Gene Expression Profiling and Genetic Markers in Glioblastoma Survival

Jeremy N. Rich,[1,2,5] Christopher Hans,[5,6] Beatrix Jones,[5,6,8] Edwin S. Iversen,[5,6] Roger E. McLendon,[3,5] B.K. Ahmed Rasheed,[3,5] Adrian Dobra,[4,6,7] Holly K. Dressman,[4,5,7] Darell D. Bigner,[3,5] Joseph R. Nevins,[4,5,7] and Mike West[5,6,7]

Departments of [1]Medicine, [2]Neurobiology, [3]Pathology, and [4]Molecular Genetics and Microbiology, [5]W.M. Keck Center for Neuro-Oncogenomics, Institutes of [6]Statistics and Decision Sciences, and [7]Genome Sciences and Policy, Duke University Medical Center, Durham, North Carolina; and [8]Institute of Information and Mathematical Sciences, Massey University, New Zealand

Abstract

Despite the strikingly grave prognosis for older patients with glioblastomas, significant variability in patient outcome is experienced. To explore the potential for developing improved prognostic capabilities based on the elucidation of potential biological relationships, we did analyses of genes commonly mutated, amplified, or deleted in glioblastomas and DNA microarray gene expression data from tumors of glioblastoma patients of age >50 for whom survival is known. No prognostic significance was associated with genetic changes in epidermal growth factor receptor (amplified in 17 of 41 patients), TP53 (mutated in 11 of 41 patients), $p16^{INK4A}$ (deleted in 15 of 33 patients), or phosphatase and tensin homologue (mutated in 15 of 41 patients). Statistical analysis of the gene expression data in connection with survival involved exploration of regression models on small subsets of genes, based on computational search over multiple regression models with cross-validation to assess predictive validity. The analysis generated a set of regression models that, when weighted and combined according to posterior probabilities implied by the statistical analysis, identify patterns in expression of a small subset of genes that are associated with survival and have value in assessing survival risks. The dominant genes across such multiple regression models involve three key genes—SPARC (Osteonectin), Doublecortex, and Semaphorin3B—which play key roles in cellular migration processes. Additional analysis, based on statistical graphical association models constructed using similar computational analysis methods, reveals other genes which support the view that multiple mediators of tumor invasion may be important prognostic factor in glioblastomas in older patients. (Cancer Res 2005; 65(10): 4051-8)

Introduction

Glioblastomas remain one of the most lethal forms of cancers with a median survival of 10 to 12 months (1). Whereas the number of patients diagnosed with primary brain tumors remains relatively small—18,500 Americans are expected to be diagnosed in 2005 (2), the morbidity and mortality of these tumors are severe. Unlike most other types of cancer, glioblastomas rarely metastasize; rather, they induce death through striking resistance to current therapies and invasion into normal brain tissues (3). Gliomas are graded based on the presence of specific histologic markers, including necrosis, nuclear pleomorphism, mitotic activity, and vascular proliferation (4). Among clinical markers, age and Karnofsky performance status are prognostic (5). Among treatment options, gross total resection (6) and radiation therapy have been shown to improve survival with limited benefit to chemotherapy (7). In many malignant glioma clinical trials, tumor histology, or patient age impact patient outcome more significantly than the therapy under investigation. Novel therapies to specific molecular targets are currently under development for many cancers including glioblastomas (8), and advances in such approaches will require the determination of the roles that specific gene products play in glioblastoma pathophysiology.

At least two genetic pathways have been delineated in glioblastoma development: *de novo* and secondary glioblastomas (9). *De novo* glioblastomas represent the most frequent presentation with an initial diagnosis of glioblastoma without evidence of preexistent lower grade tumor. These patients are commonly of older age and have a high rate of epidermal growth factor receptor (EGFR) amplification, $p16^{INK4A}$ deletion, and phosphatase and tensin homologue deleted on chromosome 10 (PTEN; mutated in multiple advanced cancers 1) mutations. In contrast, secondary glioblastomas arise after a preceding diagnosis of lower grade tumors. TP53 and RB mutations are thought to be more common in the development of secondary glioblastomas (9). Despite these genetic differences, no significant differentiation in patient survival has been noted between *de novo* and secondary glioblastomas when controlled for age. In fact, there have been no widely validated prognostic genetic markers for glioblastoma patients. Rather, several genetic changes, including PTEN and EGFR mutations, have been linked to poor prognosis in patients with anaplastic astrocytomas (10), suggesting that these are markers of transformation to glioblastomas.

Molecular profiles of glioma patient specimens have suggested that gene expression may predict patient outcome more accurately than pathologic measures (11–14). These analyses have provided large sets of genes which may be expected to regulate the process of tumor progression. To explore genome-scale expression information for potential value in defining contributors to the malignancy of gliomas with the worst prognosis—glioblastoma patients over the age of 50—we examined tumor RNA in relation to patient survival.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Requests for reprints: Jeremy N. Rich, Duke University Medical Center, Box 2900, Durham, NC 27710. Phone: 919-681-1693; Fax: 919-684-6514; E-mail: rich0001@mc.duke.edu.

Affymetrix gene chip analysis of 41 tumor specimens was examined using computational statistical methods to explore the potential for generating gene expression–based markers of survival, and to elucidate expression-based associations among any genes showing such potential. Additional analysis of the full genome-scale gene expression data using statistical graphical models that define empirical association networks over genes leads to the identification of additional genes linked to those arising in the primary predictive models. These results have been contrasted with traditional DNA studies including measurement of EGFR amplification, mutational analysis of EGFR, TP53, and PTEN, and loss of heterozygosity detection at 9p, 10p, 10q, and 17p.

Materials and Methods

Patient characteristics. The sample of 41 patients (age over 50, with sufficient resected tissue for expression analysis, and uniform surgical interventions) are summarized in Table 1. Two of the 41 patients were diagnosed with secondary glioblastomas—lower grade tumors prior to the diagnosis of glioblastoma (one anaplastic astrocytoma, one anaplastic oligodendroglioma). Both of these patients were diagnosed with grade 3 tumors less than 1 year before the diagnosis of glioblastoma. Postsurgical patient treatment information was available from medical chart review on 38 of 41 patients. Only 3 of 38 patients did not undergo additional radiation treatment (two due to poor clinical status at presentation), each suffered a rapid clinical decline and death. Median survival among patients who did not receive external beam radiation was 4.9 months, whereas median survival among the patients who received external beam radiation was 18.55 months. Eight patients underwent liquid brachytherapy with a radiolabeled monoclonal antibody (81C6; ref. 15). One patient underwent only liquid brachytherapy without external beam radiation. Median survival among patients who did not receive brachytherapy was 9.5 months, whereas median survival among those who received brachytherapy was 23.55 months. Two patients underwent nitrosourea wafer implantation (Gliadel) and one convection-enhanced delivery of a growth factor ligand-toxin chimera (TP-38). Twenty-nine of 38 patients underwent adjuvant systemic chemotherapy—one additional patient was treated with nitrosourea wafers without later systemic chemotherapy—from 1 to 7 therapeutic regimens (mean 3.0) with 1 to 17 total cycles. Patients receiving chemotherapy were categorized into two categories (one or two regimens versus three or more). Regimens commonly included a nitrosourea, temozolomide, or a topoisomerase inhibitor. External beam radiation and liquid brachytherapy were significantly associated with increased patient survival, whereas systemic chemotherapy was not (Table 1). Thus, the majority of patients were treated in a similar fashion with external beam radiation and some form of chemotherapy.

Case identification/sample collection. Cases were obtained from a survival-based study run under the auspices of the W.M. Keck Center for Neuro-Oncology at Duke University. Each block used in the analysis was independently validated for the presence of >95% tumor and graded by a neuropathologist (R.E. McLendon) using the Nelson/Burger criteria for the presence of necrosis for the diagnosis of glioblastoma. The sample of 41 glioblastomas was collected specifically from patients >50 years of age in order to bias the sampling to primary glioblastoma.

PCR-based molecular analysis. Normal DNA was extracted from lymphocytes. Tumor DNA and RNA was isolated from sections cut from the frozen block. Exons 5 to 8 of the TP53 gene and all 9 exons of PTEN were resequenced by capillary electrophoresis on ABI 3100. EGFR DNA amplification assay was done by co-PCR amplifying a 3′ untranslated region fragment of EGFR gene with a fragment of exon 3 of IFNG gene as internal control, using fluorescent tagged primers. EGFR/IFNG peak area ratios of >5 are considered as indication of EGFR gene amplification. CDKN2A ($p16^{INK4A}$) deletion assay was carried out by SYBR Green fluorescent assay on ABI 7900HT. A ΔCt CDKN2A-Globin (internal control) value of >1.7 was considered indicative of homozygous deletion of CDKN2A. Loss of heterozygosity analysis of 9p, 10, and 17p was done by comparing allele intensities of PCR amplified loci (three from each arm) from tumor and corresponding patient's lymphocyte DNA. Fragment analysis was done by capillary electrophoresis on ABI 3100. Peak height ratios (tumor/blood) <0.65 or >1.67 were considered indicative of loss of heterozygosity.

To detect the EGFR vIII variant, RNA extracted from tumor tissue was reverse-transcribed using Invitrogen Superscript II kit and PCR-amplified using primers from exons 1 and 8. The PCR products were electrophoresed in a 3% agarose gel. This assay generates a 111-bp product in vIII variants and a 912-bp product in the wild-type. Levels of SPARC and doublecortex (DCX) transcripts were assayed by SYBR Green fluorescent assay on ABI 7900HT. Normalization of input cDNA amount was done by comparing amplification of housekeeping genes glyceraldehyde-3-phosphate dehydrogenase and β2-microglobulin. ΔCt values represent average Ct SPARC or DCX minus average Ct B2M or glyceraldehyde-3-phosphate dehydrogenase.

Microarray chip RNA hybridization procedures. Total RNA was extracted from tumor tissue with Qiagen (Valencia, CA) RNEasy kits, and assessed for quality with an Agilent Lab-on-a-Chip 2100 Bioanalyzer. Hybridization target probes were prepared from total RNA according to standard Affymetrix protocols and hybridized to the human U133A GeneChip (see Supplementary Materials for full details).

Data preprocessing prior to the formal statistical analysis involved standard processes of normalization, expression intensity estimation and screening for genes showing reasonable variation across samples. The Affymetrix U133a DNA microarrays provide assay of over 20,000 probe sets. The expression intensities for all genes across the 41 samples were

Table 1. Patient and tumor characteristics and association with survival

Patient Characteristic		Significance
Age	Mean 63 (range 50-78)	0.117
Sex	15 females, 26 males	0.665
Race	36 Whites, 5 Blacks	0.858
Secondary glioblastoma	2 of 41	0.976
Resection	41 of 41	NA
External beam radiation	34 of 38	0.017
Liquid brachytherapy	8 of 38	0.021
Chemotherapy	29 of 38	0.21
Number of regimens	Mean 3 (range 1-12)	0.539
Molecular Event	Present	
EGFR amplification	17 of 41 (41%)	0.711
EGFRvIII expression	18 of 40 (45%)	0.902
TP53 mutation	11 of 41 (27%)	0.291
PTEN mutation	15 of 41 (37%)	0.517
p16INK4A deletion	15/33 (45%)	0.286

NOTE: Patients are characterized based on age at original diagnosis. Significance was analyzed for each characteristic using the Mann-Whitney test except for age, which was tested using the significance of the slope coefficient in a regression of age on log (survival time). Median survival time was only significant for treatment with external beam radiation (median survival among those not receiving radiation, 4.9 months; median survival among those receiving external beam radiation, 18.55 months) and liquid brachytherapy with radiolabeled 81C6 (median survival among those not receiving brachytherapy, 9.5 months; median survival among those receiving brachytherapy, 23.55 months). EGFR was considered amplified if values were >5.0. EGFR was also tested for significance using the slope coefficient in a regression of log (EGFR DNA amplification) on log (survival time).

estimated using robust multi-array average (16), with probe-level quantile normalization, as implemented in the Bioconductor software suite (17). The resulting robust multi-array average expression intensity estimates were then screened to identify genes whose robust multi-array average levels probe vary at least 4-fold across the samples, and whose maximum level exceeded seven on the $\log^2$ scale, leading to P = 8,408 genes/probe sets whose robust multi-array average expression intensities are the candidate predictors in the regression model analysis and computational search.

Statistical analysis. The predictive analysis evaluated linear regression models of the form $y = a_0 + a_1x_1 + \ldots + a_kx_k + e$, where y represents log survival time, each x_i represents the expression level of gene i, k is a small integer, and e represents an unexplained, random component. The challenge of statistical analysis is to search for subsets of genes that together define significant predictive regressions—that is, to select both the number k of genes, or variables, and then the specific set of genes $x_1, \ldots, x_k$ by searching over subsets. This includes the possibility of no association with any genes, i.e., $k = 0$. Technically, with many genes available, this requires some form of stochastic search. The analysis is based on a so-called *shotgun stochastic search* (18), which in a distributed computer environment, allows the rapid evaluation of many such models so long as the search is constrained to values of k that are reasonably small. The parallel computational strategies implemented are very efficient and the search over models generally focuses quickly on subsets of relevant models with higher probability (if such a model exists).

Analysis here with n = 41 samples confirms that a number of models with three to four genes are of some interest. The analysis heavily penalizes more complex models, initially very strongly favoring the null hypothesis of no significant predictors in this model context among the thousands of genes in a manner that naturally counters the false discovery propensity of purely likelihood-based model search analyses. In addition, routine calculations confirm that the false-positive rate for discovery of single variable regressions as significant as those identified among the top candidates here is tiny. Of a number of regression models involving between three and five genes that are identified, many rely on overlapping sets of genes with two of the three "key" genes—SPARC, Doublecortex, and Semaphorin3B—appearing in a larger number of most highly scoring models. This reflects inherent collinearities among gene subsets, some of which is naturally induced by coregulation of genes within common pathways, so that models based on distinct although overlapping sets of predictive genes may well reflect a single or small number of relevant biological pathways rather than distinct explanations.

The overall practical relevance of the set of regressions identified (as opposed to nominal statistical significance of any one model) is evaluated by cross-validation prediction. That is, the analysis is repeatedly done in a leave-one-out context, with the tumor left out then being predicted based on the set of models defined and weight by the analysis of the remaining $n-1$ samples, as is (or should be) standard predictive evaluation in problems where predictive value is of primary interest (19–21). Predictions are based on standard weighted model averaging: models identified are evaluated according to their relative data-based probabilities of model fit, and these probabilities provide weights to use in averaging predictions for the hold-out (or future) tumor samples.

Further statistical analysis of the gene expression data aimed to explore a number of genes implicated in the survival regressions to identify additional, statistically associated genes that would then be candidates for potential biological interpretation. A gene showing up as a marker of survival may be a statistical surrogate of other, potentially mechanistic genes. This component of the statistical analysis applied the regression model search repeatedly; now, rather than treating logged survival times as the variable to predict, we used expression of each of a selected small set of genes as the outcome variable. Genes selected as responses for this analysis are the three key genes already discussed, including each of the two probe sets representing DCX, and an additional gene, KIAA0831. These genes represent the four (really five, with the two versions of DCX) most highly scored genes, in terms of posterior probability of appearing in regression models for survival of the full set of over 8,000 genes. Exploring regression models separately for each of these genes as response generates, in each case, a set of models and ranks the genes appearing as predictors in those models according to posterior probabilities, just as in regressions for survival. The four most highly scoring genes in each case are identified in Supplementary Table 1 along with the primary genes already mentioned. Note that, for doublecortex, where two probe sets appeared as predictors of survival, each probe set was considered separately as a response, but both probe sets were removed from the set of predictors for these procedures.

Table 2. Hazard ratios associated with patient and tumor characteristics

	Genetic Analysis		Expression Analysis		Clinical Analysis	
	HR	*P*	HR	*P*	HR	*P*
Age	1.09	0.11	1.00	0.94	1.01	0.80
Race (White)	3.86	0.14	0.78	0.73	1.28	0.73
External beam radiation	0.22	0.085	0.43	0.21	0.15	0.0067
Chemotherapy	2.06	0.48	1.33	0.69	0.87	0.84
EGFR amplification	1.01	0.59	NI	NI	NI	NI
EGFRvIII	1.63	0.54	NI	NI	NI	NI
TP53 mutation	2.05	0.28	NI	NI	NI	NI
PTEN mutation	1.16	0.79	NI	NI	NI	NI
p16INK4A deletion	0.58	0.39	NI	NI	NI	NI
SPARC	NI	NI	9.51	0.00000034	NI	NI
Semaphorin3B	NI	NI	5.69	0.000010	NI	NI
Doublecortex	NI	NI	2.40	0.000019	NI	NI
KIAA0831	NI	NI	0.42	0.075	NI	NI

NOTE: Listed are *P* values and hazards ratio coefficients for three multivariate Cox proportional hazards analyses. The first is a multivariate analysis of survival time (in months) given the genetic alteration summaries corrected for the clinical variables age, race, external beam radiation, and chemotherapy. The second is of the expression data corrected for the clinical measures, and the third is of the clinical measures alone.
Abbreviations: HR, hazards ratio; NI, variable not included.

Results

Molecular p53, PTEN, p16^{INK4A}, and epidermal growth factor receptor status do not associate with patient outcome. Despite the relative uniformity of these patients for the most critical determinants of patient outcome—patient age and tumor grade—examination of the overall survival times for our patients can be stratified. These findings suggest that important determinants beyond the usual selection criteria may influence patient outcome. Previously examined molecular prognostic indicators for gliomas have included p53, PTEN, p16^{INK4A}, and EGFR. However, these markers have not been validated as independent prognostic markers as they frequently co-segregate with tumor grade or patient age. Patient tumor genomic DNA was examined for the presence of p53 and PTEN mutations, amplification of wild-type EGFR, or homozygous deletion of p16^{INK4A} or a constitutively active mutant EGFR (EGFRvIII; Tables 1 and 2). Prior patient analyses have suggested that p53 mutations are associated with gliomas presenting in younger patients and those presenting at lower tumor grades (22). Surprisingly, we found p53 mutations in almost one-third of our patients (11 of 41 = 27%). Whereas the mean survival of patients whose tumors have PTEN mutations or p16^{INK4A} deletions was slightly lower than patients with normal PTEN and p16^{INK4A} (Fig. 1), none of these directed molecular analyses yielded information that was associated with prognostic significance either singly (Table 1) or in multivariate analysis adjusted for patient variables (Table 2). As these genetic changes have frequently been associated with tumor formation or progression, it is likely that these genes may be more closely associated with tumor initiation or progression rather than modifying malignancy among glioblastomas.

Gene expression profiles associated with survival. Statistical analysis evaluated linear regression models treating logged survival times as response and logged gene expression values of multiple genes as candidate predictors, as detailed in ref. (23) and in the Statistical analysis section above. A large number of regression models involving between two and five genes were identified, weighted, and aggregated in cross-validation studies to assess strength and relevance of the association with survival. The key regression models—key in terms of receiving highest posterior probability when assessed across a large number of candidate models—involve subsets of three genes (Fig. 2*A* and Table 2). The regression model analysis was assessed by leave-one-out cross-validation, as described in Materials and Methods. Figure 2*B* provides an overall summary of this assessment and speaks to the explanatory capacity of the set of weighted regressions generated. Figure 2*B* shows the aggregate linear predictions for logged survival for each of the $n = 41$ samples, based on 41 separated re-analyses leaving each sample out of the training data (the remaining 40 samples) and then predicting the left-out case. The point predictions are accompanied by approximate 95% intervals. Note the concordance of the data with predictions, but also that there is wide uncertainty associated with predictions; this is due to the combination of a relatively small sample size and the model uncertainty arising from the combination of multiple regressions. This latter uncertainty is very important in reflecting inherent uncertainty about model form, and to ignore—by choosing, for example, a single "best" regression model—would lead to misleadingly precise intervals.

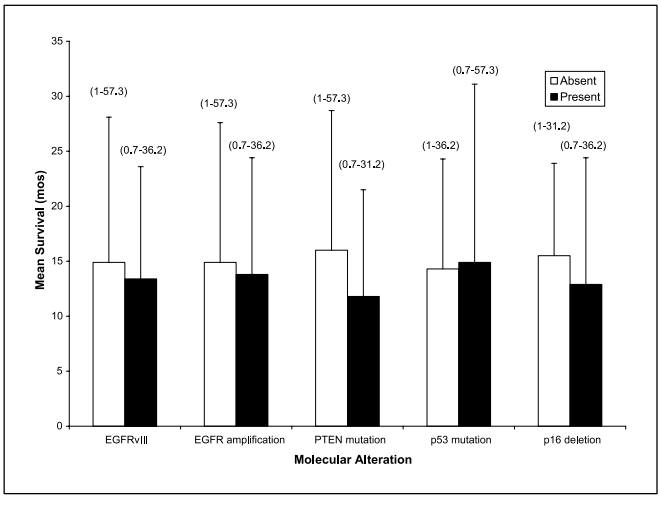

Figure 1. Median survival for patients associated with specific genetic alterations. Tumor specimens were characterized in terms of genetic changes frequently associated with gliomas as in Table 2. Mean survival with SD is displayed based on the molecular status of the tumor. Ranges of survival are shown in parentheses.

Figure 2*C* provides an indication of stratification of patients according to survival risk. The three cases identified in Fig. 2*B* represent individuals with relatively poor, moderate and higher risk in terms of the gene expression markers. Taking the gene expression data for each of these three cases, the model produces predicted log survival times that, when converted to the time axis, correspond to the three survival curves in Fig. 2*C*. The caution is that, whereas the predicted survival curves certainly do represent the differential survival outcomes related to these three regimes of gene expression, this figure does not reflect the associated uncertainty that is relevant for any specific future patient—uncertainty related to that displayed in Fig. 2*B* on the log scale.

Expression levels of Osteonectin, Doublecortex, and Semaphorin3B together associate with patient survival. Dominant regression models involve probe sets for Osteonectin, Doublecortex, and Semaphorin3B. The overall most likely model is in fact the regression on these three genes, and other models with appreciable posterior probability involve subsets of two of these three together with one other gene. Together, these three genes provide explanatory markers of survival (Table 2). Poorer survival is associated with higher levels of each of these three genes; none of them serves as a useful predictive marker alone, but the concordance of higher values together seems to associate with poorer survival (Table 3). Of note, the expression levels of individual genes were not highly correlated with one another, except for very high correlation between the two Doublecortex isoforms (Supplementary Table 2). One informative plot that summarizes the roles of these three genes as markers of survival is given in Fig. 2*A*. The metagene plotted is simply the dominant singular factor (principal component) of the expression levels of these three genes across samples, and is plotted here in a three-dimensional scatter plot together with the expression levels of two of the three—SPARC and Doublecortex (see ref. 16 for discussion of the use of metagenes defined as singular factors from groups of statistically associated genes in related contexts). The points are color-coded according to the predicted mean of log survival corresponding to the expression levels, running from blue (lower risk) to red (higher risk).

Validation of gene expression of Osteonectin and Doublecortex. To confirm the expression relationships derived from analyses of Affymetrix gene chip hybridization studies, RT-PCR confirmation of expression of these two genes was done in a subset

of our patient specimens (20 tumors). The levels of Osteonectin and Doublecortex message measured by RT-PCR were generally well correlated with the levels detected in the Affymetrix chip studies (R^2 = 0.7-0.8).

Expression profiles can derive additional relationships between genes expressed in patient specimens related to survival. Additional statistical analysis explored statistical associations in gene expression data among a few of the key genes implicated in the survival regressions and other genes that, in a regression context, showed up as predictive of expression fluctuations of this initial set of genes (see Materials and Methods). Figure 3 displays a graph summarizing the predictive relationships identified in this analysis, presented as a statistical graphical association model—a subgraph of the much larger graph relating expression levels across all genes (23, 24). The set of genes here are listed in Supplementary Table 1. Arrows are directed from a gene A to a gene B to represent the appearance of gene A as a predictor of gene B in one of the three most highly weighted regressions for expression of gene B. A dashed edge indicates that gene A had a negative regression coefficient in the highest probability model in which it was involved in predicting gene B. The number labeling an edge from gene A to gene B indicates the aggregate posterior probability of all regression models for gene B that contain gene A as an explanatory variable—an overall measure of the relevance/weight of gene A as predictive of gene B.

Discussion

Glioblastomas are genetically heterogeneous, suggesting that a diverse set of gene products may act to regulate the behavior, and thus outcome, from these tumors. Despite these limitations, we have been able to derive relationships between the expression of three genes and patient survival. The three genes that are the dominant contributors to models associating gene expression profiles with patient survival—Osteonectin, Doublecortex, and Semaphorin3B—share roles in the regulation of cellular motility suggesting that potential regulators of tumor invasion may play a part in determining patient survival after tumor progression to a glioblastoma. Unlike most other types of cancer, the morbidity and mortality from most brain tumors comes not from metastases but rather local invasion of the tumor preventing complete surgical resection (3). The majority of high-grade gliomas (80-90%) recur <2 cm of the original tumor site (25), but even local control will eventually fail due to the invasive nature of gliomas because glioma cells frequently extend through much of the neural axis prior to diagnosis. Many patients die due to malignant gliomas without a significant mass present (26). Infiltrative glioma cells are a particular therapeutic challenge due to their diffuse localization, distance from the initial site of resection, protection by an intact blood-brain barrier, and low frequency of mitosis (3). Whereas there has been a dramatic increase in the understanding of the mechanisms by which cancers initiate and grow, the process of tumor invasion remains poorly understood.

Of the genes detected in our expression studies, Osteonectin has been most clearly linked to glioma pathophysiology in prior studies. Osteonectin, also known as secreted protein acidic and rich in cysteine (SPARC) or BM-40, is an extracellular protein that plays an important role in development, tissue healing and remodeling, and

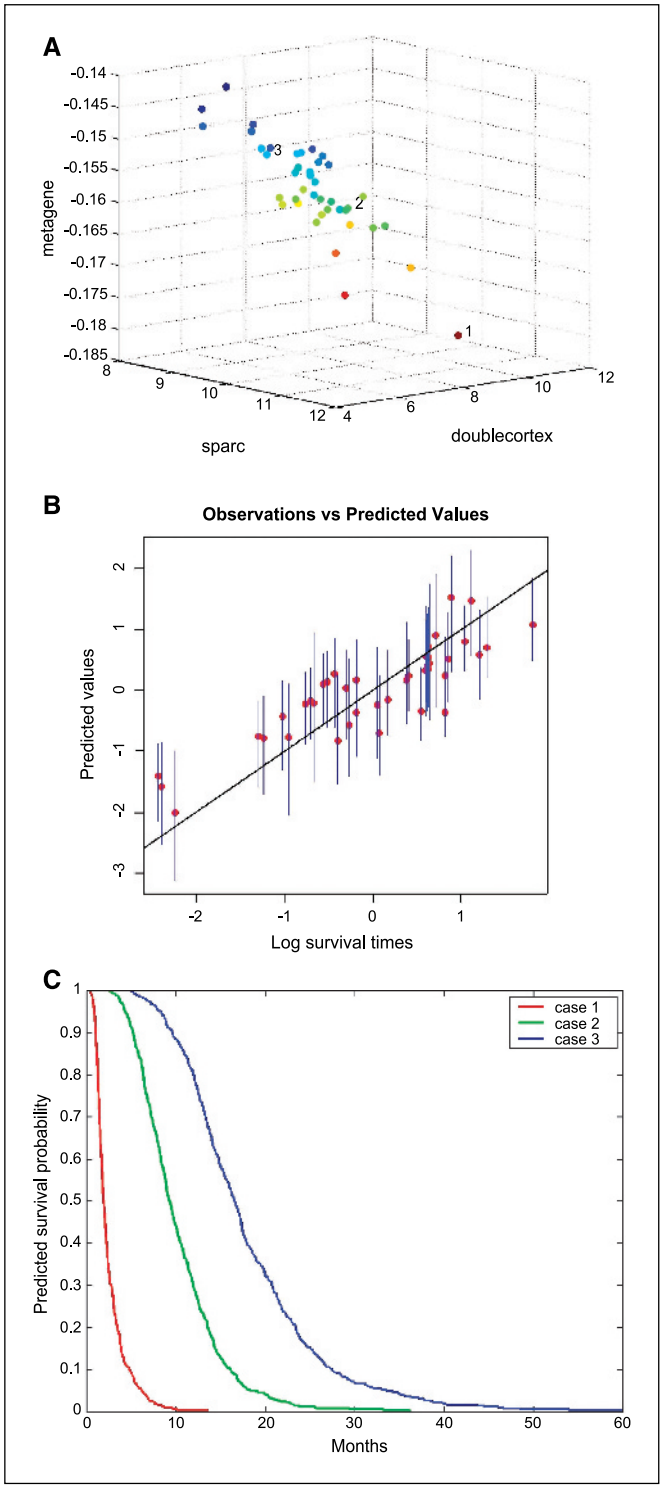

Figure 2. Expression analysis of genes related to patient survival. *A,* scatterplot of 41 glioblastoma cases according to expression levels on two of the three key genes underlying regression models evaluated. The metagene (first principal component of expression values for the three key genes SPARC, Doublecortex, and Semaphorin3B) which dominate in survival predictions is also included. The color bar/coding indicates survival time. The three samples numbered represent cases with poor (*red*), moderate (*green*), and better (*blue*) survival risks. *B,* leave-one-out cross-validation predictions from the aggregate regression model for log (base 2) survival times. For each patient, the predicted mean log survival time is plotted, with associated 95% interval, against the observed log survival time (*horizontal axis*). *C,* predicted survival functions for three hypothetical populations of individuals whose values of gene expression on the three key genes—SPARC, Doublecortex and Semaphorin3B—are the same as those of the three real patients marked in (*B*).

Table 3. Combined impact of elevated expression of SPARC, Doublecortex, and Semaphorin3B on survival

Number of genes highly expressed	Median survival	Number of patients
0	41.9	2
1	19.2	22
2	9.5	13
3	7.6	4

NOTE: A crude combined measure of expression of SPARC, Doublecortex, and Semaphorin3B was defined in terms of high (>median in the sample) versus low (≤median in the sample) expression of each of the three genes on gene chip studies. The combined variable counts the number of genes expressed at the high level. The table summarizes survival in months as a function of this variable.

angiogenesis (reviewed in ref. 27). Osteonectin/SPARC was originally discovered as an important component of bone (28) but is also expressed in epithelia exhibiting high rates of turnover (gut, skin, and glandular tissue), as well as vascular smooth muscle cells and endothelial cells (29). In addition to its normal physiologic role, Osteonectin/SPARC is abnormally expressed in cancers. Many cancers, including cancers of the gastrointestinal tract, breast, lung, kidney, adrenal cortex, prostate, bladder, and meninges (29–33), express increased SPARC levels that are associated with a conversion to invasive and metastatic tumors. Gene expression analysis of potential tumor markers in malignant gliomas by sequential analysis of gene expression found a 10-fold overexpression of SPARC (34). Immunohistochemical analysis of human gliomas reveals that SPARC is expressed specifically at sites of tumor invasion—in tumor cells at the tumor-brain interface, endothelial cells of tumor-associated vessels, and reactive astrocytes (30). *In vitro* assays of gliomas have further implicated SPARC expression in increased glioma invasion (35) and angiogenesis (36). We and others have shown that forced expression of SPARC in human glioma cell lines promotes tumor cell invasion in both cell culture and animal models associated with matrix metalloproteinase expression (37, 38). More recently, we have shown that SPARC expression in human glioma cell lines induces increased activation of AKT and promotes cell survival with serum survival (39). Thus, the prominent role of SPARC in influencing patient survival suggests that regulation of the cellular microenvironment may significantly contribute to tumor behavior beyond the progression to high-grade malignancy.

Doublecortin and doublecortex are gene products encoded by the locus linked to X-linked lissencephaly (40). Whereas males with X-linked lissencephaly display abnormally smooth brains, female patients develop brains with abnormal heterotopic gray matter regions reminiscent of a second cortical region. During brain development, neuronal precursors are generated at periventricular regions then migrate outwards to populate the cortical layers. Heterotopic gray matter regions have been linked to impaired migration of neuronal cell bodies. Doublecortin encodes a microtubule-associated protein with two actin-binding domains that regulate neuronal migration (41–43). The activity of doublecortin is suppressed by phosphorylation at specific residues (42, 43). The kinases regulating these phosphorylation events include Cdk5, MARKS, and protein kinase A (42, 43). These kinases may be abnormally expressed or regulated in some cancers, including gliomas. The expression levels and contributions of doublecortin in cancer have been previously unrecognized. Likewise, the regulation of doublecortin expression is poorly understood. A recent report linked targeted disruption of PTEN expression to changes in doublecortin expression (44). The striking relationship that we have detected between the expression levels of SOX4 and doublecortin strongly suggest co-regulation of these genes. SOX4 is a SRY box containing transcription factor that is expressed during brain development in the cerebellar external granule layer (45). The functions of SOX4 have been dissected through targeted disruption in mice, resulting in early vascular death with defects in cardiac outflow tract formation and pro-B lymphocyte generation (46). SOX4 has not been widely studied in cancer, but SOX4 has been shown to be overexpressed in another central nervous system cancer, medulloblastoma (47). The link between SOX4 expression and that of both doublecortin isoforms strongly suggests that SOX4 may regulate doublecortin expression. In a strong validation of the potential biological relationships that may be derived by the statistical analyses used in our studies, we have detected a similar relationship between SOX11 and one of the expressed transcripts of doublecortin. SOX11 is the SOX family member most closely related to SOX4 with potential overlapping functions.

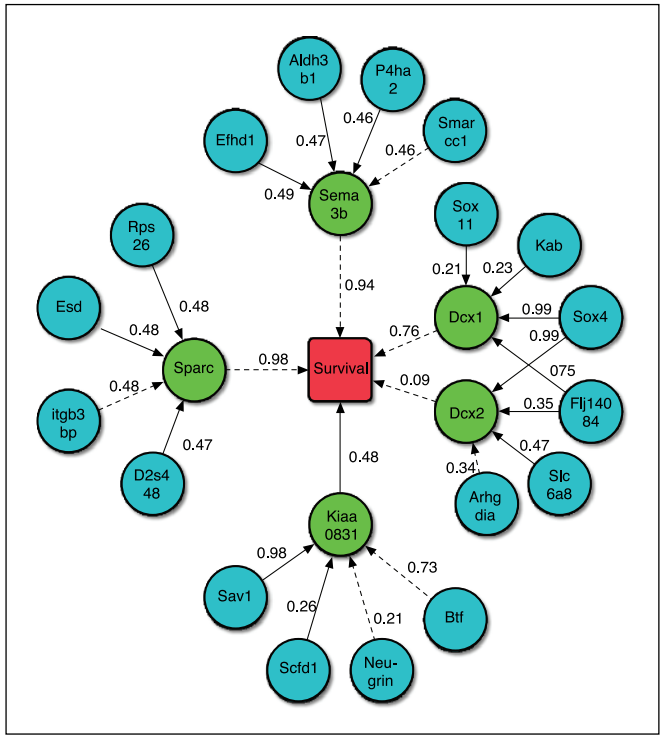

Figure 3. A graph representing a small component of the large-scale gene expression–based graphical association network of >8,000 genes. This graph displays genes that are implicated in regression models predicting survival, and additional genes that arise in regression models predicting gene expression fluctuations of that initial group. An arrow from gene A to gene B indicates that gene A appears in one of the top three most highly weighted regression models predicting gene expression fluctuations of gene B; the arrow is dashed (full) according to whether the estimation regression coefficient of gene A in the most highly weighted of those regressions is negative (positive). The number on an arrow from gene A to gene B is the estimated posterior probability of all regression models for gene B that contain gene A as an explanatory variable. See discussion in text (Statistical analysis in Materials and Methods) and Supplementary Table 1 for details on all genes.

Semaphorin3B (SEMA3B) is a class III, secreted semaphorin with SEMA, immunoglobulin, and short basic domains. In parallel to other semaphorins, SEMA3B regulates neuronal migration. SEMA3B antagonizes SEMA3A neuronal growth cone repulsion at neuropilin-1 homodimers but acts as an agonist at neuropilin-1/2 heterodimers or neuropilin-2 homodimers (48). The neuropilins are transmembrane receptors without clear independent signaling functions that may act as accessory receptors for vascular endothelial growth factor (VEGF). Although VEGF has been most closely linked to endothelial cell proliferation and increased vascular permeability, evidence of the role of VEGF in cellular migration and brain development are apparent. The SEMA3B locus is located at 3p21.3, which is a homozygous deletion region in small cell lung cancer, suggesting that SEMA3B may act as a tumor suppressor gene in some cancers (49). Reintroduction of p53 into the p53-null U373MG human malignant glioma cell line induced SEMA3B expression (50). The dichotomous role of SEMA3B parallels that of SPARC, which can also restrict tumor cell proliferation and exhibit tumor suppressive roles in cancers as well. The putative SEMA3B receptors, the neuropilins, are expressed in human gliomas and may serve biological roles in tumor malignancy.

In summary, this gene expression study provides evidence that three genes which regulate cellular motility may contribute to the poor prognosis of patients with glioblastomas. No previous studies, of which we are aware of, have elucidated the conclusive links between expression of specific genes and survival of older glioblastoma patients. Although cellular mitogenesis and resistance to apoptosis have been the targets of many biological therapies, our regression analyses using gene expression to explain the survival outcomes revealed that genes whose primary cellular effects may be the regulation of cellular migration appear as candidate markers of poor survival. Together, these results suggest that tumor migration may represent an important effector of glioblastoma malignancy and may warrant accelerated development of specific therapies. Current targeted therapies for glioblastomas have focused on cellular pathways that primarily regulate proliferation and apoptosis. Clinical experience suggests that tumor invasion is a severe challenge in the management of glioblastoma patients. Elegant studies by Berens, Bjerkvig, Rao and others have shown that glioma invasion can be the target of directed therapies and that these approaches may augment the efficacy of traditional therapies (reviewed in ref. 3). Our studies may lend further weight to these approaches and suggest that the gene products whose expression is now linked to poor survival may be useful therapeutic targets. Future studies will prospectively determine the link between the expression of SPARC, Doublecortex, and SEMA3B in gliomas of all grades and patient outcome. Additional studies under way will further dissect the contributions of these gene products to the biology of gliomas, including tumor cell invasion, proliferation, apoptosis, and secretion of angiogenic factors.

Acknowledgments

Received 11/2/2004; revised 2/4/2005; accepted 3/9/2005.

Grant support: W.M. Keck Foundation. J.N. Rich is a Damon Runyon-Lilly Clinical Investigator and a Sidney Kimmel Cancer Foundation Scholar. This work was also supported by NIH grant NS047409 (J.N. Rich).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

References

1. Legler JM, Ries LA, Smith MA, et al. Brain and other central nervous system cancers: recent trends in incidence and mortality. J Natl Cancer Inst 1999;91: 1382–90.
2. Jemal A, Murray T, Ward E, et al. Cancer statistics, 2005. CA Cancer J Clin 2005;55:10–30.
3. Giese A, Bjerkvig R, Berens ME, Westphal M. Cost of migration: invasion of malignant gliomas and implications for treatment. J Clin Oncol 2003;21:1624–36.
4. Kleihues P, Cavenee WK, editors. Pathology and genetics of tumours of the nervous system. Lyon: IARC Press; 2000. p. 3–70.
5. Lamborn KR, Chang SM, Prados MD. Prognostic factors for survival of patients with glioblastoma: recursive partitioning analysis. Neuro-oncol 2004;6: 227–35.
6. Sawaya R. Extent of resection in malignant gliomas: a critical summary. J Neurooncol 1999;42:303–5.
7. Stewart LA. Chemotherapy in adult high-grade glioma: a systematic review and meta-analysis of individual patient data from 12 randomised trials. Lancet 2002;359: 1011–8.
8. Rich JN, Bigner DD. Development of novel targeted therapies in the treatment of malignant glioma. Nat Rev Drug Discov 2004;3:430–46.
9. Kleihues P, Ohgaki H. Primary and secondary glioblastomas: from concept to clinical diagnosis. Neuro-oncol 1999;1:44–51.
10. Smith JS, Tachibana I, Passe SM, et al. PTEN mutation, EGFR amplification, and outcome in patients with anaplastic astrocytoma and glioblastoma multiforme. J Natl Cancer Inst 2001;93:1246–56.
11. Nutt CL, Mani DR, Betensky RA, et al. Gene expression-based classification of malignant gliomas correlates better with survival than histological classification. Cancer Res 2003;63:1602–7.
12. 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 2003;63:6613–25.
13. Freije WA, Castro-Vargas FE, Fang Z, et al. Gene expression profiling of gliomas strongly predicts survival. Cancer Res 2004;64:6503–10.
14. Mischel PS, Cloughesy TF, Nelson SF. DNA-microarray analysis of brain cancer: molecular classification for therapy. Nat Rev Neurosci 2004;5:782–92.
15. Reardon DA, Akabani G, Coleman RE, et al. Phase II trial of murine (131)I-labeled antitenascin monoclonal antibody 81C6 administered into surgically created resection cavities of patients with newly diagnosed malignant gliomas. J Clin Oncol 2002;20: 1389–97.
16. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 2003;31:e15.
17. Bioconductor: Open source software for bioinformatics. http://www.bioconductor.org.
18. Hans C, Dobra A, West M. Shotgun stochastic search in regression, in Discussion Paper, ISDS, Durham: Duke University; 2004.
19. Pittman J, Huang E, Dressman H, et al. Integrated modeling of clinical and gene expression information for personalized prediction of disease outcomes. Proc Natl Acad Sci U S A 2004;101:8431–6.
20. Huang E, Cheng SH, Dressman H, et al. Gene expression predictors of breast cancer outcomes. Lancet 2003; 361:1590–6.
21. Pittman J, Huang E, Nevins J, Wang Q, West M. Bayesian analysis of binary prediction tree models for retrospectively sampled outcomes. Biostatistics 2004;5: 587–601.
22. Nozaki M, Tada M, Kobayashi H, et al. Roles of the functional loss of p53 and other genes in astrocytoma tumorigenesis and progression. Neuro-oncol 1999;1: 124–37.
23. Dobra A, Jones B, Hans C, Nevins JR, West M. Sparse graphical models for exploring gene expression data. J Multivariate Analysis 2004;90:196–212.
24. Jones B, Dobra A, Carvalho C, Hans C, Carter C, West M. Experiments in stochastic computation for high-dimensional graphical models. Statistical Science. In press 2005.
25. Roth JG, Elvidge AR. Glioblastoma multiforme: a clinical study. J Neurosurg 1962;17:736–50.
26. Silbergeld DL, Rostomily RC, Alvord EC Jr. The cause of death in patients with glioblastoma is multifactorial: clinical factors and autopsy findings in 117 cases of supratentorial glioblastoma in adults. J Neurooncol 1991;10:179–85.
27. Bradshaw AD, Sage EH. SPARC, a matricellular protein that functions in cellular differentiation and tissue response to injury. J Clin Invest 2001;107: 1049–54.
28. Termine JD, Kleinman HK, Whitson SW, Conn KM, McGarvey ML, Martin GR. Osteonectin, a bone-specific protein linking mineral to collagen. Cell 1981; 26:99–105.
29. Porter PL, Sage EH, Lane TF, Funk SE, Gown AM. Distribution of SPARC in normal and neoplastic human tissue. J Histochem Cytochem 1995;43:791–800.
30. Rempel SA, Golembieski WA, Ge S, et al. SPARC: a signal of astrocytic neoplastic transformation and reactive response in human primary and xenograft gliomas. J Neuropathol Exp Neurol 1998;57:1112–21.
31. Rempel SA, Ge S, Gutierrez JA. SPARC: a potential diagnostic marker of invasive meningiomas. Clin Cancer Res 1999;5:237–41.
32. Thomas R, True LD, Bassuk JA, Lange PH, Vessella RL. Differential expression of osteonectin/SPARC during human prostate cancer progression. Clin Cancer Res 2000;6:1140–9.

33. Bellahcene A, Castronovo V. Increased expression of osteonectin and osteopontin, two bone matrix proteins, in human breast cancer. Am J Pathol 1995;146: 95–100.

34. Lal A, Lash AE, Altschul SF, et al. A public database for gene expression in human cancers. Cancer Res 1999;59: 5403–7.

35. Golembieski WA, Ge S, Nelson K, Mikkelsen T, Rempel SA. Increased SPARC expression promotes U87 glioblastoma invasion *in vitro*. Int J Dev Neurosci 1999;17:463–72.

36. Vajkoczy P, Menger MD, Goldbrunner R, et al. Targeting angiogenesis inhibits tumor infiltration and expression of the pro-invasive protein SPARC. Int J Cancer 2000;87:261–8.

37. Rich JN, Shi Q, Hjelmeland M, et al. Bone-related genes expressed in advanced malignancies induce invasion and metastasis in a genetically defined human cancer model. J Biol Chem 2003;278: 15951–7.

38. Schultz C, Lemke N, Ge S, Golembieski WA, Rempel SA. Secreted protein acidic and rich in cysteine promotes glioma invasion and delays tumor growth *in vivo*. Cancer Res 2002;62:6270–7.

39. Shi Q, Bao S, Maxwell JA, et al. Secreted protein acidic, rich in cysteine (SPARC) mediates cellular survival of gliomas through AKT activation. J Biol Chem 2004;279:52200–9.

40. Gleeson JG, Allen KM, Fox JW, et al. Doublecortin, a brain-specific gene mutated in human X-linked lissencephaly and double cortex syndrome, encodes a putative signaling protein. Cell 1998;92:63–72.

41. Francis F, Koulakoff A, Boucher D, et al. Doublecortin is a developmentally regulated, microtubule-associated protein expressed in migrating and differentiating neurons. Neuron 1999;23:247–56.

42. Tanaka T, Serneo FF, Tseng HC, Kulkarni AB, Tsai LH, Gleeson JG. Cdk5 phosphorylation of doublecortin ser297 regulates its effect on neuronal migration. Neuron 2004;41:215–27.

43. Schaar BT, Kinoshita K, McConnell SK. Doublecortin microtubule affinity is regulated by a balance of kinase and phosphatase activity at the leading edge of migrating neurons. Neuron 2004;41:203–13.

44. Li L, He F, Litofsky NS, Recht LD, Ross AH. Profiling of genes expressed by PTEN haploinsufficient neural precursor cells. Mol Cell Neurosci 2003; 24:1051–61.

45. Cheung M, Abu-Elmagd M, Clevers H, Scotting PJ. Roles of Sox4 in central nervous system development. Brain Res Mol Brain Res 2000;79:180–91.

46. Schilham MW, Oosterwegel MA, Moerer P, et al. Defects in cardiac outflow tract formation and pro-B-lymphocyte expansion in mice lacking Sox-4. Nature 1996;380:711–4.

47. Lee CJ, Appleby VJ, Orme AT, Chan WI, Scotting PJ. Differential expression of SOX4 and SOX11 in medulloblastoma. J Neurooncol 2002;57:201–14.

48. Takahashi T, Nakamura F, Jin Z, Kalb RG, Strittmatter SM. Semaphorins A and E act as antagonists of neuropilin-1 and agonists of neuropilin-2 receptors. Nat Neurosci 1998;1:487–93.

49. Lerman MI, Minna JD. The 630-kb lung cancer homozygous deletion region on human chromosome 3p21.3: identification and evaluation of the resident candidate tumor suppressor genes. The International Lung Cancer Chromosome 3p21.3 Tumor Suppressor Gene Consortium. Cancer Res 2000;60: 6116–33.

50. Ochi K, Mori T, Toyama Y, Nakamura Y, Arakawa H. Identification of semaphorin3B as a direct target of p53. Neoplasia 2002;4:82–7.