Neuro-Oncology
19(6), 786–795, 2017 | doi:10.1093/neuonc/now285 | Advance Access date 8 March 2017

Multigene signature for predicting prognosis of patients with 1p19q co-deletion diffuse glioma

Xin Hu, Emmanuel Martinez-Ledesma, Siyuan Zheng, Hoon Kim, Floris Barthel, Tao Jiang, Kenneth R. Hess, and Roel G.W. Verhaak

Department of Genomic Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas (X.H., E.M-L., S.Z., K.H., F.B., R.G.W.V.); Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas (K.R.H.); Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas (R.G.W.V.); Program of Bioinformatics and Biostatistics, The University of Texas-Houston Graduate School of Biomedical Sciences, Houston, Texas (X.H.); Department of Neurosurgery, Beijing Tiantan Hospital, Capital Medical University, Beijing, China (T.J.); Jackson Laboratory for Genomic Medicine, Farmington, Connecticut (K.H., F.B., R.G.W.V.)

Corresponding Author: Roel G.W. Verhaak, PhD, The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, Connecticut 06032 (roel.verhaak@jax.org).

Abstract

Background. Co-deletion of 1p and 19q marks a diffuse glioma subtype associated with relatively favorable overall survival; however, heterogeneous clinical outcomes are observed within this category.

Methods. We assembled gene expression profiles and sample annotation of 374 glioma patients carrying the 1p/19q co-deletion. We predicted 1p/19q status using gene expression when annotation was missing. A first cohort was randomly split into training (n = 170) and a validation dataset (n = 163). A second validation set consisted of 41 expression profiles. An elastic-net penalized Cox proportional hazards model was applied to build a classifier model through cross-validation within the training dataset.

Results. The selected 35-gene signature was used to identify high-risk and low-risk groups in the validation set, which showed significantly different overall survival (P = .00058, log-rank test). For time-to-death events, the high-risk group predicted by the gene signature yielded a hazard ratio of 1.78 (95% confidence interval, 1.02–3.11). The signature was also significantly associated with clinical outcome in the The Cancer Genome Atlas (CGA) IDH-mutant 1p/19q wild-type and IDH-wild-type glioma cohorts. Pathway analysis suggested that high risk was associated with increased acetylation activity and inflammatory response. Tumor purity was found to be significantly decreased in high-risk IDH-mutant with 1p/19q co-deletion gliomas and IDH-wild-type glioblastomas but not in IDH–wild-type lower grade or IDH-mutant, non–co-deleted gliomas.

Conclusion. We identified a 35-gene signature that identifies high-risk and low-risk categories of 1p/19q positive glioma patients. We have demonstrated heterogeneity amongst a relatively new glioma subtype and provided a stepping stone towards risk stratification.

Key words

elastic net Cox regression model | glioma | 1p/19q co-deletion | prognostic factor

According to current guidelines for brain tumors, the diagnosis of grade II-III adult diffuse glioma is assessed primarily by histopathological examination,[1] while molecular abnormalities have been evolving as supportive markers for facilitating diagnosis and management of these patients. Diffuse gliomas with mutations in the isocitrate dehydrogenase genes (*IDH1/IDH2*) may represent an entirely different type of disease than those with wild-type *IDH1/IDH2* (known as IDH–wild-type) glioma.[2–4] Within the group of IDH-mutant gliomas, presence of 1p/19q co-deletion (IDH-mutant–codel glioma) may

Importance of the study

Recent molecular studies of adult diffuse glioma have identified somatic alterations that function as biomarkers for molecular glioma subtypes including presence of IDH mutations and joint chromosome arm 1p and 19q co-deletion. The 2016 update to the World Health Organization (WHO) proposed a classification strategy based on traditional histopathology but enriched with IDH and 1p/19q co-deletion status. Co-deletion of 1p/19q was first recognized in the early 1990s, and recent clinical trials have associated this prognostic glioma subtype marker with treatment response. However, clinical outcomes in patients with 1p/19q co-deleted glioma may vary. Our study provides insights into the biology of therapy-response heterogeneity. We propose a risk index requiring prospective validation that may be useful for establishing prognosis.

present an additional prognostic marker separate from IDH-mutant glioma with intact 1p/19q chromosome arms (IDH-mutant–non-codel glioma). The unique characteristics of IDH-mutant–codel glioma led to recognition of this subtype in the 2016 WHO Classification of Tumors of the Central Nervous System.[5] A series of clinical trials revealed that standard radiation therapy followed by adjuvant chemotherapy with procarbazine, lomustine, and vincristine (PCV) delayed disease progression and increased overall survival (OS) in patients diagnosed with anaplastic oligodendroglioma. Interestingly, patients harboring 1p/19q co-deletion were more likely to respond to additional chemotherapy than those whose tumor was 1p and 19q wild-type.[6–9] Approximately 85% of diffuse gliomas with the 1p/19q co-deletion in the TCGA cohort have an oligodendroglial component and could be classified as either grade II (low-grade) or grade III (high-grade) glioma according to WHO criteria. Patients with histologically and molecularly similar glioma may reveal heterogeneous clinical characteristics and responses to treatment, which suggest that additional factors may determine clinical behavior and prognosis. The management of low-grade diffuse glioma (including the components necessary for diagnosis), the role of surveillance, and the nature of surgical intervention, radiation therapy, and chemotherapy (lacking conclusive evidence to support best practice) remain controversial.[10] Phenotypic and genomic intertumor heterogeneity of 1p/19q co-deleted gliomas may contribute to the lack of consistency between clinical observations.[10–12] Understanding the biological components associated with clinical and phenotypic heterogeneity will aid improved disease staging before treatment and tailoring of appropriate therapeutic regimens.

Molecular markers such as *IDH1/2* mutation, promoter methylation of *MGMT*, *EGFR* and *ATRX* genes mutations, and *BRAF* fusion or point mutation are increasingly recognized as integral aspects in the clinical management of patients with adult diffuse glioma.[13] There may be a role for molecular markers in risk classification of 1p/19q co-deletion glioma patients. High-risk patients could receive aggressive treatment with adjuvant chemotherapy, whereas low-risk patients might forgo intensive adjuvant chemotherapy. Several independent studies have demonstrated that gene expression profiling can be applied to identify biomarkers and molecular subtypes of glioma associated with certain clinical outcomes.[14,15] However, the prognostic profiles these studies identified have few genes in common, and the reported gene signatures are based on survival information and gene expression patterns from histopathological glioma classes. Such gene signatures might not accurately predict survival for patients whose glioma harbor the 1p/19q co-deletion because the mRNA expression patterns and underlying biological characteristics of this subgroup may be intrinsically different from those gliomas without the 1p/19q co-deletion as implied by its distinct favorable clinical outcomes.[16]

In an integrative analysis of newly diagnosed diffuse glioma patients, we observed that glioma patients harboring the 1p/19q co-deletion exhibit heterogeneous clinical outcomes.[2,4] In the present study, we sought to identify molecular markers associated with the diverse clinical outcomes in this subset of glioma patients. All datasets included in our study were previously published and patients were consented as described.

Methods

Datasets

Our approach to perform gene signature selection and validation for classification using normalized gene expression datasets is summarized in Fig. S1. We first curated gene expression and sample information from 5 publicly available glioma datasets whose tumors were assessed by microarray[14,17,18] or RNA-Seq,[2,19] (summarized in Table S1). Normalized RNA-Seq by Expectation-Maximization (RSEM) values for The Cancer Genome Atlas (TCGA) glioma samples were retrieved from the LGG-GBM project data portal (https://tcga-data.nci.nih.gov/docs/publications/lgggbm_2015). Reads Per Kilobase of transcript per Million mapped reads (RPKM) values for CGGA1 RNA-seq data[19] were calculated using in-house software (PRADA).[20] We used Affymetrix Human Genome U133 Set annotation provided by the Bioconductor library hgu133plus2.db and hgu133a.db, and Illumina HumanHT-12 WG-DASL to convert microarray probe signals to gene expression levels. Multiple probe sets were mapped to a single gene by averaging the signals. We also curated and combined 2 gene expression datasets measured by microarray and used this as a second validation dataset[21,22] (Table S1). Affymetrix CEL files in training and first validation dataset (but not the second validation dataset) were normalized together. OS time was defined as the time from diagnosis to death; the patients who were alive at the end of each study period

were censored at the date of last follow-up. Tumor grade was also established at primary diagnosis.

Microarray expression values were retrieved from GEO and log2 transformed. RNA-sequencing derived RPKM/RSEM values from the original publications were log2 transformed. Since each expression dataset contained a slightly different set of genes, we merged the expression datasets and retained the genes commonly present in all datasets in order to use elastic net to perform feature (ie, gene) selection. Each dataset was then globally scaled across all genes and samples to obtain a mean of zero and standard deviation of one, and an empirical Bayes framework (combat) was applied to adjust batch effects on the merged dataset.[23] Using frequency-matched random sampling, we then assigned glioma samples with the 1p/19q co-deletion to the training dataset or validation dataset.

Predicting 1p/19 Status Using Gene Expression

DNA copy number profiles determining 1p/19q status were available for a subset of the cohort (Table S1). For cases in which it was absent, we applied a Gaussian window-smoothing algorithm to infer the pattern of chromosome arm-sized copy number variations (CNVs). Using the expression values of the genes located on Chr-1p and Chr-19q sorted by their genomic locations from start to end, we used a sliding 100 gene window to determine chromosome arm-wide 1p/19q expression levels. We applied the following equation to the resulting gene-specific expression patterns to determine 1p/19q status: $CNV_k\,(i) = \sum_{j=i-50}^{i+50} E_k\,(g_j)\,/101$

where (i) is the estimated copy number (relative value) of sample k at ith gene in the genomic-ordered gene list, g_j is the jth gene in the genomic-ordered gene list, and (g_j) is the relative gene expression value of that gene in sample k. Note that the estimated 1p/19q status is often consistent with the chromosome centromere borders, with increased or decreased values within specific chromosomes suggesting that it accurately represents chromosomal changes. We then applied hierarchical cluster analysis to CNVk(i) values to assign all samples from each dataset into a group reflecting 1p/19q co-deletion and another group with 1p/19q wild-type copy number.

Correlation of Somatic Mutations and Clinical Outcome

We applied the Kaplan-Meier estimator to assess the prognostic value of *CIC, FUBP1, NOTCH1,* and *PIK3CA* mutations (the most prevalent mutations in glioma) on OS. Two-sided log-rank tests were used to assess the differences of OS between the patients with and without any of these mutations. P values < .05 were considered statistically significant. We conducted this analysis using the TCGA dataset, which included mutation information. Although OS might be affected by treatment bias at the time of tumor progression, OS data are generally more accurate than progression-free survival data; therefore, we used OS to represent clinical outcomes that more accurately reflected disease aggressiveness in each glioma patient.

Gene Signature Selection and Risk-based Classification

Using the training set, we first prefiltered the genes based on Wald P values generated from univariate Cox models, selected the 1,000 most significant genes, and then applied the Cox proportional hazards model with elastic net penalty for variable selection. The univariate and multivariable Cox models were built using the R package "survival," and elastic net regression (ie, the combination of L1 regularization and L2 regularization) was performed using the R package "glmnet."[24,25] The penalty parameter λ was chosen based on 3-fold cross validation within the training set, which produced the minimum mean cross-validated error for the Cox model. Thus, we used shrinkage-based regularization combined with a univariate Cox model to obtain the gene signature.

Using the training dataset, we fit a multivariable Cox proportional hazards model with the genes identified using the above penalty-based method. We then computed a prognostic index for each patient in the validation set by multiplying his/her gene expression values by the corresponding regression coefficients estimated from the training data. This resulted in a risk score for each patient in the validation dataset according to a linear combination of the mRNA expression level from the validation data weighted by the multivariable Cox model-derived regression coefficients from the training data. We calculated the concordance index (C-index) for the gene signature, age, grade, and gene signature combined with age and grade, respectively, using the R package "survcomp."[26] We also calculated the hazard ratios (HRs) and their 95% confidence intervals between 2 groups of patients with risk scores above and below the median risk score computed from the training dataset.

Association of Risk Classification and Clinical Outcome

We divided the patients from the validation dataset into high-risk and low-risk groups based on their risk scores derived from the linear prediction and using the median risk score in the training set as the cutoff value. We used the Kaplan–Meier estimator and the 2-sided log-rank test to evaluate the differences in OS between the high-risk and low-risk patients. To examine the robustness of the risk-based classification using selected genes, we divided the patients into subgroups using a series of different risk scores as cutoff value and evaluated the difference of OS between high-risk and low-risk groups using the Kaplan–Meier estimator and HR. To further investigate the trend of the OS pattern to align with the predicted risk scores, we fit a smoothing spline to ascertain the association of risk scores with the OS of the patients in the validation dataset.

Top Gene Ontology and Gene Set Variation Analysis of Associated Genes

We first used a Student's t test to identify the genes differentially expressed between the high-risk and low-risk

groups, only including genes with a *P* value <.05 and an absolute difference in median gene expression of 0.4. We then mapped the gene ontology (GO) terms of the corresponding 260 genes that presented the most variance in expression between the 2 risk groups. We applied gene set variation analysis (GSVA)[27] to obtain the enrichment scores for each gene set that corresponded to the GO terms containing those genes in all of the patients.

Evaluation of Tumor Purity with ESTIMATE Gene Signatures

We inferred tumor purity of each sample using ESTIMATE,[28] which reflects the enrichment of stromal and immune cell gene signatures in a transcriptional profile.

Results

Effects of Somatic Mutations on Patient Outcome

Recent studies by TCGA and others have revealed genes frequently mutated in IDH-mutant and 1p/19q co-deleted glioma including the 1p gene *FUBP1* and the 19q gene *CIC*. Mutations in these genes fulfill the classic Knudson tumor suppressor 2-hit model in which one allele is lost and the second is inactivated through somatic mutation. hus, glioma carrying *CIC/FUBP1* mutations may have progressed further compared to those that are *CIC/FUBP1* wild type. To test this hypothesis, we performed univariate survival analyses of 151 diffuse IDH-mutant–codel gliomas from TCGA. We found significant correlation between OS and the presence of *FUBP1* mutation (*n* = 40 of 151; log-rank test *P* value = .05) but not *CIC* mutation (*n* = 71 of 151; log-rank test *P* value = .71). No associations were observed for other gene mutations frequently detected in IDH-mutant–codel gliomas such as in *NOTCH1* (23/151; log-rank test *P* value = .46), or *PIK3CA* (20/151; log-rank test *P* value = .06). Despite the relatively small numbers of death events in the TCGA dataset (15 of 151), our results suggest that these mutations do not significantly affect disease progression or clinical outcomes in IDH-mutant–codel glioma patients.

Constructing a Gene Expression Dataset of 1p/19q Co-deleted Glioma

Since only *FUBP1* somatic point mutations showed a significant but inconclusive association with patient outcomes, we set out to identify a gene expression signature with the potential to identify high-risk IDH-mutant-codel patients. We curated gene expression and clinical information from 7 publicly available datasets of adult diffuse glioma patients whose tumors were assessed by microarray or RNA sequencing. Where available, we used annotation on 1p/19q co-deletion available per the respective publications or data from genome-wide DNA copy number profiling to identify IDH-mutant–codel cases. Because these data were unavailable in some datasets, we applied a Gaussian window smoothing algorithm to infer the signal of large scale CNVs for each sample. By suppressing individual gene-specific expression patterns and averaging relative expression levels over large genomic regions, we selected the samples harboring the 1p/19q co-deletion based on the hierarchical clustering of CNVs estimated from each dataset (Fig. 1). We found that our method could predict 1p/19q codel status with high accuracy in the TCGA dataset (sensitivity = 0.97, specificity = 0.97, Mathews correlation coefficient = 0.94). Using the window sliding method in the Weller *et al*. dataset, which included CGH-based 1p/19q status, resulted in a specificity of 0.91. We curated 411 of 2,231 gliomas to contain the 1p/19q co-deletion and retained those samples with survival data, resulting in the dataset consisting of 374 1p/19q codel IDH mutant glioma gene expression profiles (Table S1).

The patient characteristics of the 1p/19q co-deletion glioma cohort are summarized in Table 1. The patients' median age at diagnosis was 43 years (range, 17–87 years). The median survival time was 75.7 months (range 1–248 months), and 121 events occurred. Among censored cases, the median follow up was 22.7 months (range: 0–182.3 months). After performing scale normalization using the 13,345 genes common to all datasets, we found no distinct clustering in any of the 5 gene expression datasets in the training and first validation set, suggesting that any platform or batch variance across the different datasets had been mostly eliminated (Fig. S2).

Identification of a 35-Gene Signature Associated with Overall Survival

We divided 333 patients from 5 datasets (Table S1) into training and validation datasets by frequency-matched random sampling so that each consisted of comparable numbers of astrocytic, oligodendrocytic, and mixed histological tumor subtypes. We also balanced for chemotherapy and radiotherapy treatment. Through controlled randomized sampling, the training dataset (*n* = 170, death events = 64) included 105 samples with treatment annotation of whom 26 patients received radiotherapy and 26 patients received chemotherapy, with eleven of those cases undergoing both radio- and chemo-therapy. The remaining 105 patients (*n* = 64) were surgically debulked without further treatment. The first validation dataset (*n* = 163, death events = 57) comprised 84 samples with treatment information specifying which 22 patients received radiotherapy and which 18 received chemotherapy. To build the training model, we selected the 1,000 genes with the most significant linear correlation with OS. We then applied a linear regression function that fits the Cox model regularized by an elastic net penalty (Fig. S3) to select 35 genes as active covariates of the Cox model to assess the prognostic index in the validation sets (Table S1). To assess performance of the signature genes as classifier, we computed a linear combination of the 35 genes using the coefficients of multivariable Cox regression derived from the training set to calculate the risk scores for the patients in the first validation dataset. By using the median risk score amongst

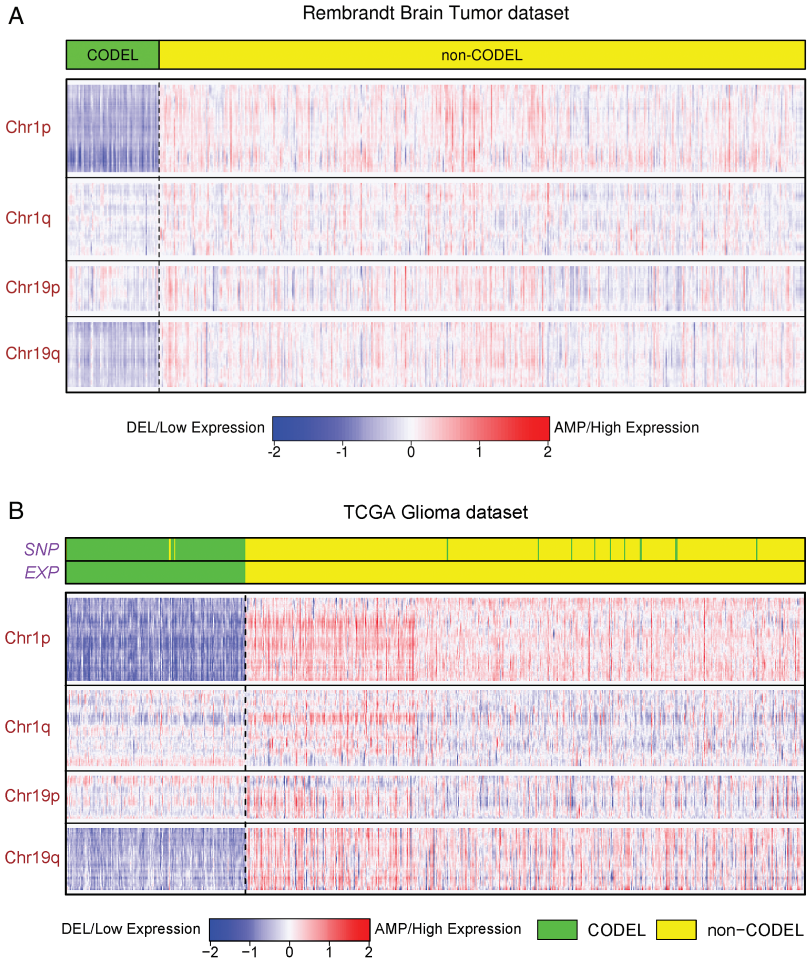

Fig. 1 Co-deletion of 1p/19q inferred by gene expression profiling. Normalized gene expression levels of chromosome 1 and chromosome 19. Top panel (A) shows Rembrandt glioma dataset ($n = 69$ with co-deletion, $n = 550$ total). Bottom panel (B) shows the TCGA dataset. The top bar denotes the CODEL status based on SNP6 DNA copy number arrays, the bottom bar denotes the CODEL status inferred by our method using gene expression data ($n = 162$ with co-deletion, $n = 667$ total). The green bar denotes the samples classified as CODEL,and the yellow bar denotes the samples classified as non-CODEL. The averaged expression level is shown in red-white-blue scale.

samples from the training dataset as the cutoff value to divide the first validation dataset into high-risk and low-risk groups, we found a significant difference in OS time between the 2 groups (log-rank test $P = .014$)(Fig. 2A). The high-risk group associated with a HR of 2.03 (95% confidence interval, 1.14–3.60). The median OS duration was 75.2 months for the patients with high-risk prognostic indices and 118.2 months for those with low-risk indices. We also found a significant difference in OS when dividing the first validation set using the top and bottom quartile risk scores (log-rank test $P = .0085$; HR = 2.9, 95% confidence interval, 1.3- 6.6;, Fig. S6A). We evaluated several different arbitrary risk score cutoffs to define the high-risk and low-risk patient groups and found that the OS of the low-risk group was significantly better than that of the high-risk group regardless of the cutoff value chosen (Fig. S6B, Fig. S6C). We then computed risk scores on 41 samples with predicted 1p/19q co-deletion from 2 datasets that were not included in the training set (Table S1). There was no significant difference in survival between the 2 resulting groups in this second validation set (log-rank test $P = .25$, HR = 2.7), which was likely due to the low number of death events ($n = 5$)(Fig. S7). There was a trend in the high-risk group toward reduced survival compared with the low-risk group. Combining both validation sets into a single analysis showed a highly significant difference in survival between the high-risk and low-risk groups (log-rank test $P = .00058$, HR = 2.65)(Fig. 2B). We performed scaled Schoenfeld residuals to verify proportional hazards assumption (Fig. S4) and martingale residuals analysis to verify the linearity (Fig. S5) as well as variance inflation factor (VIF) analysis to assess the potential for multicollinearity on those 35 signature genes (Table S2). The results of residual analysis with overall VIF were acceptable without high correlation, with all VIF values being < 10 in the training dataset. These results suggest that the 35-gene signature is significantly associated with the survival of 1p/19q co-del patients.

Table 1 Clinical characteristics of glioma patients harboring the 1p/19q co-deletion

Characteristic	No. of Patients
Age, years (range =17, 87; median = 43)	
< 43	175
> 43	179
43	13
Not available	7
Sex	
Male	217
Female	148
Not available	9
Time to death/last follow-up for event-free subjects: median 22.7 mo, range: 0–182.3 mo	
WHO grade	
II	179
III	159
IV	16
Not available	20
Histologic subtype	
Oligodendroglioma	217
Astrocytoma	32
Oligoastrocytoma	96
GBM	9
Not available	13

Abbreviations: GBM, glioblastoma multiforme; WHO, World Health Organization.
Note: The dataset of glioma patients with the 1p/19q co-deletion was curated and selected based on copy number variation information from The Cancer Genome Atlas (15 death events/151 patients), the Chinese Glioma Genome Atlas (22 death events/81 patients),[12,17] Gravendeel et al. (35 death events/40 patients),[14] Rembrandt et al. (49 death events/61 patients),[15] Weller et al. (4 death events/31 patients),[22] and Guan et al (1 death event/10 patients).[21]

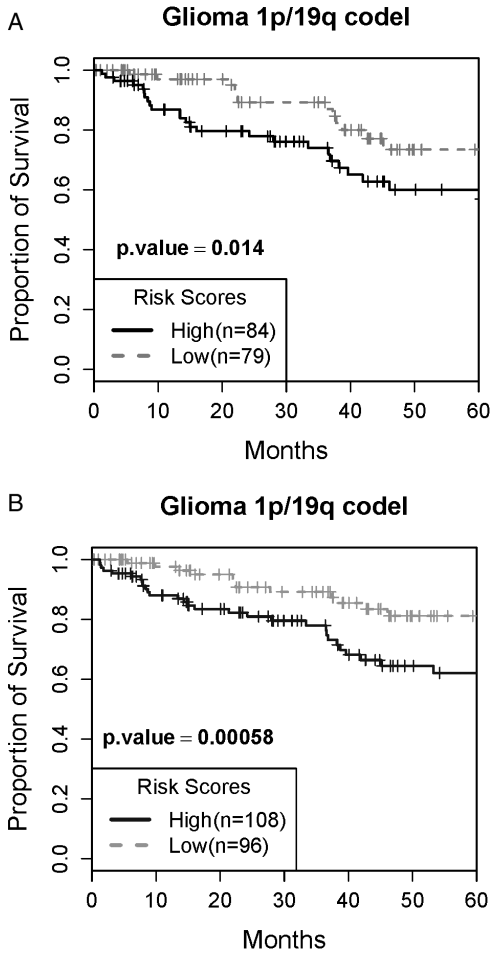

Fig. 2 **A.** Kaplan-Meier survival analysis of glioma patients harboring 1p/19q co-deletion according to 35-gene signature derived risk scores. Kaplan-Meier cumulative survival curves for first validation set glioma patients with 1p/19q co-deletion tumors, classified in 2 groups based on 35-gene signature derived risk scores. The survival of the high-risk patients (solid line) was significantly worse than that of the low-risk patients (dashed line; $P = .014$, log-rank test) **B.** Kaplan-Meier survival analysis of codel glioma in the combined validation dataset according to 35-gene signature derived risk scores. Kaplan-Meier survival curves for 1p/19q co-deletion glioma patients from the first and second validation dataset, separated into 2 groups based on risk score. The survival of the high-risk patients (solid line) was significantly worse than that of the low-risk patients (dashed line; $P = .0058$, log-rank test).

Multivariable Analysis Shows Prognostic Power of 35-Gene Signature

By using the prognostic index (risk score) as a continuous covariate, we determined the predictive accuracy by computing the C-index of the gene signature, age, and grade. We also examined the C-index value of the gene signature combined with age and grade. Both analyses were performed in the joint validation set and were restricted to the partcipants whose age and grade information were available. The C-index of the gene signature (0.626 ± 0.044) was comparable to that of age (0.640 ± 0.048) or grade (0.640 ± 0.073) alone (Table 2). The highest C-index was achieved when the 3 variables (0.663 ± 0.041) were combined. These results suggest that risk prediction was most accurate when the 35-gene signature was combined with age and tumor grade.

Functional Annotation of 35-gene Signature

We compared gene expression between the high-risk and low- risk groups and found that 32 of 35 signature genes showed a significant difference (Fig. S8, Table S3). Gene set variation analysis revealed that the differentially expressed genes in two groups were associated with inflammation, acetylation activity, response to copper ions, prostaglandins (Fig. 3). The corresponding biological functions of protein acetylation, inflammatory response, and copper homeostasis may contribute to these patients' high risk and poor clinical outcome.

Table 2 Performance of multivariable analysis in the validation dataset

Predictor	Gene Signature	Age	Grade	Age + Grade	Gene Signature + Age + Grade
C-Index ± SE	0.626 ± 0.044	0.640 ± 0.048	0.640 ± 0.073	0.656 ± 0.041	0.663 ± 0.041
C-Index(CI)	0.540, 0.712	0.545, 0.734	0.497, 0.785	0.574, 0.737	0.583, 0.743
HR (95% CI)	1.78(1.02–3.11)	1.71(0.99–2.97)	1.26 (0.74–2.15)	2.06(1.18–3.60)	3.23 (1.73–6.04)

Abbreviations: C-index, concordance index; HR, hazard ratio; CI, confidence interval.
Estimates are based on data from patients in the combined validation dataset (n = 191, first validation dataset + second validation dataset), with both age and grade information available. The hazard ratios (HRs) and their 95% confidence intervals between 2 groups of patients in the validation dataset were calculated based on their risk scores above and below the median risk score computed from the training dataset.

The presence of the inflammation category among the differentially activated GO terms suggested differences in the tumor microenvironment between the high-risk and low-risk groups. We applied the ESTIMATE algorithm to predict tumor purity using the gene expression profiles[28] and found a significant increase in ESTIMATE scores in the high-risk group (Fig. S9), suggesting that a greater presence of inflammatory microenvironment components is associated with progressive tumorigenesis.

Applying the 35-gene Signature Across Glioma

We asked whether the 35-gene signature model is also associated with patient survival in patients with IDH-wild-type or IDH-mutant non-codel gliomas. From TCGA, we obtained the gene expression profiles of 223 IDH-mutant gliomas that were wild type for chromosome arms 1p and 19q and the transcriptional profiles from 221 IDH–wild-type gliomas. The analysis was restricted to cases with available outcome data and expression data generated by RNA sequencing. After computing risk scores for all samples, we separated the 2 datasets into low-risk and high-risk groups based on median risk score respectively and observed significant differences in OS for both glioma categories (Fig. 4).

To gain insight into the universal relevance of the 35-gene signature across different molecular subtypes of glioma, we also applied the ESTIMATE algorithm to compare tumor purity between high-risk and low- risk groups. As for IDH-mutant–codel gliomas, we found that ESTIMATE-based tumor purity scores were significantly lower in the low-risk group of IDH–wild-type glioma samples compared with their high-risk counterpart (Fig. S10). This was not the case for IDH–wild-type LGG nor for IDH-mutant non-codel gliomas regardless of grade (Fig. S10). The difference in microenvironment presence between the high-risk and low-risk groups of IDH-wildtype glioblastoma emphasizes the facilitating role that tumor-associated microglia play in promoting disease progression.[29]

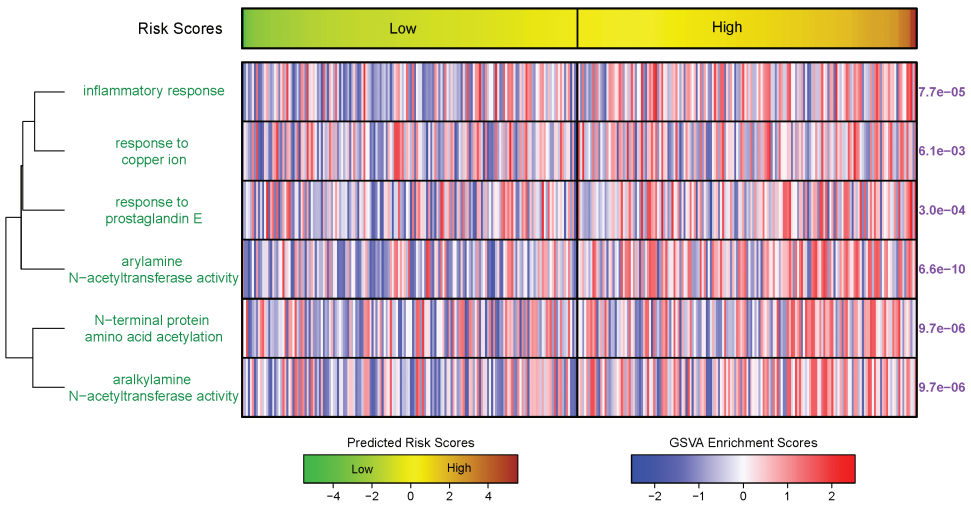

Fig. 3 Association of risk groups with gene ontology (GO) function. Risk scores for each patient (top bar; in ascending order, from left to right) were derived from a multivariable Cox model. Gene set variance analysis was used to calculate gene set enrichment scores (bottom). The P values on the right were obtained using a t test of enrichment scores from high-risk and low-risk groups for each GO term.

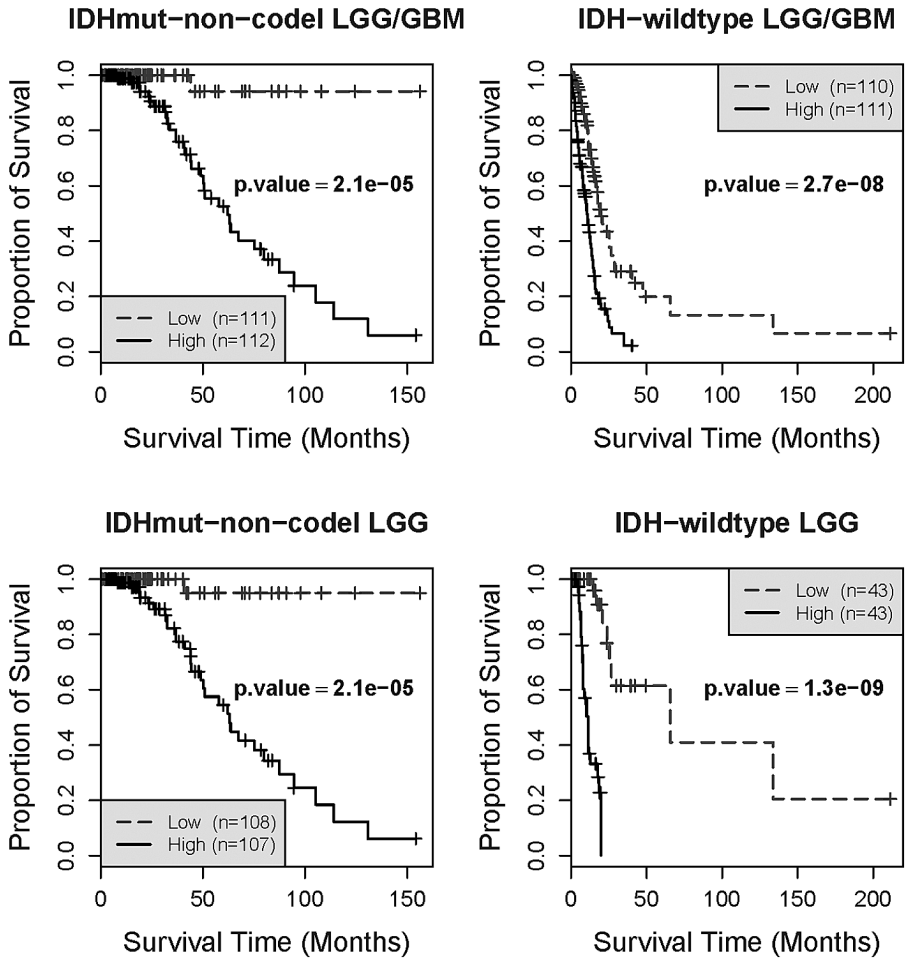

Fig. 4 Prediction of outcome in non-codel IDH-mutant glioma and IDH-wild-type glioma. Kaplan-Meier cumulative survival curves for TCGA diffuse glioma patients whose tumors carry IDH-mutation but not the 1p/19q co-deletion (left) and with IDH-wild-type tumors (right), classified into 2 groups based on 35-gene signature-derived risk scores. *P* value is the result of a log-rank test between the 2 groups shown in each panel.

Discussion

High-throughput gene profiling and sequencing have yielded new insights on the molecular aberrations underlying glioma.[4,30] As our perspective on the optimal clinical and molecular marker-based classification of adult diffuse glioma harboring 1p/19q co-deletion progresses, biomarkers for risk-based classification may provide additional value. Our systematic analysis identified a 35-gene signature, which classified 1p/19q codel glioma patients according to their OS. Remarkably, the median survival of the group of patients classified as high risk was 75 months, confirming that 1p/19q codel patients have a favorable OS and suggesting a relatively homogeneous disease subtype. Our gene signature was derived from multiple data sources and represented patients from a mixture of diffuse glioma grades and histologies. We validated the prognostic performance in independent validation datasets and showed the signature to have added predictive signal when combined with known prognostic markers such as age and grade. While normalization of the raw data of training and validation should be performed independently, this was not possible due to the nonavailability of several of the files needed. Patients in our cohort were treated using a variety of different modalities, and treatment annotations were lacking for a substantial portion of the dataset. With the recent introduction of a potential standard of care for low-grade glioma,[8] it is important to repeat and validate the gene signature on a coherently treated patient dataset while considering additional prognostic factors such as tumor size, location, and extent of resection. In order to pursue validation studies, risk scores can be computed using the gene signature and regression coefficients provided in Table S2.

The univariate Cox model alone is insufficient for feature selection through estimation of survival as clinical endpoints when solving regression problems with high dimensional data. To prevent overfitting, the ridge regression Cox model demonstrates the best performance in

tested datasets.[31] Therefore, we applied the univariate Cox model for filtering genes related to OS time and used regularized regression coefficients, which were calculated by an elastic net regression Cox model that combined the algorithm of ridge and lasso regression to increase the predictive performance of the prognostic index on independent data. Owing to the relatively small number of events (64 deaths/170 patients in the training dataset), we applied a 3-fold cross-validation for Cox proportional hazards regression and selected the signature genes with optimized λ based on penalty regularization. While none of the individual genes showed an exceptionally high coefficient in our Cox model, multiple genes cumulatively exhibited an effect on survival prediction. Finally, we used multivariable Cox regression to adjust the selected genes for the clinical factors of age and grade and to generate our prognostic index. We would have preferred to adjust for the clinical factors while selecting the genes in the elastic net Cox model but could not find software to facilitate this approach. As treatment information was unavailable for a considerable portion of our cohort (43%), we did not consider treatment variables in our statistical modeling.

Pathway analysis suggested that N-terminal acetyltransferases (NATs), protein acetylation, response to copper ions, prostaglandins, and inflammation may be involved in 1p/19q glioma progression. Therapeutic agents (including tamoxifen and cisplatin) have been reported to demonstrate their anticancer effects through NAT inhibitory activity,[32–35] suggesting that NATs be targeted as a potential therapeutic strategy in high-risk 1p/19q co-deleted cases. In addition, copper depletion may act as an effective antiangiogenesis strategy,[36] and prostaglandins play an important role in cell adhesion, migration, and invasion during cancer development.[37] Accordingly, the genes involved in protein acetylation and response to inflammation and copper are highly expressed in high-risk glioma patients.[38] These data indicate that alterations in the expression levels of these signature genes might exert significant roles in glioma progression by promoting growth and conveying cell survival advantages. In addition to our pathway analyses, we noted that the stromal and immune-related signals quantitated via ESTIMATE scores were significantly increased in the high-risk group relative to the low-risk group of 1p/19q codel glioma as well as in IDH–wild-type glioblastoma. This observation implies an association between the survival risk predicted by our gene signature and the infiltration by tumor-associated normal cells, which play a critical role in microenvironment regulation during tumor progression.[39] Four genes (*ITIH3, TRAT1, FRZB, IL32*) from 35 genes' signature overlap with the stromal and immune gene signatures use to define ESTIMATE scores, further nominating the tumor microenvironment as a potential risk factor for subsets of glioma patients.

Collectively, our findings highlighted signature genes that might be involved in critical tumor progression and fundamental biological functions in gliomas with the 1p/19q co-deletion. The lack of treatment standardization among our patient cohort means that further research is needed to determine whether this or other gene signatures could serve as treatment biomarkers. Ideally, clinical decisions are based on a predictive model that involves integrating clinical variables, tumor phenotypic, and molecular factors. While further and prospective validation is needed, the gene signature approach may provide a starting point to better understanding of prognostic risk factors in 1p/19q co-deletion glioma. The results described here provide a first report investigating the heterogeneity of the relatively novel entity of 1p/19q codel glioma.

Supplementary Material

Supplementary data are available at *Neuro-Oncology* online.

Acknowledgments

This work was supported by grants from the U.S. National Institutes of Health P50 CA127001, R01 CA190121, P01 CA085878; and Cancer Prevention & Research Institute of Texas (CPRIT) R140606.

Conflicts of interest statement. KRH is a consultant for Angiochem, Inc.

References

1. Weller M, van den Bent M, Hopkins K, et al.; European Association for Neuro-Oncology (EANO) Task Force on Malignant Glioma. EANO guideline for the diagnosis and treatment of anaplastic gliomas and glioblastoma. *Lancet Oncol*. 2014;15(9):e395–e403.
2. Brat DJ, Verhaak RG, Aldape KD, et al.; Cancer Genome Atlas Research Network. Comprehensive, Integrative Genomic Analysis of Diffuse Lower-Grade Gliomas. *N Engl J Med*. 2015;372(26):2481–2498.
3. Eckel-Passow JE, Lachance DH, Molinaro AM, et al. Glioma Groups Based on 1p/19q, IDH, and TERT Promoter Mutations in Tumors. *N Engl J Med*. 2015;372(26):2499–2508.
4. Ceccarelli M, Barthel FP, Malta TM, et al.; TCGA Research Network. Molecular Profiling Reveals Biologically Discrete Subsets and Pathways of Progression in Diffuse Glioma. *Cell*. 2016;164(3):550–563.
5. Louis DN, Perry A, Reifenberger G, et al. The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol*. 2016;131(6):803–820.
6. van den Bent MJ, Brandes AA, Taphoorn MJ, et al. Adjuvant procarbazine, lomustine, and vincristine chemotherapy in newly diagnosed anaplastic oligodendroglioma: long-term follow-up of EORTC brain tumor group study 26951. *J Clin Oncol*. 2013;31(3):344–350.
7. Dubbink HJ, Atmodimedjo PN, Kros JM, et al. Molecular classification of anaplastic oligodendroglioma using next-generation sequencing: a report of the prospective randomized EORTC Brain Tumor Group 26951 phase III trial. *Neuro Oncol*. 2016;18(3):388–400.
8. Buckner JC, Shaw EG, Pugh SL, et al. Radiation plus Procarbazine, CCNU, and Vincristine in Low-Grade Glioma. *N Engl J Med*. 2016;374(14):1344–1355.

9. Cairncross G, Wang M, Shaw E, et al. Phase III trial of chemoradiotherapy for anaplastic oligodendroglioma: long-term results of RTOG 9402. *J Clin Oncol*. 2013;31(3):337–343.
10. Zadeh G, Khan OH, Vogelbaum M, Schiff D. Much debated controversies of diffuse low-grade gliomas. *Neuro Oncol*. 2015;17(3):323–326.
11. Figarella-Branger D, Mokhtari K, Dehais C, et al.; POLA Network. Mitotic index, microvascular proliferation, and necrosis define 3 groups of 1p/19q codeleted anaplastic oligodendrogliomas associated with different genomic alterations. *Neuro Oncol*. 2014;16(9):1244–1254.
12. Alentorn A, Dehais C, Ducray F, et al.; POLA Network. Allelic loss of 9p21.3 is a prognostic factor in 1p/19q codeleted anaplastic gliomas. *Neurology*. 2015;85(15):1325–1331.
13. Siegal T. Clinical impact of molecular biomarkers in gliomas. *J Clin Neurosci*. 2015;22(3):437–444.
14. Yan W, Zhang W, You G, et al. Molecular classification of gliomas based on whole genome gene expression: a systematic report of 225 samples from the Chinese Glioma Cooperative Group. *Neuro Oncol*. 2012;14(12):1432–1440.
15. Freije WA, Castro-Vargas FE, Fang Z, et al. Gene expression profiling of gliomas strongly predicts survival. *Cancer Res*. 2004;64(18):6503–6510.
16. Huang YT, Hsu T, Kelsey KT, Lin CL. Integrative analysis of micro-RNA, gene expression, and survival of glioblastoma multiforme. *Genet Epidemiol*. 2015;39(2):134–143.
17. Madhavan S, Zenklusen JC, Kotliarov Y, Sahni H, Fine HA, Buetow K. Rembrandt: helping personalized medicine become a reality through integrative translational research. *Mol Cancer Res*. 2009;7(2):157–167.
18. Gravendeel LA, Kouwenhoven MC, Gevaert O, et al. Intrinsic gene expression profiles of gliomas are a better predictor of survival than histology. *Cancer Res*. 2009;69(23):9065–9072.
19. Bao ZS, Chen HM, Yang MY, et al. RNA-seq of 272 gliomas revealed a novel, recurrent PTPRZ1-MET fusion transcript in secondary glioblastomas. *Genome Res*. 2014;24(11):1765–1773.
20. Torres-García W, Zheng S, Sivachenko A, et al. PRADA: pipeline for RNA sequencing data analysis. *Bioinformatics*. 2014;30(15):2224–2226.
21. Guan X, Vengoechea J, Zheng S, et al. Molecular subtypes of glioblastoma are relevant to lower grade glioma. *PLoS One*. 2014;9(3):e91216.
22. Weller M, Weber RG, Willscher E, et al. Molecular classification of diffuse cerebral WHO grade II/III gliomas using genome- and transcriptome-wide profiling improves stratification of prognostically distinct patient groups. *Acta Neuropathol*. 2015;129(5):679–693.
23. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics*. 2007;8(1):118–127.
24. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*. 2010;33(1):1–22.
25. Hughey JJ, Butte AJ. Robust meta-analysis of gene expression using the elastic net. *Nucleic Acids Res*. 2015;43(12):e79.
26. Schröder MS, Culhane AC, Quackenbush J, Haibe-Kains B. survcomp: an R/Bioconductor package for performance assessment and comparison of survival models. *Bioinformatics*. 2011;27(22):3206–3208.
27. Hänzelmann S, Castelo R, Guinney J. GSVA: gene set variation analysis for microarray and RNA-seq data. *BMC Bioinformatics*. 2013;14:7.
28. Yoshihara K, Shahmoradgoli M, Martínez E, et al. Inferring tumour purity and stromal and immune cell admixture from expression data. *Nat Commun*. 2013;4:2612.
29. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. *Cell*. 2011;144(5):646–674.
30. Verhaak RG, Hoadley KA, Purdom E, et al.; Cancer Genome Atlas Research Network. Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*. 2010;17(1):98–110.
31. Bøvelstad HM, Nygård S, Størvold HL, et al. Predicting survival from microarray data–a comparative study. *Bioinformatics*. 2007;23(16):2080–2087.
32. Lu KH, Lin KL, Hsia TC, et al. Tamoxifen inhibits arylamine N-acetyltransferase activity and DNA-2-aminofluorene adduct in human leukemia HL-60 cells. *Res Commun Mol Pathol Pharmacol*. 2001;109(5-6):319–331.
33. Kalvik TV, Arnesen T. Protein N-terminal acetyltransferases in cancer. *Oncogene*. 2013;32(3):269–276.
34. Lee JH, Lu HF, Wang DY, et al. Effects of tamoxifen on DNA adduct formation and arylamines N-acetyltransferase activity in human breast cancer cells. *Res Commun Mol Pathol Pharmacol*. 2004;115-116:217–233.
35. Ragunathan N, Dairou J, Pluvinage B, et al. Identification of the xenobiotic-metabolizing enzyme arylamine N-acetyltransferase 1 as a new target of cisplatin in breast cancer cells: molecular and cellular mechanisms of inhibition. *Mol Pharmacol*. 2008;73(6):1761–1768.
36. Goodman VL, Brewer GJ, Merajver SD. Copper deficiency as an anticancer strategy. *Endocr Relat Cancer*. 2004;11(2):255–263.
37. Menter DG, Dubois RN. Prostaglandins in cancer cell adhesion, migration, and invasion. *Int J Cell Biol*. 2012;2012:723419.
38. Di Cerbo V, Schneider R. Cancers with wrong HATs: the impact of acetylation. *Brief Funct Genomics*. 2013;12(3):231–243.
39. Quail DF, Joyce JA. Microenvironmental regulation of tumor progression and metastasis. *Nat Med*. 2013;19(11):1423–1437.