# Identification and validation of a gene expression signature that predicts outcome in malignant glioma patients

ATSUSHI KAWAGUCHI[4], NAOKI YAJIMA[1], YOSHIHIRO KOMOHARA[3], HIROSHI AOKI[1], NAOTO TSUCHIYA[1], JUMPEI HOMMA[1], MASAKAZU SANO[1], MANABU NATSUMEDA[1], TAKEO UZUKA[1], AKIHIKO SAITOH[1], HIDEAKI TAKAHASHI[1], YUKO SAKAI[5], HITOSHI TAKAHASHI[2], YUKIHIKO FUJII[1], TATSUYUKI KAKUMA[4] and RYUYA YAMANAKA[5]

Departments of [1]Neurosurgery and [2]Pathology, Brain Research Institute, Niigata University; [3]Department of Cell Pathology, Graduate School of Medical Sciences, Kumamoto University; [4]Biostatistics Center, Kurume University; [5]Graduate School for Health Care Science, Kyoto Prefectural University of Medicine, Kyoto 602-8566, Japan

Received August 16, 2011; Accepted October 3, 2011

DOI: 10.3892/ijo.2011.1240

**Abstract.** Better understanding of the underlying biology of malignant gliomas is critical for the development of early detection strategies and new therapeutics. This study aimed to define genes associated with survival. We investigated whether genes selected using random survival forests model could be used to define subgroups of gliomas objectively. RNAs from 50 non-treated gliomas were analyzed using the GeneChip Human Genome U133 Plus 2.0 Expression array. We identified 82 genes whose expression was strongly and consistently related to patient survival. For practical purposes, a 15-gene set was also selected. Both the complete 82 gene signature and the 15 gene set subgroup indicated their significant predictivity in the 3 out of 4 independent external dataset. Our method was effective for objectively classifying gliomas, and provided a more accurate predictor of prognosis. We assessed the relationship between gene expressions and survival time by using the random survival forests model and this performance was a better classifier compared to significance analysis of microarrays.

## Introduction

Glioblastoma, which is pathologically the most aggressive form of glioma, has a median survival range of only 9-15 months (1,2). Advances in basic knowledge of cancer biology and surgical techniques, chemotherapy, and radiotherapy, have led to little improvement in survival rates of patients suffering from glioblastoma (1). Poor prognosis is attributable to difficulties in early detection and to a high recurrence rate after initial treatment. Therefore, it is important to devise more effective therapeutic approaches, to reveal more clearly the biological features of glioblastoma, and to identify novel target molecules for diagnosis and therapy of the disease. Several histological grading schemes exist, and the World Health Organization (WHO) system is currently the most widely used (3). A high WHO grade correlates with clinical progression and decreased survival. However, there are still many individual variations within diagnostic categories, resulting in a need for additional prognostic markers. The inadequacy of histopathological grading is evidenced, in part, by the inability to recognize these patients prospectively.

Recently, microarray technology has permitted development of multi-organ cancer classification including gliomas (4-6), identification of glioma subclasses (7-15), discovery of molecular markers (16-23) and prediction of disease outcomes (24-27). Unlike clinicopathological staging, molecular staging can predict long-term outcomes of any individual based on the gene expression profile of the tumor at diagnosis, helping clinicians make an optimal clinical decisions. The analysis of expression profiles of genes in clinical materials is an essential step towards clarifying the detailed mechanisms of oncogenesis and the discovery of target molecules for the development of novel therapeutic drugs.

In the present study, we describe an expression profiling study on a panel of 50 patients with glioma for the identification of genes predictive of overall survival using random survival forests model, with validation in independent data sets.

## Materials and methods

*Samples*. Tissues were snap-frozen in liquid nitrogen within 5 min of harvesting, and stored thereafter at -80˚C. Clinical stage was estimated from accompanying surgical pathology and clinical reports. Samples were specifically re-reviewed by a board-certified pathologist in Niigata University according to the WHO 2000 criteria, using observation of sections of paraffin-embedded tissues that were adjacent or in close proximity to the frozen sample from which the RNA was subsequently extracted. The histopathology of each collected specimen was reviewed to confirm the adequacy of the sample (i.e., minimal contamination with non-neoplastic elements), and to assess the extent of tumor necrosis and cellularity. Informed consent was obtained

*Correspondence to:* Dr Ryuya Yamanaka, Kyoto Prefectural University of Medicine, Graduate School for Health Care Science, 465 Kajii-cho, Kamigyoku, Kyoto 602-8566, Japan
E-mail: ryaman@cmt.kpu-m.ac.jp

*Key words:* glioma, gene expression profile, prognostic marker

from all patients for the use of the samples, in accordance with the guidelines of the Ethics Committee on Human Research, Niigata University Medical School (Protocol no. 70). Overall survival was measured from the date of the first operation for diagnosis. Survival endpoints corresponded to dates of death or last follow-up.

*RNA extraction and array hybridization*. Approximately 100 mg of tissue from each tumor was used to extract total RNA using the Isogen (Nippongene, Toyama, Japan) method, following the manufacturer's instructions. The quality of RNA obtained was verified with the Bioanalyzer System (Agilent Technologies, Tokyo, Japan) using RNA Pico Chips. Only samples with 28S/18S ratios >0.7 and with no evidence of ribosomal peak degradation were included in the study. Six micrograms of RNA were processed for hybridization on the GeneChip Human Genome U133 Plus 2.0 Expression arrays (Affymetrix, Inc., Tokyo, Japan), which comprised ~47,000 genes. After hybridization, the chips were processed using a Fluidics Station 450, a High-Resolution Microarray Scanner 3000, and a GCOS Workstation Version 1.3 (Affymetrix, Inc.).

*Validation of differential expression by real-time quantitative PCR*. Quantitative PCR (QPCR) was performed on a StepOne Real-Time PCR Systems (Applied Biosystems, Tokyo, Japan) using the TaqMan Universal PCR Master Mix (Applied Biosystems) according to the manufacturer's protocol. The TaqMan Gene Expression Assay Mix contained primers and TaqMan probes: Hs99999905-m1 (GAPDH), Hs00933163-m1 (PIK3R1), Hs00934330-m1 (SERPING1), and Hs00162558-m1 (TAGLN) from Applied Biosystems. Total RNA (5 $\mu$g) was subjected to reverse-transcription into cDNA using Super-Script II (Invitrogen, Tokyo, Japan). One microliter of this cDNA was used for QPCR. Validation was performed on a subset of tumors that were part of the original tumor data set assessed. Assays were done in duplicate. The raw data produced by QPCR referred to the number of cycles required for reactions to reach exponential phase. Expression of GAPDH was used for normalization of the QPCR data. Mean expression fold change differences between tumor groups were calculated using the $2^{-\Delta\Delta CT}$ method (28). Mean expression fold changes between short- (survival time <2 years; n=30) and long-term (survival time >2 years; n=14) survivors were compared.

*Immunohistochemistry*. Five-micron sections from formalin-fixed, paraffin-embedded tissue specimens were used for immunohistochemistry. Immunohistochemistry for PIK3R1 (antibody dilution 1:200; Abcam, Tokyo), SERPING 1 (M81) (antibody dilution 1:50; Abcam) and TAGLN (SM22$\alpha$) (antibody dilution 1:200; Abcam) was performed as described previously (21). Staining intensity was classified as none or weakly positive (0 point), moderately positive (1 point), or strongly positive (2 points). Averages of three independent measurements were calculated to the first decimal. Observers were not aware of case numbers.

*Statistical analysis*. To rank genes, the Cox score for each gene was obtained from the univariate Cox proportional hazards regression model. To obtain stable scores, we took the mean of scores computed from the 5-fold cross-validated samples. We excluded genes with low scores ($p>0.0025$). Genes that passed the filter criteria were considered for further analysis. All statistical analysis was performed in the R software (29) and Bioconductor (30). To select predictors for survival time, we first set filtered gene expressions and phenotypes [age, WHO grade, Karnofsky performance status score (KPS), and gender] to be initial candidates and iteratively fitted random survival forests model (31), at each iteration building a new forest after discarding those predictors with the smallest variable importances. For parameters in random survival forests model such as the number of tree and the number of valiables selected randomly at each node, we gave the default setting in the rsf function within the randomSurvivalForest package before the selection. We selected the set of predictors with the smallest 5-fold cross-validated error rate, which is one minus the Harrell's concordance index (32) and the '1 s.e. rule' used in the classification trees literature (33). That is, the error rate plus its 1 s.e. was used as the threshold for the selection. The cross-validated error rate was computed in each set of predictors.

We classified samples into two survival groups by a Ward's minimum variance cluster analysis, with its inputs being ensemble cumulative hazard functions for each individual for all unique death time-points estimated from the fitted random survival forests model. The Kaplan-Meier method (34) was used to estimate the survival distribution for each group. A log-rank test was used to test the difference between survival groups. For the purpose of comparison, the same analysis was conducted for two different groups based on WHO grade and gene expression clustering. The relationship between the grouping by the random survival forests model and gene expressions for the selected to genes was visualized by a heatmap.

GSEA (gene set enrichment analysis) was performed by the GSA method (35) on KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway (36). The pathways used in this analysis were selected to the one which the selected genes belong. We picked up pathways with FDR = 0.3.

To quantify the accuracy of prediction, we utilized independent test data sets from four studies (25,26,37,38). Firstly, ensemble mortalities for each individual in the test data were computed using the random survival forests model fitted to our data set with selected genes. Secondly, the relationship between computed ensemble mortalities and survival times in the test data was analyzed using Harrell's concordance index. Predictive accuracy was assessed using a weighted combination of p-values for the index, with weights being based on the sample sizes of the studies. These were implemented using the concordance.index and test.hetero.test functions within the survcomp package (39). Since, there were some missing genes in the test data, the genes were treated as missing values and imputed through the proximity approach (40). For signatures found in other studies (25,26,37), we perfumed the overlap analysis, and applied random survival forests with the signatures and compared based on survival curves described above. For this analysis, a $p<0.05$ was considered to indicate statistical significance.

## Results

*Patients characteristics*. Fifty non-treated glioma specimens [five astrocytoma (grade II), seven anaplastic astrocytoma, six anaplastic oligoastrocytoma or oligodendroglioma (grade III),

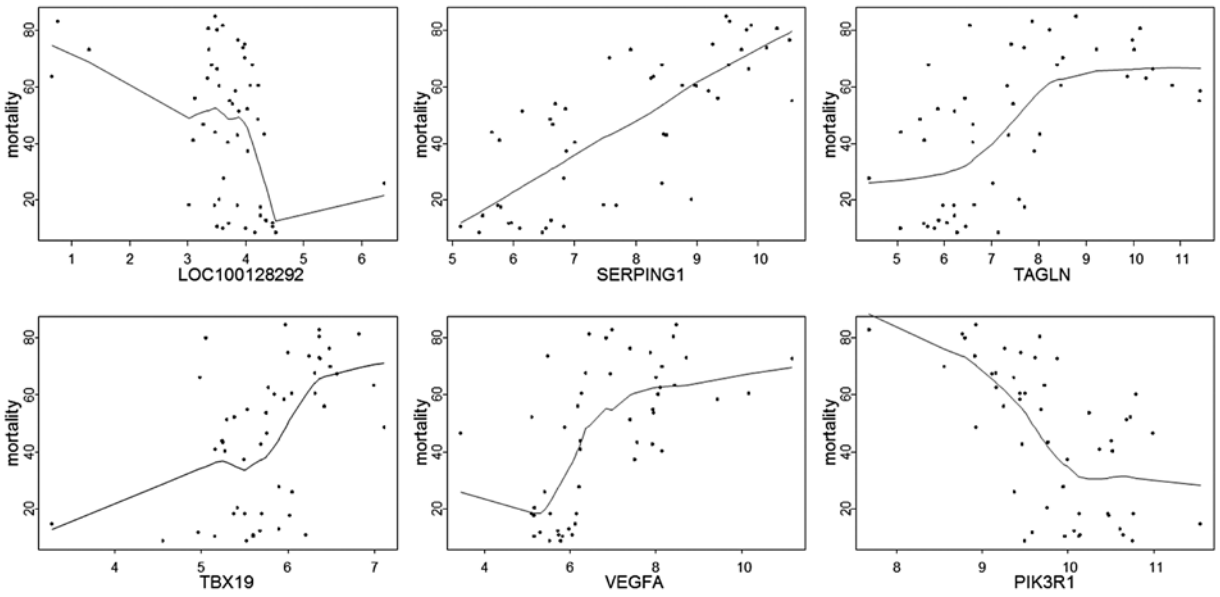


Figure 1. Marginal plots for the six genes. Values on the vertical axis represent the expected number of deaths for a given gene expression and the horizontal axis represents the gene expression value. Each point corresponds to an individual, patients with a higher value on the vertical axis have a higher risk. Lines represent a LOWESS (locally weighted scatterplot) smoothing to show the relationship between predicted values by random survival forests model and gene expressions based on the scatter plot.

and 32 glioblastoma (grade IV), corresponding to the WHO criteria] were obtained from patients who underwent surgical resection between 2000 and 2005. The mean age of the patients was 65 (range, 18-80). Thirty-four patients were males, and 16 were females. The preoperative Karnofsky Performance Status (KPS) was at least 50 in 48 (96%) patients. For anaplastic astrocytoma and glioblastoma, after maximum surgical resection of the tumor, patients had a course of external beam radiation therapy (standard dose of 60 Gy to the tumor with a 2-cm margin) and 1st line chemotherapy with nimustine, and temozolomide at recurrence. For grade II astrocytoma, after surgical resection of the tumor, patients had standard dose of 60 Gy radiation therapy to the tumor with a 2-cm margin at initiation and chemotherapy with temozolomide at recurrence. In most anaplastic oligodendroglioma cases, patients were treated by radiotherapy with chemotherapy of the modified PCV regimen (procarbazine, nimustine and vincristine). Patients were monitored for recurrences of the tumor during the initial and maintenance therapy by magnetic resonance imaging (MRI) or computed tomography (CT). Treatments were carried out at the Department of Neurosurgery, Niigata University Hospital. The median survival time with grade II was 57 months, grade III 29.5 months, and Grade IV 13.5 months, respectively. The median follow-up for survivors was 4.7 years (range 3.7-8.3).

*Selection of predictive genes*. Eighty-two genes and no phenotypes were selected as the predictor. Table I shows the list of the genes with the obtained variable importances (VI). The scatter plot (Fig. 1) shows the relationship between estimated ensemble mortalities and expressions for six selected genes (LOC100128292, TBX19, VEGFA, PIK3R1, TAGLN, and SERPING1). The heatmap (Fig. 2) consists of patients clustered by the estimated ensemble mortalities in the column and genes clustered by their expressions in the row. For patients with poor survival (represented by the blue bar at the top), the upper located genes are overexpressed while the lower are underexpressed. For patients with good survival (represented by the red bar at the top), there were no clear distinguishing patterns. However, the expression pattern in patients with very good survival (located in the right cluster) was reversed from the patterns observed in the poor survivors. Thus, from the heatmap, the selected genes might be effective in distinguishing between poor and very good survivors.

*Survival analysis using the selected gene classifiers reveals a prognostic value for tumor subtype*. Kaplan-Meier curves (Fig. 3) were drawn for groups classified by WHO grades III and IV (Fig. 3A), by the clustering analysis based on the gene expressions selected by the SAM (significance analysis of microarrays) (41) with FDR (false discovery rates) <0.0005 (Fig. 3B), and by random survival forests model (Fig. 3C). The corresponding test statistics (Q) and p-values (p) for the log-rank test were Q=8.6, p=0.0034 for WHO grade, Q=14.4, p=0.0001 for the SAM, and Q=45.6, p<0.0001 for random survival forests model with the 82-gene set. These results show that the random survival forests model is more useful than the direct use of gene expressions.

*GSEA*. The six pathways were identified as: focal adhesion, glycosaminoglycan degradation, leukocyte transendothelial migration, complement and coagulation cascades, starch and sucrose metabolism, and other glycan degradation.

*Survival analysis using the selected gene classifiers in independent data sets*. The results for the accuracy of prediction are summarized in Table II. The computed Harrell's concordance indexes (their 95% confidence intervals and p-values) are shown. The p-value of the combined tests was p<0.0001, which indicates that the selected set of genes was significantly predictive.

Table I. Identification of survival related 82 genes.

| Probe | Symbol | Description | VI |
|---|---|---|---|
| **1562598_at** | **LOC100128292[a]** | **Hypothetical LOC100128292** | **0.0051 (0.00674536)** |
| 242769_at | | ESTs | 0.0051 |
| **206838_at** | **TBX19[a]** | **T-box 19** | **0.0042 (0.00590219)** |
| 1563453_at | | cDNA DKFZp686J113 (from clone DKFZp686J113) | 0.0034 |
| **222477_s_at** | **TM7SF3[a]** | **Transmembrane 7 superfamily member 3** | **0.0034 (0.0134907)** |
| **235691_at** | **LOC729970[a]** | **Similar to hCG2028352** | **0.0034 (0.00337268)** |
| **238596_at** | **C10orf4[a]** | **Chromosome 10 open reading frame 4** | **0.0034 (-0.00252951)** |
| 200776_s_at | KIAA0005 | KIAA0005 gene product | 0.0025 |
| **220459_at** | **MCM3APAS[a]** | **MCM3AP antisense RNA (non-protein coding)** | **0.0025 (0.00421585)** |
| **223319_at** | **GPHN[a]** | **Gephyrin** | **0.0025 (-0.00505902)** |
| **225867_at** | **VASN[a]** | **Vasorin** | **0.0025 (-0.00505902)** |
| **232975_at** | **HCG18[a]** | **HLA complex group 18** | **0.0025 (0.00252951)** |
| **200820_at** | **PSMD8[a]** | **Proteasome (prosome, macropain) 26S subunit, non-ATPase, 8** | **0.0017 (-0.00421585)** |
| **201590_x_at** | **ANXA2[a]** | **Annexin A2** | **0.0017 (-00337268)** |
| **203234_at** | **UPP1[a]** | **Uridine phosphorylase 1** | **0.0017 (0.0109612)** |
| **205918_at** | **SLC4A3[a]** | **Solute carrier family 4, anion exchanger, member 3** | **0.0017 (-0.00084317)** |
| 211976_at | | FLJ22515 fis | 0.0017 |
| **212171_x_at** | **VEGFA[a]** | **Vascular endothelial growth factor A** | **0.0017 (0.0193929)** |
| **212240_s_at** | **PIK3R1[a]** | **Phosphoinositide-3-kinase, regulatory subunit 1 (α)** | **0.0017 (0.0084317)** |
| 218540_at | THTPA | Thiamine triphosphatase | 0.0017 |
| 221839_s_at | UBAP2 | Ubiquitin associated protein 2 | 0.0017 |
| 225367_at | PGM2 | Phosphoglucomutase 2 | 0.0017 |
| 201798_s_at | MYOF | Myoferlin | 0.0008 |
| 203957_at | E2F6 | E2F transcription factor 6 | 0.0008 |
| 206172_at | IL13RA2 | Interleukin 13 receptor, alpha 2 | 0.0008 |
| 213011_s_at | TPI1 | Triosephosphate isomerase 1 | 0.0008 |
| 213309_at | PLCL2 | Phospholipase C-like 2 | 0.0008 |
| 215566_x_at | LYPLA2 | Lysophospholipase II | 0.0008 |
| 218145_at | TRIB3 | Tribbles homolog 3 (*Drosophila*) | 0.0008 |
| 219194_at | SEMA4G | Sema domain, immunoglobulin domain (Ig), transmembrane domain (TM) and short cytoplasmic domain, (semaphorin) 4G | 0.0008 |
| 228253_at | LOXL3 | Lysyl oxidase-like 3 | 0.0008 |
| 231967_at | PHF20L1 | PHD finger protein 20-like 1 | 0.0008 |
| 238021_s_at | CRNDE | Colorectal neoplasia differentially expressed (non-protein coding) | 0.0008 |
| 238563_at | | ESTs | 0.0008 |
| 239144_at | B3GAT2 | β-1,3-glucuronyltransferase 2 (glucuronosyltransferase S) | 0.0008 |
| 240806_at | RPL15 | Ribosomal protein L15 | 0.0008 |
| 243024_at | ZNF789 | Zinc finger protein 789 | 0.0008 |
| 244688_at | | ESTs | 0.0008 |
| 1553645_at | CCDC141 | Coiled-coil domain containing 141 | 0.0000 |
| 1554340_a_at | C1orf187 | Chromosome 1 open reading frame 187 | 0.0000 |
| 1559529_at | PTK2 | PTK2 protein tyrosine kinase 2 | 0.0000 |
| 203620_s_at | FCHSD2 | FCH and double SH3 domains 2 | 0.0000 |
| 203834_s_at | TGOLN2 | Trans-golgi network protein 2 | 0.0000 |
| 203930_s_at | MAPT | Microtubule-associated protein tau | 0.0000 |
| 205547_s_at | TAGLN | Transgelin | 0.0000 |
| 207198_s_at | LIMS1 | LIM and senescent cell antigen-like domains 1 | 0.0000 |
| 211956_s_at | EIF1 | Eukaryotic translation initiation factor 1 | 0.0000 |
| 215952_s_at | OAZ1 | Ornithine decarboxylase antizyme 1 | 0.0000 |

Table I. Continued.

| Probe | Symbol | Description | VI |
|---|---|---|---|
| 217936_at | ARHGAP5 | Rho GTPase activating protein 5 | 0.0000 |
| 218454_at | PLBD1 | Phospholipase B domain containing 1 | 0.0000 |
| 220988_s_at | C1QTNF3 | C1q and tumor necrosis factor related protein 3 | 0.0000 |
| 226186_at | TMOD2 | Tropomodulin 2 (neuronal) | 0.0000 |
| 229178_at | PRTG | Protogenin homolog (Gallus gallus) | 0.0000 |
| 230826_at | MMD2 | Monocyte to macrophage differentiation-associated 2 | 0.0000 |
| 233117_at | | FLJ14328 fis, clone PLACE4000252 | 0.0000 |
| 238360_s_at | | ESTs | 0.0000 |
| 55662_at | C10orf76 | Chromosome 10 open reading frame 76 | 0.0000 |
| 200827_at | PLOD1 | Procollagen-lysine 1, 2-oxoglutarate 5-dioxygenase 1 | -0.0008 |
| 202185_at | PLOD3 | Procollagen-lysine, 2-oxoglutarate 5-dioxygenase 3 | -0.0008 |
| 202709_at | FMOD | Fibromodulin | -0.0008 |
| 208740_at | SAP18 | Sin3A-associated protein, 18 kDa | -0.0008 |
| 217933_s_at | LAP3 | Leucine aminopeptidase 3 | -0.0008 |
| 225986_x_at | CPSF2 | leavage and polyadenylation specific factor 2, 100 kDa | -0.0008 |
| 226547_at | MYST3 | MYST histone acetyltransferase (monocytic leukemia) 3 | -0.0008 |
| 227407_at | TAPT1 | Transmembrane anterior posterior transformation 1 | -0.0008 |
| 227719_at | SMAD9 | SMAD family member 9 | -0.0008 |
| 228906_at | TET1 | Tet oncogene 1 | -0.0008 |
| 230987_at | | ESTs | -0.0008 |
| 231031_at | KGFLP2 | Keratinocyte growth factor-like protein 2 | -0.0008 |
| 237817_at | SSR3 | Signal sequence receptor, γ (translocon-associated protein γ) | -0.0008 |
| 238603_at | LOC254559 | Hypothetical LOC254559 | -0.0008 |
| 201676_x_at | PSMA1 | Proteasome (prosome, macropain) subunit, α type, 1 | -0.0017 |
| 219648_at | MREG | Melanoregulin | -0.0017 |
| 221589_s_at | ALDH6A1 | Aldehyde dehydrogenase 6 family, member A1 | -0.0017 |
| 1554018_at | GPNMB | Glycoprotein (transmembrane) nmb | -0.0025 |
| 207805_s_at | PSMD9 | Proteasome (prosome, macropain) 26S subunit, non-ATPase, 9 | -0.0025 |
| 212695_at | CRY2 | Cryptochrome 2 (photolyase-like) | -0.0025 |
| 241255_at | | ESTs | -0.0025 |
| 201576_s_at | GLB1 | Galactosidase, β 1 | -0.0034 |
| 200986_at | SERPING1 | Serpin peptidase inhibitor, clade G (C1 inhibitor), member 1 | -0.0042 |
| 202207_at | ARL4C | ADP-ribosylation factor-like 4C | -0.0042 |
| 224711_at | YY1 | YY1 transcription factor | -0.0042 |

VI, variable importance. [a]The top 15 genes are shown in bold. Parenthesis in VI, the variable importance score based on reduced set of genes.

*Further optimization of gene set predictor.* To optimize the multigene predictor for application in clinical samples, the top 15 genes were selected on the basis of the strength and significance of their survival association as presented in Table I with asterix and bold, by applying the random survival forests model again from 82-gene set with symbol name. P-values for the 15-gene set selected using Kaplan-Meier curves were $p<0.0001$ (Q=30) (Fig. 3D), but were slightly worse compared with the entire 82-gene set. In 13 grade III cases, 1 patient was in the poor-prognosis group (Fig. 3C) and 2 were in the poor-prognosis group (Fig. 3D). In 32 grade IV cases, 14 patients were in the good-prognosis group (Fig. 3C) and 12 were in the good-prognosis group (Fig. 3D).

The 15-gene profile was tested for prediction of outcome in the independent data set groups (Table II). Kaplan-Meier curves comparing groups classified by the fitted random survival forests model with the 15-gene model in these data set are shown in Fig. 4. Our signature was validated in 3 out of 4 independent studies, although neither the 82 nor the 15 gene list was predictive in the Freije dataset. The p-value of combined tests was $p<0.0001$, which indicates that the selected set of genes was significantly predictive. There were no differences of expression in 15 selected genes between anaplastic astrocytomas and anaplastic oligo-astrocytoma or anaplastic oligodendroglioma (data not shown). Thus, 15 genes that have been identified as a predictor for better prognosis could not be the one just overexpressed in oligoden-

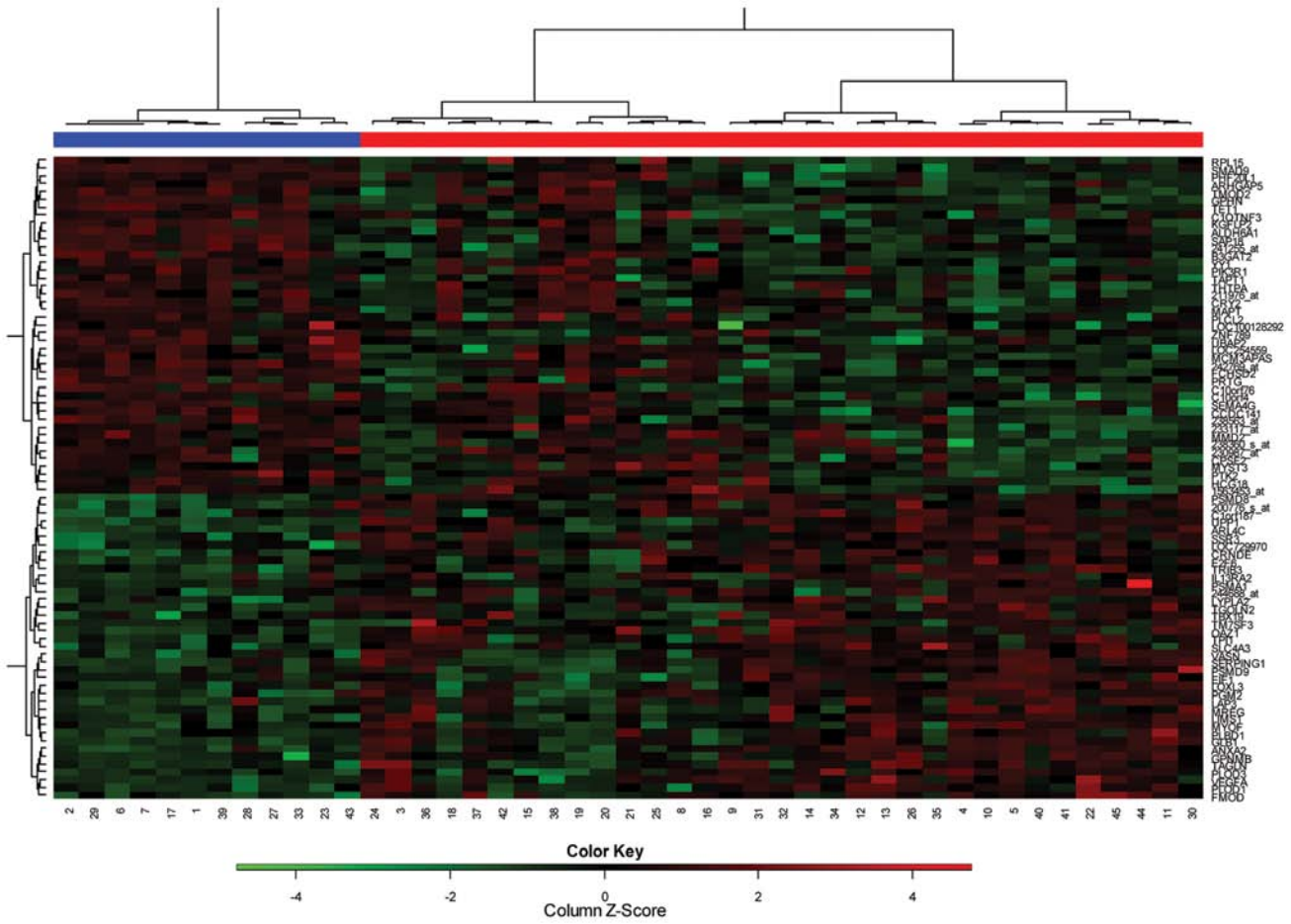


Figure 2. Heatmap for the selected genes. The blue and red bars at the top indicate the two groups classified by the fitted random survival forests model. The heatmap consists of patients sorted by the estimated mortalities in the column and genes sorted by expressions in the row. The blue bar at the top represents classified patients as poor survivors. The red bar at the top represents classified patients as good survivors.

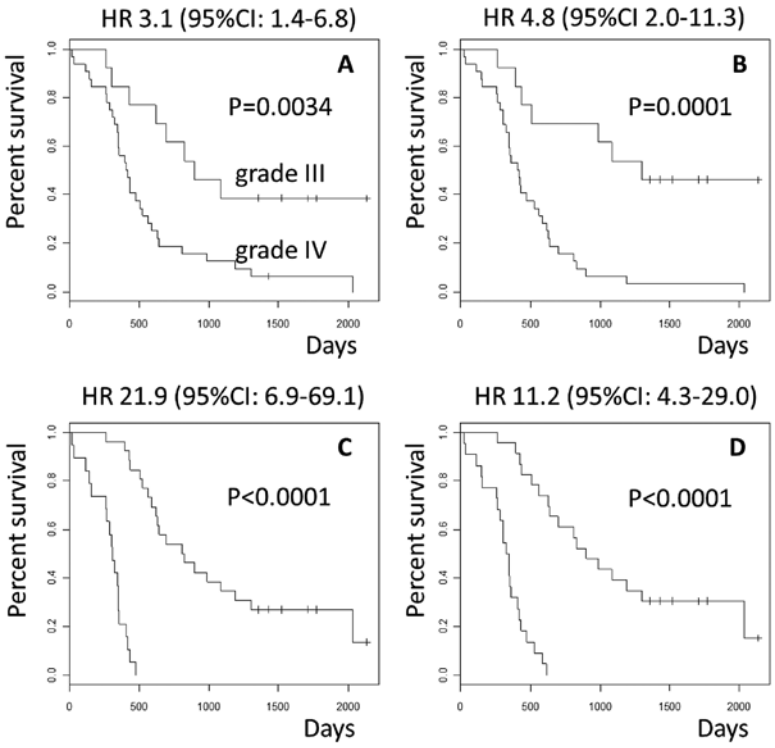


Figure 3. (A) Kaplan-Meier curves comparing WHO grades III and IV. (B) Comparison of groups classified by the clustering analysis based on the gene expressions selected by the SAM. (C) Comparison of groups classified by the fitted random survival forests model with the 82-gene model. (D) Comparison of groups classified by the fitted random survival forests model with the 15-gene model.

Table II. Performance of prediction in four independent data sets.

| | 82 genes | | | | | 15 genes | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | C. index | SE | 95% CI | | p-value | C. index | SE | 95% CI | | p-value |
| TCGA (n=347) | 0.5506 | 0.0188 | 0.5137 | 0.5874 | 0.0036 | 0.5547 | 0.0189 | 0.5177 | 0.5917 | 0.0019 |
| Petalidis *et al* (26) (n=57) | 0.6671 | 0.0390 | 0.5907 | 0.7435 | <0.0001 | 0.6771 | 0.0333 | 0.6119 | 0.7423 | <0.0001 |
| Phillips *et al* (37) (n=76) | 0.6273 | 0.0400 | 0.5489 | 0.7057 | 0.0007 | 0.6213 | 0.0388 | 0.5452 | 0.6974 | 0.0009 |
| Freije *et al* (25) (n=85) | 0.5116 | 0.0384 | 0.4364 | 0.5869 | 0.3810 | 0.5478 | 0.0408 | 0.4678 | 0.6278 | 0.1209 |

C. index, concordance index; SE, standard error; CI, confidence interval.

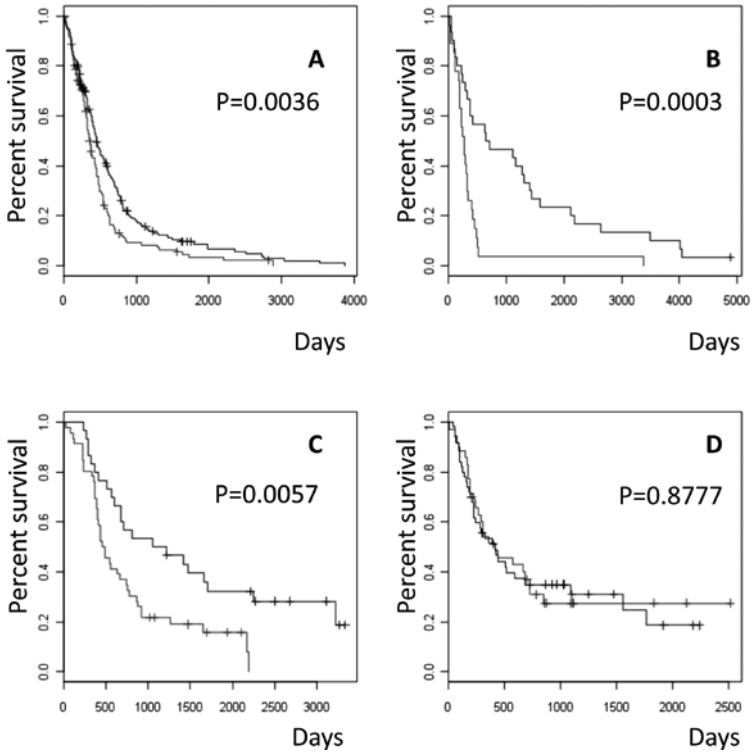


Figure 4. Kaplan-Meier curves comparing groups classified by the fitted random survival forests model with the 15-gene model in (A) TCGA, (B) Petalidis (26), (C) Phillips (37) and (D) Freije (25).

droglioma. The 15-gene profile only performed slightly worse than the 82 gene set in the Phillips *et al* (37), dataset. Therefore, the 15-gene set contains a subpopulation of genes that predict the survival of patients similarly to the 82-gene set.

*Gene classifiers of particular biological interest*. In addition to the identification of 82 genes, PI3KR1, TAGLN, and SERPING 1 were of particular biological interest and/or novelty. Their expression changes were validated by both QPCR and immunohistochemistry (Fig. 5). These genes were also found to be differentially expressed between short- (survival time <2 years) and long-term (survival time >2 years) survivors ($p<0.01$; Fig. 5A). A 15-gene set, was successfully validated by QPCR (data not shown). Representative immunohistochemistry results for PI3KR1 showed positive staining on grade II and negative staining on grade IV tumors. TAGLN and SERPING 1 showed positive staining on grade IV and negative staining on grade II, III. PI3KR1, TAGLN, and SERPING1 showed cytoplasmic staining (Fig. 5B).

## Discussion

Several studies have been reported on gene expression profiles of malignant gliomas using SAM. Our study has a small number of samples, however, this is the first study using gene-gene interactions in the gene expression prognostic classification context in glioma patients. We assessed the relationship between gene expressions and survival time by using the random survival forests model and it was a better classifier compared to SAM.

For this purpose, there may be several other choices for the statistical analysis. In Petalidis *et al* and Phillips *et al* studies (26,37), Pearson's correlation coefficient was used to assess

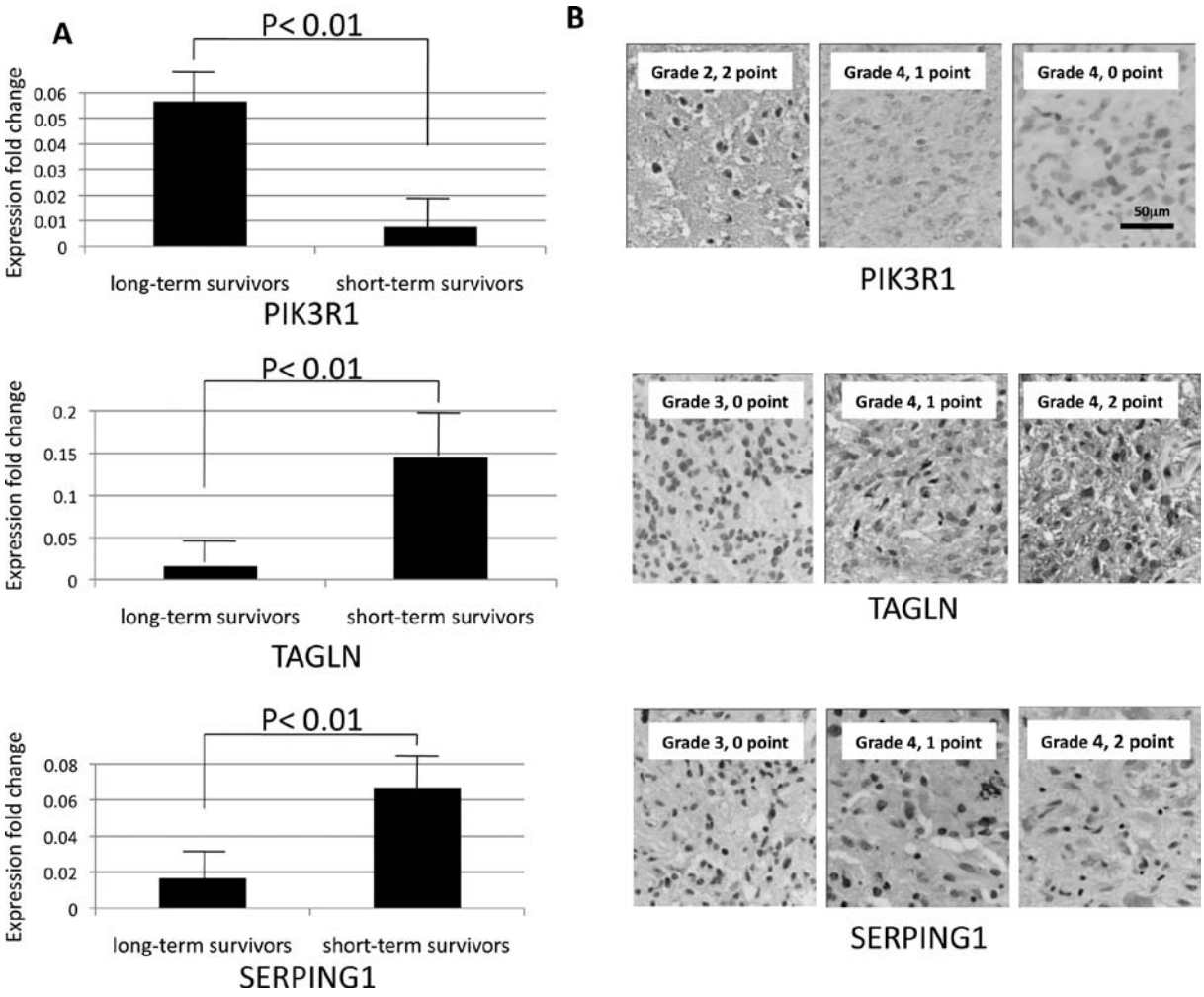


Figure 5. Expression of PI3KR1, TAGLN, and SERPING1 changes with tumor progression at both transcript and protein levels. (A) Validation of expression changes as assessed by QPCR expression technology. Mean expression fold changes between short- (survival time <2 years; n=30) and long-term (survival time >2 years; n=14) survivors. (B) Representative immunohistochemistry results for PI3KR1, TAGLN, and SERPING 1 on grade II, III, and IV tumors. PI3KR1, TAGLN, and SERPING1 showed cytoplasmic staining.

associations of individual genes with survival time. This disadvantage might be lack of consideration for censoring. As more appropriate method, the SAM and Rankproduct were able to deal with censoring. However, it is not possible to construct a functional form of genes representing the patient's prognosis used in the prediction. As discussed in Cordell *et al* (42), the functional form should contain gene by gene interaction terms. Although the multiple regression model such as the Cox regression model and partial least squares (PLS) used in Freije *et al* study (25) is applicable for this purpose, the correlation among genes may induce the multicolinearlity in the Cox model. The PLS is appropriate for the correlation but it is difficult to incorporate higher order interactions because of a finite sample size. The random forests method is classified into the tree-based method which has an advantage in detecting interaction. It has been developed to apply to data with several variables (genes) much larger than the number of patients. In this regard, the framework of random forests which overcomes this problem would be necessary in the analysis. In the overlap analysis, there may a few identified signatures. We consider the different approach to assess the relationship to survival time as the reason. We applied essentially the same method for variable selection to Diaz-Uriarte *et al* (43). In a recent report (44), other methods such as minimal depth and variable hunting were introduced and the applicability discussed (available in the recent version of randomSurvivalForest package). Their consideration was based on the open microarray datasets with relatively large sample size (the minimum was 78). Including the method in this paper, comparison between methods in small sample size would be helpful but beyond the scope of this report.

The primary role of the 82-gene and 15-gene panel for optimization of therapy would be to prospectively identify patients who are more likely to have durable survival to standard therapy. Eighty-two-gene set contains a large population of genes that predicts survival of patients, and the 15-gene set contains a subpopulation of genes that similarly predicts the survival of patients. Although these set of 15 and 82 genes were not particularly unique and that other subsets of genes would likely perform similarly. Among our candidate genes, two are of particular interest, VEGF and PI3KR1. Malignant gliomas display striking vascularity with high expression of vascular endothelial growth factor (VEGF), a key growth factor for new blood vessel formation (45). Also, recent clinical trials combining bevacizumab (an anti-VEGF-A antibody), with

chemotherapy reported very encouraging response rates (46). PI3K activates downstream target molecules such as AKT and the mammalian target of rapamycin (mTOR), which results in cell proliferation and survival of glioma cells (38,47). Ruano *et al* reported the activation of the PI3K/Akt pathway was survival-related in glioblastoma patients (48). The PI3K complex is activated by upstream signals from receptor tyrosine kinase (RTK), these consist of a p110α, encoded by PIK3CA, and a regulatory protein, p85α, encoded by PIK3R1. In a TCGA (38) cohort, nine PIK3R1 somatic mutations were detected among the 91 glioblastomas. It is speculated that spatial constraints due to these mutations might prevent inhibitory contact of p85α with p110α, causing constitutive PI3K activity. In our study, expression of PIK3R1 and patient survival were inversely co-related. Low expression of PIK3R1 might contribute to constitutive PI3K activity in malignant glioma. Thus, we are now trying to reveal the biological roles of the transcripts of these interesting candidate genes in glioma. Focal adhesion and leukocyte transendothelial migration were characteristics of malignant glioma by GSEA analysis. These pathways are related to mesenchymal transformation of gliomas which is malignant phenotype of the disease (37,49).

Although our predictor was mainly based on cases from 1st line nitrosourea-based chemotherapy, the results with four external data sets (25,26,37,38) where 1st line temozolomide-based chemotherapy was carried out, support the universal performance of the predictor, irrespective of chemotherapeutic regimen. Survival benefit by chemotherapy is relatively small in most malignant gliomas; therefore it is important to elucidate the differences in the intrinsic biological characters of the tumors. In addition, genetic differences within malignant gliomas underpin the heterogeneity of these tumor types.

The value of gene-expression-based predictors for prognosis of malignant glioma patients will not be fully realized until additional therapies are available for patients destined to have poor survival following conventional chemotherapy. In this regard, expression profiles might not only predict the likelihood of long-term survival, but might also yield clues on individual genes involved in tumor development, progression, and response to therapy (37). Moreover, the ability to distinguish histologically-ambiguous gliomas will enable appropriate therapies to be tailored to specific tumor subtypes. Class prediction models based on defined molecular profiles allow classification of malignant gliomas in a manner that will better correlate with clinical outcomes. Therefore, identification of these molecular subclasses of glioma could greatly facilitate prognosis and our ability to develop effective treatment protocols.

In conclusion, we identified gene signatures associated with outcome in patients with glioma. Adaptation of subsets of these genes for use in clinical assays could result in improved outcome prediction. We have extended our observations in the validation of these signatures in independent data sets from other institutions. In conclusion, our profiling results will help to construct a new classification scheme that better assesses clinical malignancies.

## Acknowledgments

There is no potential conflict of interest. This work was supported in part by Grants-in-Aid from the Ministry of Education, Culture, Sport, Science and Technology of Japan (21700312 to A.K., 14370428 and 17390394 to R.Y.).

## References

1. Stewart LA: Chemotherapy in adult high-grade glioma: a systematic review and meta-analysis of individual patient data from 12 randomised trials. Lancet 359: 1011-1018, 2002.
2. Stupp R, Hegi ME, Mason WP, *et al*: Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial. Lancet Oncol 10: 459-466, 2009.
3. Kleihues P, Louis DN, Wiestler OD, *et al*: WHO grading of tumours of the central nervous system. In: World Health Organization Classification of Tumours of the Nervous System. Louis DN, Ohgaki H, Wiestler OD and Cavenee WK (eds). IARC Press, Lyon, pp10-11, 2007.
4. Ramaswamy S, Tamayo P, Rifkin R, *et al*: Multiclass cancer diagnosis using tumor gene expression signatures. Proc Natl Acad Sci USA 98: 15149-15154, 2001.
5. Rickman DS, Bobek MP, Misek DE, *et al*: Distinctive molecular profiles of high-grade and low-grade gliomas based on oligonucleotide microarray analysis. Cancer Res 61: 6885-6891, 2001.
6. Kim S, Dougherty ER, Shmulevich I, *et al*: Identification of combination gene sets for glioma classification. Mol Cancer Ther 1: 1229-1236, 2002.
7. Khan J, Wei JS, Ringner M, *et al*: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 7: 673-679, 2001.
8. Mischel PS, Shai R, Shi T, *et al*: Identification of molecular subtypes of glioblastoma by gene expression profiling. Oncogene 22: 2361-2273, 2003.
9. Nigro JM, Misra A, Zhang L, *et al*: Integrated array-comparative genomic hybridization and expression array profiles identify clinically relevant molecular subtypes of glioblastoma. Cancer Res 65: 1678-1686, 2005.
10. Shai R, Shi T, Kremen TJ, *et al*: Gene expression profiling identifies molecular subtypes of gliomas. Oncogene 22: 4918-4923, 2003.
11. Sorlie T, Tibshirani R, Parker J, *et al*: Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci USA 100: 8418-8423, 2003.
12. Liang Y, Diehn M, Watson N, *et al*: Gene expression profiling reveals molecularly and clinically distinct subtypes of glioblastoma multiforme. Proc Natl Acad Sci USA 102: 5814-5819, 2005.
13. Wong KK, Chang YM, Tsang YT, *et al*: Expression analysis of juvenile pilocytic astrocytomas by oligonucleotide microarray reveals two potential subgroups. Cancer Res 65: 76-84, 2005.
14. Shirahata M, Iwao-Koizumi K, Saito S, *et al*: Gene expression-based molecular diagnostic system for malignant gliomas is superior to histological diagnosis. Clin Cancer Res 13: 7341-7356, 2007.
15. Li A, Walling J, Ahn S, *et al*: Unsupervised analysis of transcriptomic profiles reveals six glioma subtypes. Cancer Res 69: 2091-2099, 2009.
16. Sallinen SL, Sallinen PK, Haapasalo HK, *et al*: Identification of differentially expressed genes in human gliomas by DNA microarray and tissue chip techniques. Cancer Res 60: 6617-6622, 2000.
17. Godard S, Getz G, Delorenzi M, *et al*: Classification of human astrocytic gliomas on the basis of gene expression: a correlated group of genes with angiogenic activity emerges as a strong predictor of subtypes. Cancer Res 63: 6613-6625, 2003.
18. Rich JN, Hans C, Jones B, *et al*: Gene expression profiling and genetic markers in glioblastoma survival. Cancer Res 65: 4051-4058, 2005.
19. Somasundaram K, Reddy SP, Vinnakota K, *et al*: Upregulation of ASCL1 and inhibition of Notch signaling pathway characterize progressive astrocytoma. Oncogene 24: 7073-7083, 2005.
20. Horvath S, Zhang B, Carlson M, *et al*: Analysis of oncogenic signaling networks in glioblastoma identifies ASPM as a molecular target. Proc Natl Acad Sci USA 103: 17402-17407, 2006.
21. Yamanaka R, Arao T, Yajima N, *et al*: Identification of expressed genes characterizing long-term survival in malignant glioma patients. Oncogene 25: 5994-6002, 2006.
22. Soroceanu L, Kharbanda S, Chen R, *et al*: Identification of IGF2 signaling through phosphoinositide-3-kinase regulatory subunit 3 as a growth-promoting axis in glioblastoma. Proc Natl Acad Sci USA 104: 3466-3471, 2007.

23. Reddy SP, Britto R, Vinnakota K, *et al*: Novel glioblastoma markers with diagnostic and prognostic value identified through transcriptome analysis. Clin Cancer Res 14: 2978-2987, 2008.
24. Nutt CL, Mani DR, Betensky RA, *et al*: Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res 63: 1602-1607, 2003.
25. Freije WA, Castro-Vargas FE, Fang Z, *et al*: Gene expression profiling of gliomas strongly predicts survival. Cancer Res 64: 6503-6510, 2004.
26. Petalidis LP, Oulas A, Backlund M, *et al*: Improved grading and survival prediction of human astrocytic brain tumors by artificial neural network analysis of gene expression microarray data. Mol Cancer Ther 7: 1013-1024, 2008.
27. Marko NF, Toms SA, Barnett GH and Weil R: Genomic expression patterns distinguish long-term from short-term glioblastoma survivors: a preliminary feasibility study. Genomics 91: 395-406, 2008.
28. Livak KJ and Schmittgen TD: Analysis of relative gene expression data using real-time quantitative PCR and the 2(-Delta Delta C(T)) method. Methods 25: 402-408, 2001.
29. R Development Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2009 ISBN 3-900051-07-0, URL http://www.R-project.org.
30. Gentleman R, Carey V, Bates D, *et al*: Bioconductor: open software development for computational biology and bioinformatics. Genome Biol 5: R80, 2004.
31. Ishwaran H, Kogalur UB, Blackstone EH and Lauer MS: Random survival forests. Ann Appl Statist 2: 841-860, 2008.
32. Harrel FE Jr, Lee KL and Mark DB: Tutorial in biostatistics: multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing error. Stat Med 15: 361-387, 1996.
33. Ripley BD: Pattern recognition and neural networks. Cambridge: Cambridge University Press, 2008.
34. Kaplan EL and Meier P: Nonparametric estimation from incomplete observations. J Am Statist Assoc 53: 457-481, 1958.
35. Efron B and Tibshirani R: On testing the significance of sets of genes. Ann Appl Stat 1: 107-129, 2007.
36. Kanehisa M, Goto S, Furumichi M, *et al*: KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res 38: 355-360, 2010.
37. Phillips HS, Kharbanda S, Chen R, *et al*: Molecular subclasses of high-grade glioma predict prognosis, delineate a pattern of disease progression, and resemble stages in neurogenesis. Cancer Cell 9: 157-173, 2006.
38. Cancer Genome Atlas Research Network: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061-1068, 2008.
39. Haibe-Kains B, Desmedt C, Sotiriou C and Bontempi G: A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? Bioinformatics 24: 2200-2208, 2008.
40. Breiman L: Random forests. Machine Learning 45: 5-32, 2001.
41. Tusher VG, Tibshirani R, Chu G: Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci USA 98: 5116-5121, 2001.
42. Cordell HJ: Detecting gene-gene interactions that underlie human diseases. Nat Rev Genet 10: 392-404, 2009.
43. Diaz-Uriarte R and Alvarez de Andres S: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7: 3, 2006.
44. Ishwaran H, Kogalur UB, Gorodeski EZ, *et al*: High-dimensional variable selection for survival data. J Am Statist Assoc 105: 205-217, 2010.
45. Hlobilkova A, Ehrmann J, Knizetova P, *et al*: Analysis of VEGF, Flt-1, Flk-1, nestin and MMP-9 in relation to astrocytoma pathogenesis and progression. Neoplasma 56: 284-290, 2009.
46. Vredenburgh JJ, Desjardins A, Herndon JE II, *et al*: Bevacizumab plus irinotecan in recurrent glioblastoma multiforme. J Clin Oncol 25: 4722-4729, 2007.
47. Cheng CK, Fan QW and Weiss WA: PI3K signaling in glioma-animal models and therapeutic challenges. Brain Pathol 19: 112-120, 2009.
48. Ruano Y, Mollejo M, Camacho FI, *et al*: Identification of survival-related genes of the phosphatidylinositol 3'-kinase signaling pathway in glioblastoma multiforme. Cancer 112: 1575-1584, 2008.
49. Carro MS, Lim WK, Alvarez MJ, *et al*: The transcriptional network for mesenchymal transformation of brain tumours. Nature 463: 318-325, 2010.