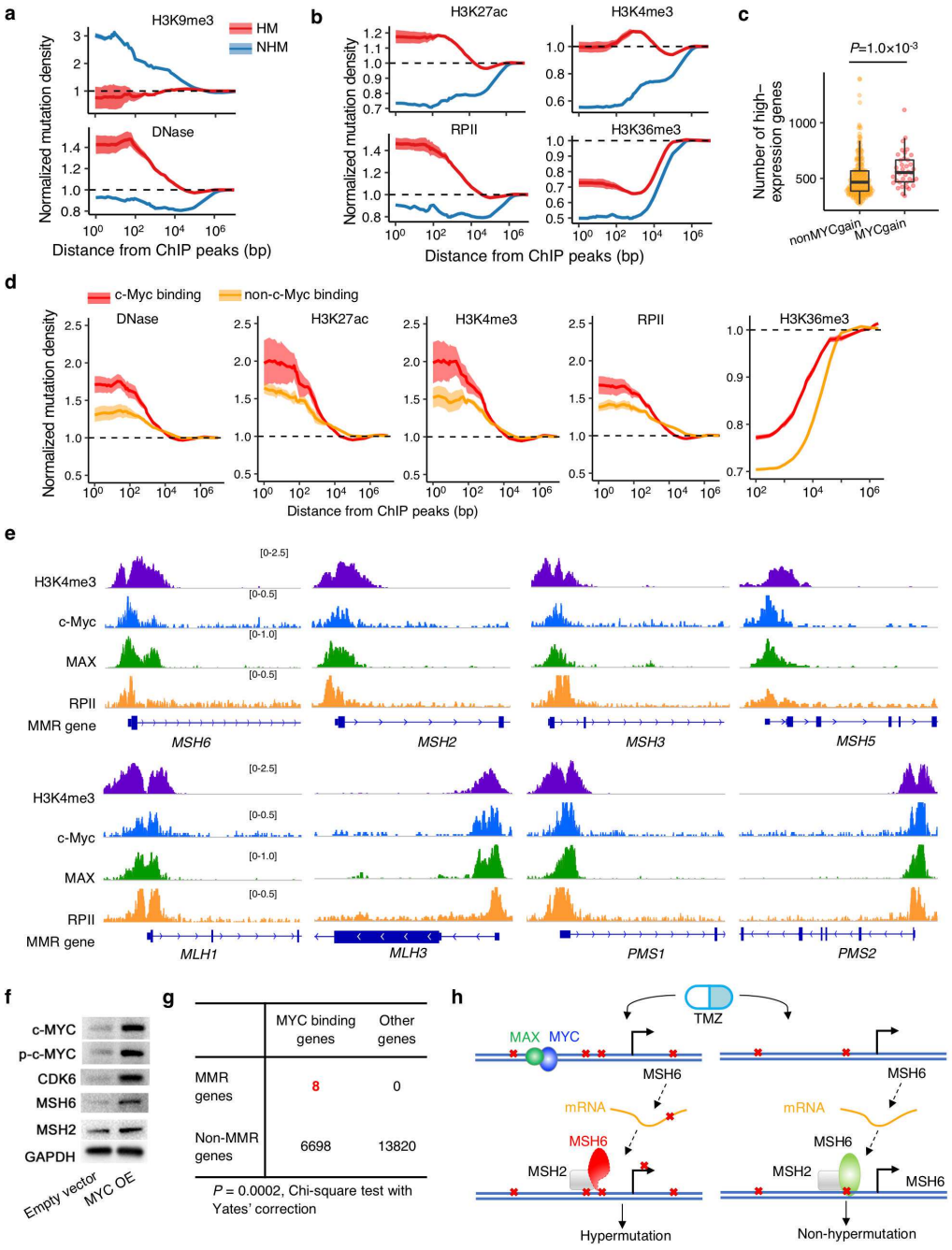


Figure 4. TMZ-induced mutations are enriched in open active regions, while c-Myc binding further increases mutation density

(a) Normalized mutation density of TMZ-associated hypermutations (HM) and non-TMZ associated recurrence-specific somatic mutations (NHM) around H3K9me3 modification sites (upper panel) and DNase hypersensitive sites (lower panel). The dashed horizontal

line represents the random level. **(b)** Normalized mutation density of HM and NHM around H3K27ac modification sites (upper left), H3K4me3 modification sites (upper right), RNA polymerase II (RPII) binding sites (lower left) and H3K36me3 modification sites (lower right). **(c)** Number of high-expression genes in gliomas with and without MYC gain. The *P* value was calculated by Wilcoxon's rank-sum test. **(d)** Normalized mutation density of HM around DNase hypersensitive sites, H3K27ac modification sites, H3K4me3 modification sites, RPII binding sites and H3K36me3 modification sites with (red color) and without (orange color) c-Myc binding. **(e)** Chromatin immunoprecipitation (ChIP) intensity of H3K4me3 modification, c-Myc, MAX and RPII binding at eight mismatch repair (MMR) genes. **(f)** Contingent table of c-Myc binding and MMR genes. **(g)** Western blot showing c-MYC, p-c-MYC, CDK6, MSH6, MSH2, and MLH1 in U87 cell lines with empty vector or MYC overexpression. OE: overexpression. **(h)** Proposed working model of how MYC activation increases chance of hypermutation.

Methods

Human Patients and Samples

In this study we collected longitudinally paired glioma from at least two time points, namely the initial tumor and the recurrent tumor. Where possible, blood samples were also collected. The newly collected samples are from Capital Medical University Beijing Tiantan Hospital (China), Samsung Medical Center (SMC, Korea) and CUHK Otto Wong Brain Tumour Centre in Hong Kong.

Glioma patients from Beijing Tiantan Hospital were enrolled as part of the Chinese Glioma Genome Atlas project (CGGA, http://cgga.org.cn). Ethical approval was obtained from the institutional review board in Capital Medical University (IRB ID: KYSQ2019-200-01), and informed consent for each patient was obtained before surgery. For each specimen, the pathological diagnosis is confirmed independently by two neuropathologists based on the 2016 edition of WHO classification of central nervous system tumors. The specimen was flash-frozen within 5 mins after resection and restored in liquid nitrogen until being used for DNA/RNA extraction and other experiments.

The SMC cohort includes both published cases and newly enrolled samples. The study is approved by the institutional review board of SMC (IRB file No. 2010-04-004 & 2005-04-001), and written consent is obtained from each patient. The glioma specimens were immediately snap-frozen after surgical resection and stored in in liquid nitrogen for further analysis. This study also included samples from six patients with recurrent glioma from CUHK Otto Wong Brain Tumour Centre. Ethical approval was obtained from the Joint

Chinese University of Hong Kong-New Territories East Cluster Clinical Research Ethics Committee with Reference No. 2018.389.

In addition to these newly collected specimens, we also curated sequencing data of longitudinally paired glioma from published cohorts. This include the Istituto Neurologico C. Besta cohort (INCB, n=19)[16], the MD Anderson Cancer Center cohort (MD Anderson, n=10)[37], the Kyoto University cohort (KU, n=10)[7], the University of California San Francisco cohort (UCSF, n=23)[9], The Cancer Genome Atlas cohort (TCGA, n=27)[38], the Yale cohort (n=41)[19], the Columbia University cohort (Columbia, n=14)[39], the Memorial Sloan Kettering Cancer Center cohort (MSKCC, n=67)[35]. The original Glioma Longitudinal AnalySiS (GLASS) cohort[8] included some samples from the above collection, so we included the rest 139 patients for which sequencing data of both initial and recurrent tumors were available and designated as GLASS cohort.

Sample processing and Illumina sequencing

For the samples from Tiantan hospital, DNA and/or RNA from the tumor were extracted using the same protocol as described previously[15,40]. Where available, DNA from matched blood was also extracted. For patients PS115-PS132, the exome regions of both tumor and blood DNA were captured using the Agilent SureSelect XT Human All ExonV5 kit, and then sequenced using Illumina HiSeq 2500 platform. For patients PS132-PS157, whole genome libraries were prepared and sequenced using Illumina HiSeq 2500 platform, generating about 180Gb and 90Gb 150 bp paired end reads for tumor and normal DNA, respectively. For RNA, the extracted total RNA was firstly depleted for tRNA and rRNA,

then reverse-transcribed to cDNA, and sequenced using Illumina HiSeq 2500 platform, generating around 6 Gb 150 bp paired end reads for each sample.

Samples from SMC were processed and sequenced as previously described[41]. For samples from CUHK, both total DNA and RNA were isolated from tumor samples while only total DNA was purified from blood samples. Frozen tumors were homogenized in Buffer RLT (Qiagen) with 40mM DTT and subsequently filtered by QIAshredder (Qiagen). Total DNA/RNA from tissue lysate and from blood were then isolated using AllPrep DNA/RNA Mini kit (Qiagen) according to manufacturer's protocol. The quantity and purity of the nucleic acids were determined using Qubit 3 Fluorometer with either dsDNA HS Assay Kit or RNA HS Assay Kit. For DNA samples with sub-optimal purity determined by A260/280 ratio, they were further cleaned up using DNA Clean & Concentrator-5 kit (ZYMO Research). RNA integrity was assessed by RNA Quality Number (RQN) using Fragment Analyzer (Advanced Analytical) with HS RNA Kit (15NT). DNA samples with a A260/280 of ~1.8, and RNA samples with a A260/280 of ~2.0 and RQN > 5 were subjected to sequencing. Exome DNA were subsequently captured using Agilent SureSelect XT Human All ExonV5 kit. DNA and RNA were sequenced using the same protocol as samples from Tiantan Hospital, but generated 24Gb reads for tumor DNA, 12 Gb reads for blood DNA, and 12 Gb reads for RNA.

For the published samples, a detailed list of the sequencing types and protocols were summarized in Table S1.

Sequencing data preprocessing and mapping

The quality of the sequencing reads was first checked by running FastQC v0.11.5. The low-quality reads (containing at least one ambiguous base, or average quality <20) were then removed using fastp v0.20.1. For DNA sequencing, the clean reads were mapped to the human reference genome hg19 using Burrows Wheeler Aligner (BWA, v0.7.15-r1140) mem algorithm, then sorted by coordinates using samtools v1.2. Subsequently, duplicate reads were marked by running Picard MarkDuplicates 2.9.2 tool (https://broadinstitute.github.io/picard/), generating the final bam file for mutation and copy number identification. For RNA-seq, the reads were mapped to hg19 using STAR 2.6.1d. The gene annotation is ENSEMBL GRCh37.75.

For published samples, sequencing reads were either downloaded from NCBI SRA or extracted from the published bam files that we downloaded from EGA, and then realigned to hg19 using the same parameters as the newly sequenced samples.

Identification of somatic mutations in longitudinal samples

Somatic mutations in the newly sequenced samples were identified using SAVI[42]. Briefly, raw mutation calls as compared to the normal sample were generated from samtools 1.2 mpileup, and then annotated using SNPEff for their impacts[43]. SNPSIFT was used to add further annotation such as whether the mutations are common SNPs (per dbSNP[44]), cancer related somatic mutations (per COSMIC[45]), or observed in normal patients. A statistical test was performed for each mutation call to determine the significance of being a somatic mutation. SAVI is able to analyze mutations from multiple samples in one single run, which is especially useful for longitudinal studies.

For the published data, we identified somatic mutations using the same pipeline as the newly sequenced ones when raw sequencing data were available. For the YALE, MSKCC and GLASS cohort, we retrieved mutation call information (including mutation loci, reference and alterative alleles, reference and alternative depth) from individual sources, annotated with SAVI using the same parameters, and then applied the same filters to identify somatic mutations. Mutations in known glioma driver genes were reported.

Copy number alteration detection

Copy number variations in the newly sequenced samples were detected using CNVkit 0.9.5[46], which is suitable for both WES and WGS data. In both cases, only the exon regions were considered. The BED file containing information about the captured regions was downloaded from Agilent website (https://earray.chem.agilent.com/suredesign/index.htm). The final results were segmented into SEG files.

Detection of genes fusions, *MET*ex14 and *EGFR*vIII from RNA-seq data

RNA sequencing data from previous publications were downloaded, and the reads were extracted using samtools 1.2[47]. Starting from FASTQ files, STAR-fusion 1.5.0[48] was utilized to identify and annotate gene fusion candidates, using the fastq files as input. The fusion candidates were then filtered by removing fusions that were present in normal tissues, fusions involving mitochondria genes and uncharacterized genes, and fusions of two paralog genes.

*MET*ex14 and *EGFR*vIII were identified by counting the spanning reads over the junction of *MET* exon 13 and exon 15, and *EGFR* exon 1 and exon 8, respectively. Briefly, RNA sequencing reads were aligned to the reference genome (hg19), and then the spanning reads were extracted based on the CIGAR record. The spanning reads were manually checked in Integrative Genomic Viewer to remove false positives such as PCR artifacts and potential mapping errors. Only samples with at least three supporting reads were considered as positive.

Integration of mutation profiles

For samples from the Yale, MSKCC and GLASS cohorts, all the mutations were obtained from the original publications, which contains information about the coordinates of the mutations, the reference and alternative alleles, as well as the read count of the reference and alternative alleles. All the mutations were reannotated using SAVI, and the same mutation filter as those used in the new samples were applied. This unified mutation annotation and filtering method ensures the results from multiple cohorts comparable.

The copy number alteration calls from all samples were first transformed to SEG format which included the coordinate of each segment and the estimated segment mean value (reflecting the magnitude of copy number change). Subsequently, each segment was annotated for the carried genes using the reference human genome GRCh37.75. The same cutoffs for copy number gain, amplification and deletion were applied.

For gene fusions, where available, the raw sequencing reads were downloaded or extracted, and then subjected to gene fusion detection using the same protocol.

Construction of Tumor Evolution Directed Graphs

Tumor Evolution Directed Graphs[49] were constructed using CELLO[20]. The input to CELLO was the mutation calls from SAVI, and the samples were organized in the normal-initial-recurrence order. CELLO automatically extracted the shared, private to initial and private to recurrence mutations in each patient, which will be used to construct the tree. The evolution trees were visualized in Cytoscape.

Calculation of entropy and mutual information

For each patient, an alteration might be gained or lost during evolution, both were counted as one "status change event". If the alteration was present at both time points, or absent at both time points, it is counted one "status stable event". For a given alteration X, suppose the number of status change events in the cohort is *m*, and the number of "status stable event" is *n*, then the entropy is

$$H(X) = -\frac{m}{m+n} log_2\left(\frac{m}{m+n}\right) - \frac{n}{m+n} log_2\left(\frac{n}{m+n}\right).$$

The mutual information between two alterations, X and Y, is calculated as below:

$$I(X,Y) = H(X) + H(Y) - H(X,Y).$$

In our analysis, the maximum number of features was set to four. All the possible combinations were enumerated to find the four features with the highest mutual information.

Machine learning model training, validation, and feature importance

XGBoost was used to train and test the models[50]. XGboost is a boosting machine learning methodology, supporting influential supervised classification tasks. The dataset was

divided into training set and test set two times for the prediction of HM and Grade with the ratio of 3:1. For the training dataset, we leveraged five-fold cross validation to select the optimal parameters, which were applied to the final configuration of XGBoost with best validation AUC (Area Under receiver operating characteristic Curve). Then, the independent test dataset was imported into the model to evaluate the performance and generalization ability of the model for unseen data. The model training and testing was implemented in R using the *xgboost* package, and SHAP (SHapley Additive exPlanations) score[17] was extracted from the XGBoost models using *SHAPforxgboost* package.

Survival analysis

Survival analysis was performed using the *survival* package in R. The significance levels were calculated by two-sided log-rank test, and the Kaplan-Meier curves were plotted using *survminer* package.

Identification of hypermutation

Hypermutation was detected using CELLO, with default parameters. CELLO determines the hypermutation status based on mutation load and a hypermutation score which we defined previously[20]. For the MSKCC cohort which includes panel sequencing data, the mutation load reported in the publication was directed used, while the hypermutation score was calculated using the same methodology as other samples.

Gene expression analysis

Differential gene expression analysis was performed by the DEseq package 1.26.0. Gene Set Enrichment Analysis was performed by the *fgsea* package (version 1.12.0) in R 3.6.3. The gene sets regarding human cancer were gene sets which were downloaded from MSigDB database version 6.1 (http://software.broadinstitute.org/gsea/msigdb). To quantify global transcription activity, the mean expression level of 14 housekeeping genes: ACTB, GAPDH, PGK1, PPIA, RPL13A, RPLP0, B2M, YWHAZ, SDHA, TFRC, GUSB, HMBS, HPRT1 and TBP in each sample was calculated and the genes with expression levels higher than this reference value was designated as "high-expression" genes.

ChIP-seq data acquisition and intersection with hypermutation loci

ChIP-seq data (listed in Table S3), including both binding density and binding peaks, were downloaded from ChIP-Atlas (https://chip-atlas.org/). The peaks were extended by 1-5,000,000 bp upstream and downstream, and the overlaps, regions in the centromeres, telomeres or out of chromosome borders were removed.

Recurrent-specific mutations from WGS data were divided into MYC-gained HM, non-MYC-gained-HM and non-HM mutations, and randomly down-sampled to 150,000 for each type. The sampling process was repeated for five times. The mutation sites were transformed to BED format, and their intersection with the ChIP peaks were calculated using bedtools v2.26.0. The mutation density was calculated by dividing the total number of mutations with the total width of the regions. The mean and standard deviation of the five replicates were calculated.

Cell cultures

The human glioma cell line U87 and U251were obtained from the Cell Resource Center, Peking Union Medical College (Beijing, China), and the cell has been authenticated by the short tandem repeat analysis Chronic temozolomide treatment cells. Both the two cells were routinely cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% fetal bovine serum (HyClone, Logan, Utah), 100 units/ml penicillin and 100 mg/ml streptomycin (Invitrogen, Carlsbad, CA) at 37℃ in a humidified atmosphere of 5% CO_2. The PDC T2-4, was obtained from fresh surgical specimens of a human primary GBM and cultured as tumor spheres in DMEM/F12 medium supplemented with B27 supplement (Life Technologies), bFGF and EGF (20 ng/ml each) at 37℃ in a humidified atmosphere of 5% CO_2.

Chronic temozolomide treatment of glioma cell lines

The human glioma cell line U87MG and U251MG were obtained from the Cell Resource Center, Peking Union Medical College (Beijing, China). U87MG and U251MG cells were treated with TMZ on Initial concentration of 100 μM for 72h. TMZ-sensitive cells generated death, and the survived cells are collected for further resistant induction with increasing TMZ concentration (≤200 μM). After more than 20 generations and 6 months later, each cell line's cells survived and growth in DMEM culture medium containing TMZ of final concentration of 200 μM, were referred as U87TR and U251TR cells.

The PDCs were treated with TMZ on an initial concentration of 50 μM for 5 days, and the remain cells were incubated in medium with increasing concentration of TMZ until to 200 μM. The IC50 values of TMZ were determined in the following situations: 1) after 6

generations of growth in medium with 200 μM TMZ; 2) the PDCs reached the status that they have similar growth rates in normal medium and medium containing 200 μM TMZ. Finally, the PDCs were considered TMZ-resistant when they have significantly elevated IC50 values compared with their corresponding initial cells.

Temozolomide sensitivity test

The temozolomide sensitivity were studied by CCK-8 kit (Dojindo Laboratories, Kumamoto, Japan) according to the manufacturer's protocol. 3000 cells were grown in each well of 96-well plates with normal medium for 24hr. Then the cells were incubated at medium containing TMZ at predesigned concentrations for 72 hrs, and there were six-repeated wells for each predesigned concentrations. Next, 10 μl CCK-8 reagent was added to the medium of each well, and the absorbances at 450 nm and 630 nm of the medium were measured after incubation for 2 hours. The number of living cells could be reflected by the difference between optical density of 450 (OD_{450}) and 630 (OD_{630}) nm wavelength. The response curves were fitted using the *drc* package 3.0-1.

Western blot

The whole-cell lysates were using RIPA buffer [150 mM NaCl, 0.1% (wt/vol) NP-40, 50 mM Tris (pH 8.0), 0.5% (wt/vol) sodium deoxycholate,1% (wt/vol) sodium dodecyl sulfate, 1 mM dithiothreitol, 0.1 mM phenylmethylsulfonyl fluoride]. A bicinchoninic acid (BCA) array was used to measure the total protein content. The equal total proteins (20 μg) were boiled and then were electrophoresed on a 8% and 12% sodium dodecyl sulfate–polyacrylamide gel electrophoresis gel, respectively. After membrane transfer and

blocking, the primary antibodies and used to probe the target proteins in blocking solution overnight at 4℃. The membrane were detected using an ECL Western Blotting Detection System (Bio-Rad) after Goat anti-rabbit IgG-HRP or goat anti-mouse IgG-HRP incubation 1 hr at room temperature. Glyceraldehyde 3-phosphate dehydrogenase (GAPDH) was used as the loading control.

Generation of lentiviral vectors encoding c-MYC

The sequencing of c-MYC (NM_002467) was cloned from glioma samples. The sequence was inserted into the lentiviral vector plasmid GV358 (Genechem). The GV358 plasmid carrying the c-MYC CDS sequencing was transduced into 293T cells with the lentiviral packaging plasmid mix (Genechem). The culture medium was collected once a day on two consecutive days and centrifuged with an ultracentrifuge at 25,000 rpm for 1.5 h. The precipitate was resuspended, aliquoted and stored at -80℃. This preparation was added to the U87 cell culture or PDCs supplemented with 3 mg/ml polybrene (Sigma) in the medium. The medium was changed after 24 h. After 72 h, cells containing vectors were selected during 3 days of treatment with puromycin. The c-MYC expression in the U87 cells and PDCs were confirmed by Western blot.

Validation of MSH6 mutation by Sanger sequencing

DNA extraction and PCR were performed to get the fragment including the MHS6 mutation site identified by the whole exon sequencing, the primers sequences were as follows: forward 5'-CTCCCCATGGGCTGCTAAG-3', Reverse 5' -TATGTCCTAGGCGCACAGC-3'. The product bands were extracted from agarose gel after electrophoresis and verified by Sanger sequencing with the forward primer.

Code availability

The hypermutation status, and the Tumor Evolution Directed Graphs were determined using CELLO which is available at https://github.com/WangLabHKUST/CELLO. All other scripts are available from the authors upon reasonable request.

Data availability

The processed genomic and transcriptomic data are available in the CELLO2 website (https://wanglab.shinyapps.io/cello). Raw sequencing data of the newly sequenced samples will be deposited to European Genome-Phenome Archive (EGA, https://ega-archive.org/). Data from previous SMC samples were available in EGA (https://www.ebi.ac.uk/ega/datasets/EGAS00001001800), data from TCGA were downloaded from NCI Genomics Data Commons (GDC) data portal (https://portal.gdc.cancer.gov). Previously published CGGA data have been uploaded to the Genome Sequence Archive in BIG Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession number BioProject ID: PRJCA001636 (https://bigd.big.ac.cn/bioproject/browse/PRJCA001636) and PRJCA001747 (https://bigd.big.ac.cn/bioproject/browse/PRJCA001747). All the other data supporting the findings of this study are available within the article and its information files and from the corresponding author upon reasonable request.

Additional information

Supplementary Information is available for this paper. Correspondence and requests for materials should be addressed to Jiguang Wang (jgwang@ust.hk) or Tao Jiang (taojiang1964@163.com).

Supplementary item titles and legends

Supplementary Table S1: Clinical characteristics of the longitudinal glioma cohort, **related to Figure 1**.

Supplementary Table S2: Sequencing platforms used for the samples in the longitudinal glioma cohort, **related to Figure 1**.

Supplementary Table S3: List of ChIP-seq data used in hypermutation distribution analysis, **related to Figure 4**.

Figures

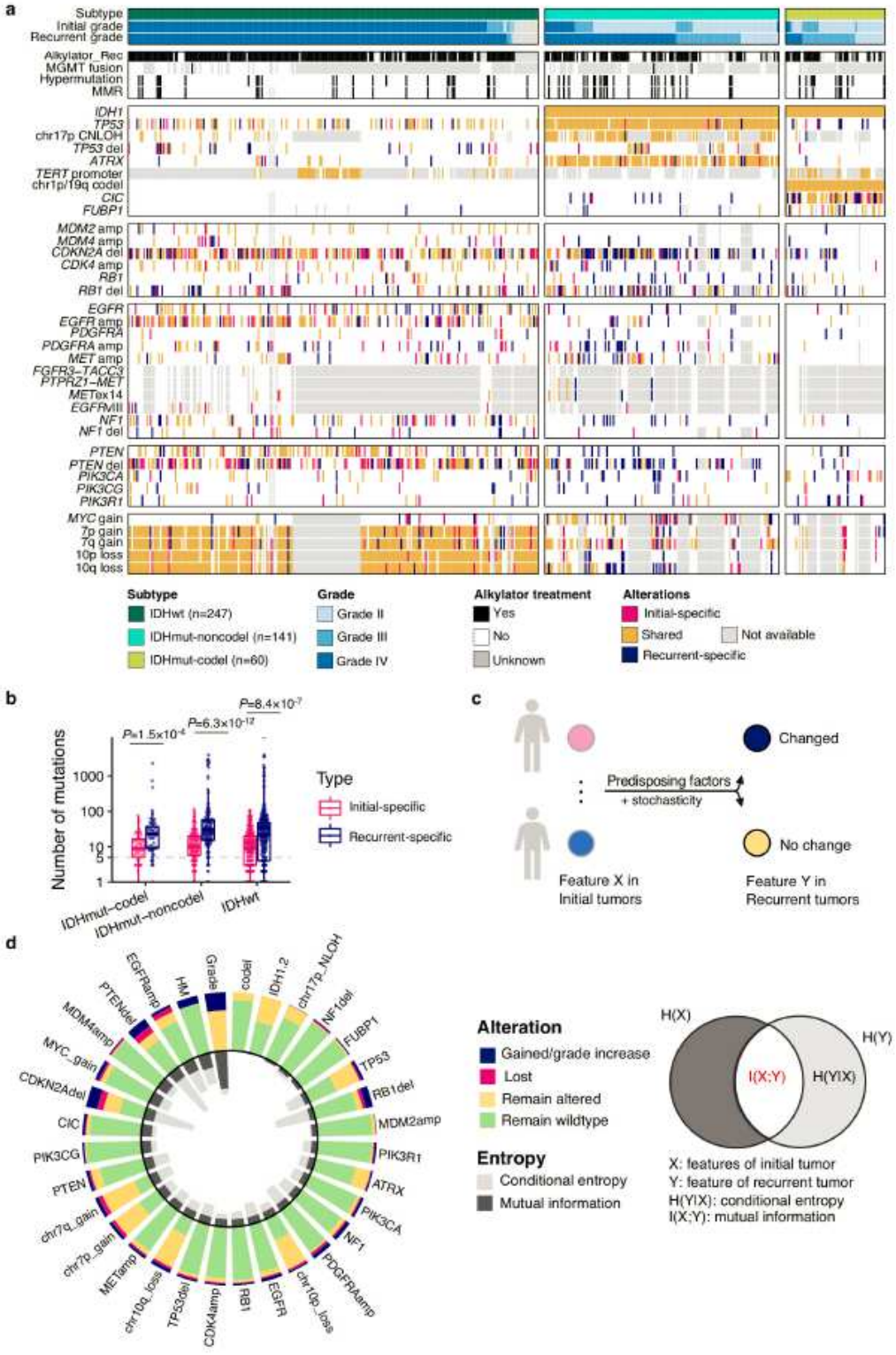


Figure 1

Identification of predictable events from pan-glioma evolutionary landscape. (a) Longitudinal molecular landscape of glioma patients with paired DNA sequencing data. Each column represents a patient which were stratified by the molecular subtype, and each row represents a clinical feature or molecular variant.

The color of each variant indicates whether it is specific to initial, recurrence or shared. MMR: mismatch repair; CNLOH: copy-neutral loss of heterozygosity. (b) Comparison of the number of initial- and recurrent specific somatic mutations during evolution in each glioma subtype. The horizontal dashed line represents five mutations. The P values were calculated by Wilcoxon's rank-sum test. (c) Proposed model explaining the changes of the clinical and molecular features over time. Predisposing factors in the initial tumor, together with stochasticity, determine whether the status of the feature will change at recurrence. (d) Alteration status, entropy and mutual information of clinical and genomic features. Stacked bars outside the circle summarizes the frequency of a clinical feature or molecular variant to be lost, gain, remaining wildtype or remaining altered. Inside the circle shows the overlapped bar plots of mutual information (black) and conditional entropy (grey), where high proportion of overlap indicates better predictability. The features are ranked by mutual information.

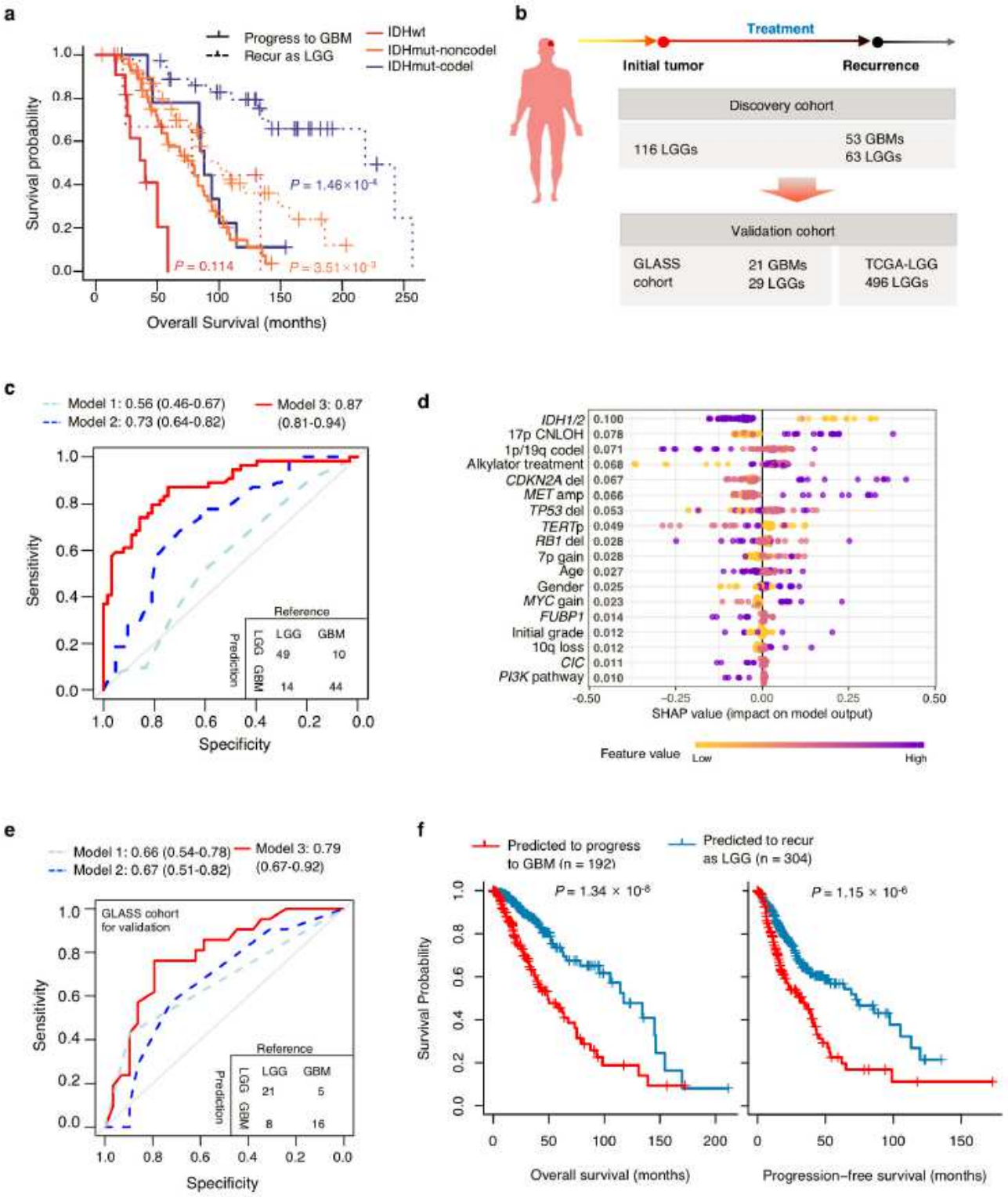


Figure 2

Genomic characteristics of initial glioma informs grade progression after treatment. (a) Overall survival of low-grade gliomas that progressed to glioblastoma (GBM) and those that recurred as low-grade glioma (LGG). The colors represent glioma subtypes, while the line type (solid or dashed) show whether the patient progressed to GBM at recurrence. The P values were calculated by log-rank test. (b) Design of the machine learning tasks. The discovery cohort, including 117 LGGs, was used to train machine learning

models to predict whether the initial LGG patients would progress to GBM at recurrence. The models were then applied to two independent cohorts for validation. (c) Receiving operation characteristic (ROC) curves showing the cross-validation performance of the models trained using different combinations of features. Model 1: the initial grade as the single predictor; Model 2: the initial grade, age, gender, and treatment as predictors; Model 3: the initial grade, age, gender, treatment, and genomic alterations as predictors. (d) Shapley additive explanation (SHAP) scores of features in the prediction model of grade progression. The color of the points represents value of each feature. (e) ROCs showing the three models' prediction performance in the GLASS cohort. A confusion matrix is shown at the right bottom. (f) Overall survival (left panel) and progression-free survival (right panel) of LGGs from TCGA-LGG cohort, stratified by their predicted risk of progressing to GBM. Patients that were in this longitudinal cohort were excluded. The P values were calculated by log-rank test.

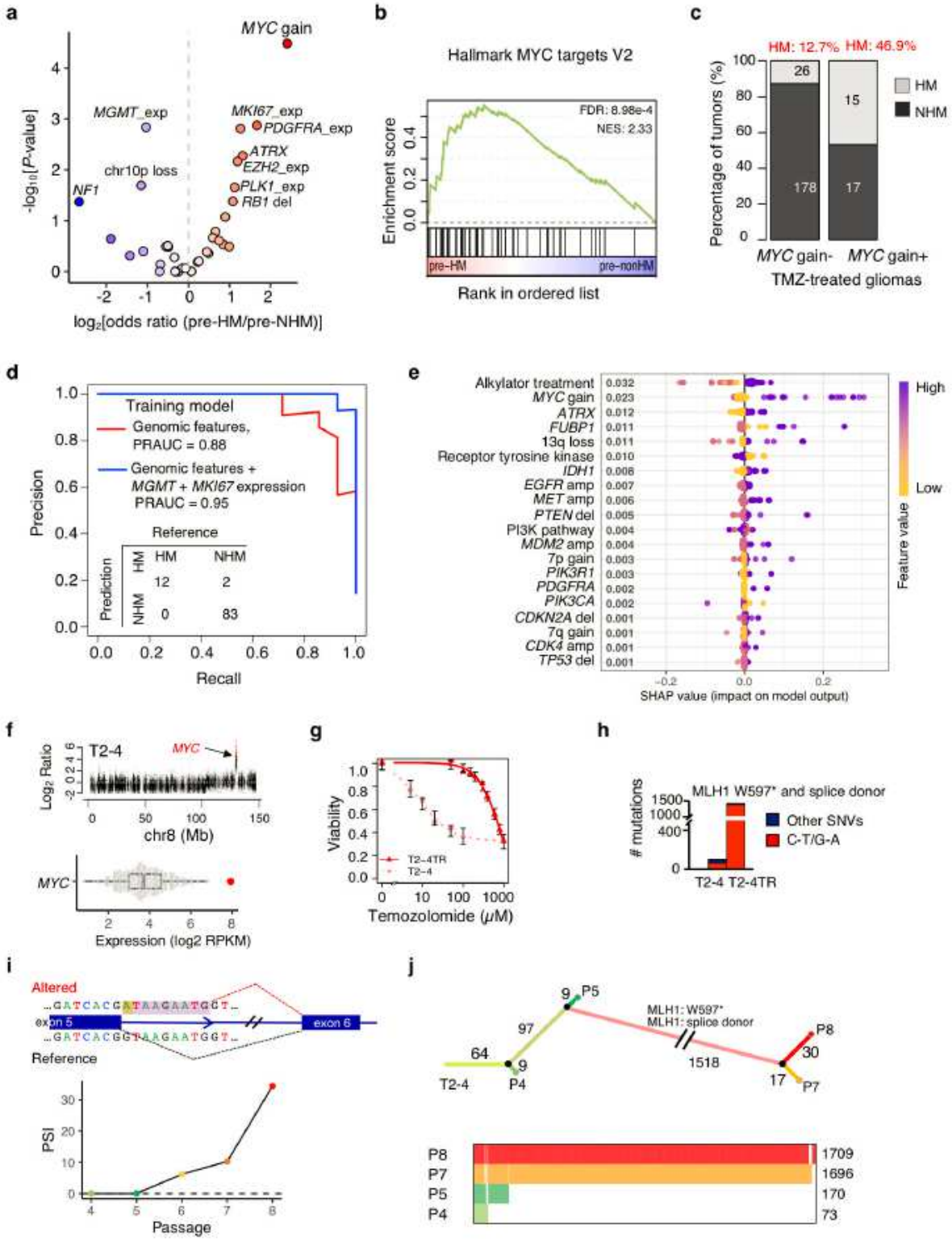


Figure 3

Features of initial tumors predispose TMZ-associated hypermutation at recurrence. (a) Comparison of genomic and transcriptomic features of the TMZ-treated initial gliomas preceding hypermutators (pre-HM) and initial tumors preceding non-hypermutators (pre- NHM). (b) Gene set enrichment analysis. Gene expression profile of pre-HM gliomas were compared to pre-NHM gliomas. FDR: false-discovery rate; NES: normalized enrichment score. (c) The frequency of developing hypermutation in TMZ treated gliomas

with or without MYC copy number gain. (d) Precision-recall curve of HM prediction the model trained using genomic features (red curve) and genomic features plus expression of MGMT and MKI67 (blue curve). (e) SHAP value of each feature in the model trained with genomic features plus expression of MGMT and MKI67 in (c). (f) The DNA copy number of MYC in T2-4 (upper panel), and the expression level (lower panel) of T2-4 (the red dot) and other RNA sequenced gliomas in this study (grey dots). (g) Response curves of T2-4 and the induced T2-4TR cell line to TMZ. (h) Number of somatic mutations in T2-4 and T2-4 TR. Dark red represents C to T or G to A mutations, while dark blue represents other somatic mutations. (i) Aberrant splicing related to the MLH1 splicing donor mutation in T2-4. The upper panel shows the aberrant splicing which included nine additional base pairs in the spliced mRNA. The lower panel shows the percentage of spliced-in (PSI) reads of the aberrant isoform in different passages of T2-4. (j) Mutations in T2-4 at different time points during the TMZ-inducing experiment. The upper panel shows the phylogenetic tree constructed from the mutations in passages 4, 5, 7 and 8 as compared to the untreated T2- 4 PDC, and the lower panel shows a heatmap of the mutations.

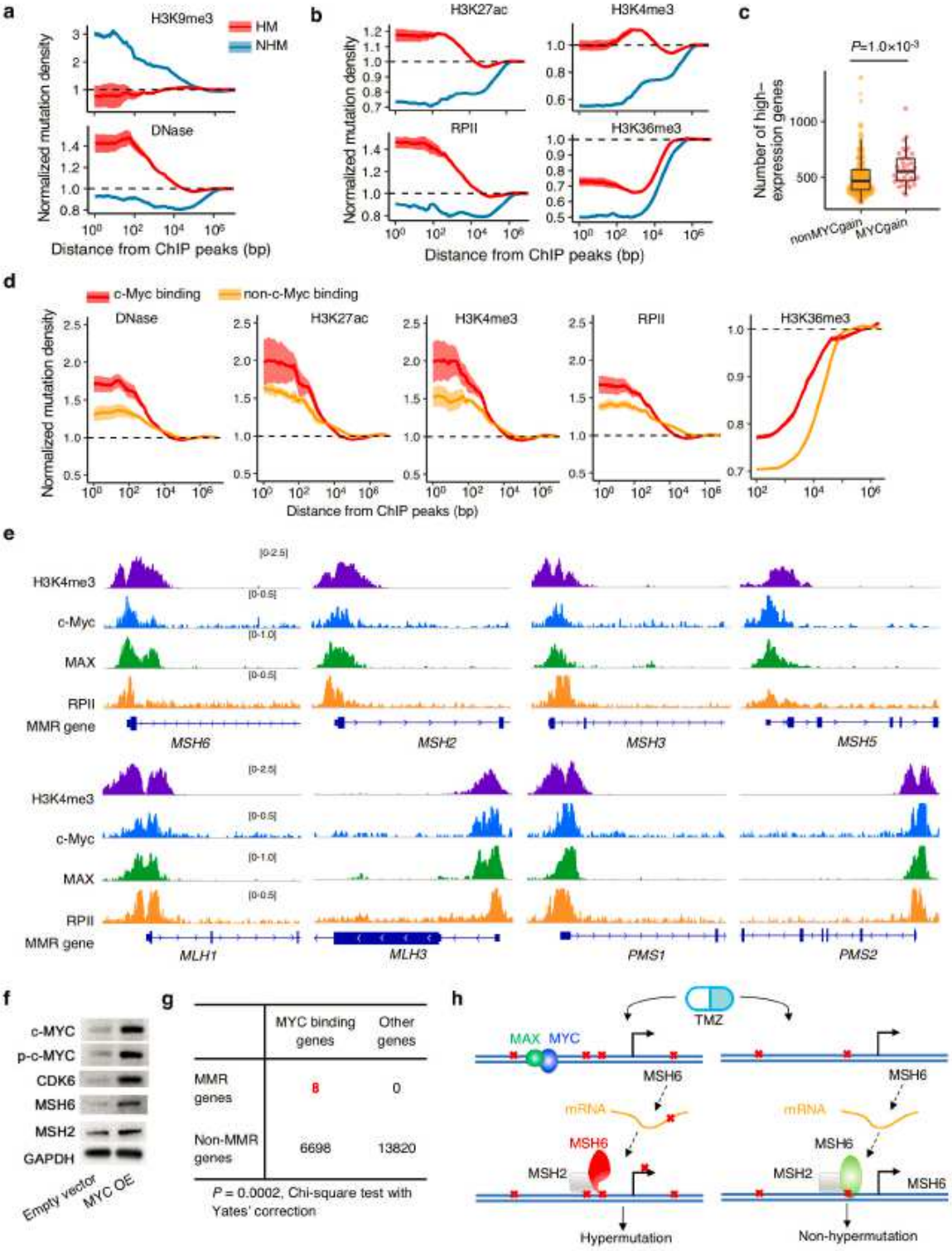


Figure 4

TMZ-induced mutations are enriched in open active regions, while c-Myc binding further increases mutation density (a) Normalized mutation density of TMZ-associated hypermutations (HM) and non-TMZ associated recurrence-specific somatic mutations (NHM) around H3K9me3 modification sites (upper panel) and DNase hypersensitive sites (lower panel). The dashed horizontal line represents the random level. (b) Normalized mutation density of HM and NHM around H3K27ac modification sites

(upper left), H3K4me3 modification sites (upper right), RNA polymerase II (RPII) binding sites (lower left) and H3K36me3 modification sites (lower right). (c) Number of high-expression genes in gliomas with and without MYC gain. The P value was calculated by Wilcoxon's rank-sum test. (d) Normalized mutation density of HM around DNase hypersensitive sites, H3K27ac modification sites, H3K4me3 modification sites, RPII binding sites and H3K36me3 modification sites with (red color) and without (orange color) c-Myc binding. (e) Chromatin immunoprecipitation (ChIP) intensity of H3K4me3 modification, c-Myc, MAX and RPII binding at eight mismatch repair (MMR) genes. (f) Contingent table of c-Myc binding and MMR genes. (g) Western blot showing c-MYC, p-c-MYC, CDK6, MSH6, MSH2, and MLH1 in U87 cell lines with empty vector or MYC overexpression. OE: overexpression. (h) Proposed working model of how MYC activation increases chance of hypermutation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- TableS1.Patient.ClinicalInformation.xlsx
- TableS2.Sequencing.Data.xlsx
- SupplementaryInformationDec30.docx