Oncogene (2003) 22, 4918–4923

www.nature.com/onc

Gene expression profiling identifies molecular subtypes of gliomas

Ruty Shai[1], Tao Shi[1,2], Thomas J Kremen[3], Steve Horvath[1,2], Linda M Liau[3], Timothy F Cloughesy[4], Paul S Mischel*[,5,6] and Stanley F Nelson[1,6]

[1]Department of Human Genetics, Henry E. Singleton Brain Tumor Program, UCLA School of Medicine, Los Angeles, CA 90095, USA; [2]Department of Biostatistics, Henry E. Singleton Brain Tumor Program, UCLA School of Medicine, Los Angeles, CA 90095, USA; [3]Department of Neurosurgery, Henry E. Singleton Brain Tumor Program, UCLA School of Medicine, Los Angeles, CA 90095, USA; [4]Department of Neurology, Henry E. Singleton Brain Tumor Program, UCLA School of Medicine, Los Angeles, CA 90095, USA; [5]Department of Pathology and Laboratory Medicine, Henry E. Singleton Brain Tumor Program, UCLA School of Medicine, Los Angeles, CA 90095, USA

Identification of distinct molecular subtypes is a critical challenge for cancer biology. In this study, we used Affymetrix high-density oligonucleotide arrays to identify the global gene expression signatures associated with gliomas of different types and grades. Here, we show that the global transcriptional profiles of gliomas of different types and grades are distinct from each other and from the normal brain. To determine whether our data could be used to uncover molecular subtypes without prior knowledge of pathologic type and grade, we performed K-means clustering analysis and found evidence for three clusters with the aid of multidimensional scaling plots. These clusters corresponded to glioblastomas, lower grade astrocytomas and oligodendrogliomas ($P<0.00001$). A predictor constructed from the 170 genes that are most differentially expressed between the subsets correctly identified the type and grade of all samples, indicating that a relatively small number of genes can be used to distinguish between these molecular subtypes. These results further define molecular subsets of gliomas which may potentially be used for patient stratification, and suggest potential targets for treatment.
Oncogene (2003) **22,** 4918–4923. doi:10.1038/sj.onc.1206753

Keywords: glioblastoma; gene expression profiling; microarray; glioma

Introduction

The development of inhibitors that are targeted to specific genetic lesions or pathway alterations represents an important new approach to cancer therapy. The success of this approach depends largely on identifying the right subset of patients for each type of treatment. Therefore, one of the critical challenges in cancer biology is to develop classifications of tumors that reflect the underlying molecular abnormalities and which can be used as the basis for patient stratification rather than relying solely on histologic classification of tumors.

Standard approaches to cancer diagnosis and therapy have relied primarily on pathologic/morphological assessment. In some types of cancer, this approach has provided useful prognostic and therapeutic information. Unfortunately, this approach has been only partially successful for gliomas, the most common malignant brain tumor of adults (Mischel and Cloughesy, 2003). Current pathologic categories of glioma provide relatively good prognostic information about overall survival, but have generally not been useful for determining optimum therapy, or likely response to treatment. Further, the current classification system does not take into consideration the underlying molecular lesions. The availability of large-scale genomic approaches and new bioinformatics analysis methods now make it possible to develop molecularly defined classifications of tumors. Large-scale gene expression profiling can be used to identify tumor subtypes with distinct molecular and/or clinical phenotypes or responses to therapy (Golub *et al*., 1999; Perou *et al*., 2000; Alizadeh *et al*., 2001; MacDonald *et al*., 2001; Sorlie *et al*., 2001; Pomeroy *et al*., 2002; Shipp *et al*., 2002), including in gliomas (Sallinen *et al*., 2000; Ljubimova *et al*., 2001; Rickman *et al*., 2001; Lal *et al*., 2002; Zhang *et al*., 2002; Mischel *et al*., 2003).

We analysed the expression of 12 555 probe sets encoding ~10 000 genes (Affymetrix U95Av2 oligonucleotide arrays) in 35 glioma samples of varying pathologic type and grade (grade II astrocytoma, grade III astrocytomas, grade IV astrocytomas (glioblastoma), grade II oligodendrogliomas), as well as seven normal brain samples taken from the subcortical white matter. Each tumor was examined by a neuropathologist and dissected into two portions; one portion was used for tissue diagnosis and the other for RNA extraction. The diagnosis on each tumor was confirmed by two neuropathologists according to the WHO classification system (Kleihues *et al*., 2002). To confirm that the tissue

*Correspondence: PS Mischel; E-mail: pmischel@mednet.ucla.edu
[6]These authors codirected this work
Received 26 December 2002; revised 25 April 2003; accepted 10 May 2003

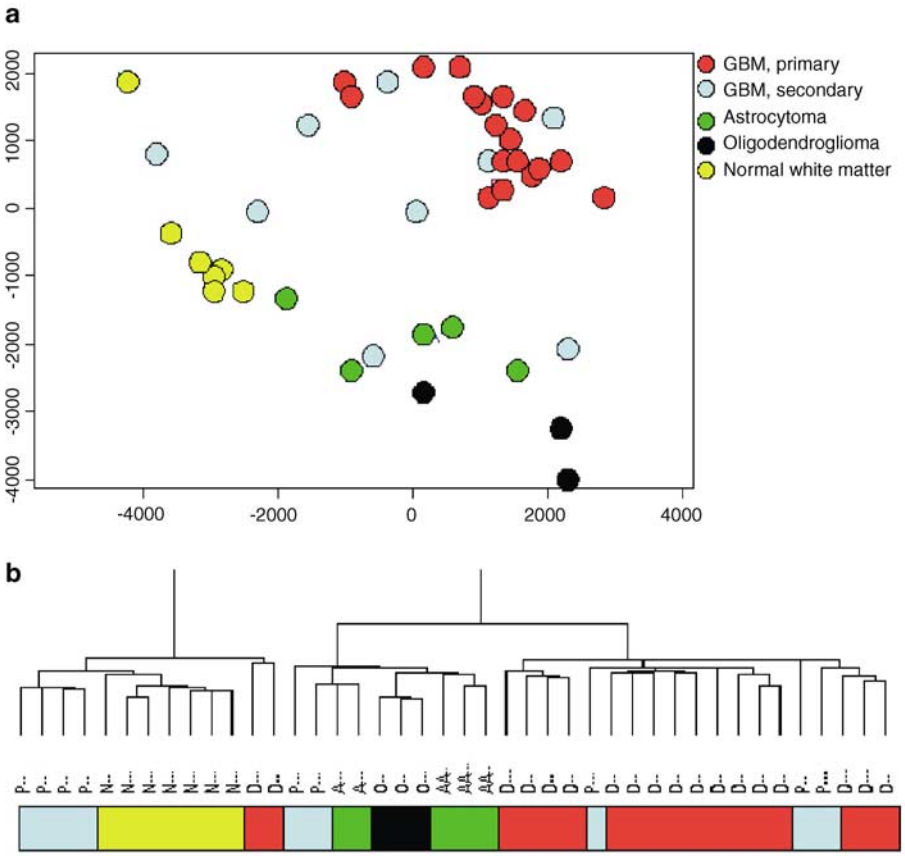

Figure 1 Uninstructed grouping of glial tumor samples. RNA was extracted and biotin-labeled cDNA was generated as previously described (Mischel *et al*., 2003). Labeled cRNA was fragmented, hybridized to Affymetrix U95Av2 GeneChip and scanned to generate an image file (Mischel *et al*., 2003). Model-based expression indices were calculated (Li and Wong, 2001; Mischel *et al*., 2003) and multidimensional scaling and hierarchical clustering were performed (Kaufmann and Rousseeu, 1990; Venables and Ripley, 1999; Mischel *et al*., 2003). (**a**) Multidimensional scaling plot of all 42 tissue samples plotted in two-dimensional space using expression values from all 12 555 probesets. (**b**) The same 42 tissue samples were grouped into hierarchical clusters. Tissue samples are color-coded. Red: primary GBM, blue: secondary GBM, green: grades II and III astrocytomas, black: oligodendroglioma, yellow: normal white matter

from which RNA was extracted was reflective of the pathologic diagnosis, a hematoxylin and eosin-stained section from the frozen tumor piece was analysed. RNA extraction, generation of labeled cRNA and hybridization to Affymetrix U95Av2 oligonucleotide arrays were performed as previously described (Mischel *et al*., 2003). We computed model-based expression indices using the dCHIP software (Li and Wong, 2001). As a first step in the analysis, we asked whether the global transcriptional signatures of the different pathologic subtypes of gliomas were molecularly distinct. We performed multidimensional scaling, an unsupervised method of data reduction, in which high-dimensional gene expression data are projected onto two viewable dimensions representing linear combinations of genes that provide the most variation in the data set (Venables and Ripley, 1999). In this approach, the distance between data points provides a measure of their differences. Multidimensional scaling analysis of our samples based on expression of all 12 555 probe sets demonstrated that the global gene expression profiles of gliomas of different type and grade have distinctive global gene expression signatures. The glioblastomas, lower grade astrocytomas and oligodendrogliomas were all separable from each other, and from normal brain tissue (Figure 1a). The multidimensional scaling data also indicate that primary glioblastomas, which arise as *de novo* grade IV tumors, are not molecularly distinct from secondary glioblastomas, which develop from lower grade gliomas. However, the secondary GBMs are more diverse than the primary GBMs.

We further analysed the global gene expression signatures by performing hierarchical clustering, another unsupervised learning method (Kaufmann and Rousseeu, 1990; Hastie *et al*., 2001a; Hastie *et al*., 2001b) (Figure 1b). The main branch of the dendrogram identified one cluster enriched for normal brain tissue and the other from glioma samples ($P = 0.00006$, Fisher's exact test). Within the glioma-enriched branch of the dendrogram, there were two main branches, one enriched for lower grade gliomas and the other for glioblastomas ($P = 0.00001$). Further, within the lower grade glioma-enriched branch of the dendrogram, the low-grade astrocytomas (grade II), the anaplastic astrocytomas (grade III) and the oligodendrogliomas were separable ($P = 0.003$). Thus, the global gene expression signatures of these different pathologic subsets are distinct. We have previously shown that

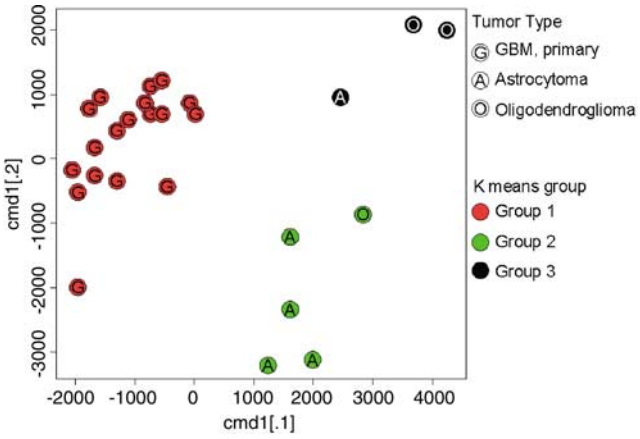

Figure 2 Grouping of tumors. All tumor samples were plotted using multidimensional scaling using all 12 555 probesets. The identity of the histology of the tumor sample is indicated by a letter: G indicates glioblastoma, A indicates astrocytoma, O indicates oligodendroglioma. We performed nonhierarchical K-means clustering (Kaufmann and Rousseeu, 1990). Three groups were defined. Each tumor is assigned to one of three cluster groups by color: red is group 1, green is group 2 and black is group 3

there are distinct molecular subsets of primary glioblastomas associated with distinct global gene expression signatures (Mischel *et al*., 2003). The data presented here suggest that secondary glioblastomas are an even more heterogeneous group than primary glioblastomas that awaits further molecular subclassification. The extent of heterogeneity among secondary glioblastomas is not surprising, considering that these tumors arise from lower grade gliomas by accumulation of additional mutations, typically in the setting of p53 mutation.

Next, we asked if our data might be used to uncover molecular subtypes of gliomas without prior knowledge of their pathologic type or grade. That is, how many categories of glioma are suggested by the gene expression data? When the data were represented in a two-dimensional multidimensional scaling plot based on all 12 555 genes, we found that there is a clear separation of samples into three clusters (Figure 2). We then used K-means clustering (Kaufmann and Rousseeu, 1990) to classify the samples into three clusters based on the Euclidean distances between them. Cluster one was enriched for glioblastomas, cluster two for astrocytomas (grades II and III) and cluster three for oligodendrogliomas ($P<0.0001$) (Figure 2). These data indicate that there are three main molecular subsets of gliomas, which correspond to glioblastomas, astrocytomas (grades II and III) and oligodendrogliomas.

Lastly, we constructed a gene voting predictor for glioma subtype by performing multiple pair-wise comparisons between pathologic groups (normal vs glioma; grade II + III astrocytomas vs glioblastoma; grade II astrocytomas vs grade III astrocytomas; grade II and III astrocytomas vs oligodendrogliomas and normal brain vs oligodendrogliomas) (Golub *et al*., 1999). To add additional prognostic relevance to the predictor, we added a pair-wise comparison between short surviving glioblastomas (less than 1 year) vs longer surviving glioblastomas (greater than 3 years). The genes most differentially expressed between subsets were selected based on a *t*-test of the quantitative gene expression values of the 4000 most variable genes, and the 30 genes most differentially expressed in each pair-wise comparison were selected. To assess the validity of this approach, we estimated crossvalidation error rates (Table 1). Our gene selection procedure, coupled with a weighted gene voting prediction method (Golub *et al*., 1999), accurately distinguished tumors from normals (7% error rate) and lower grade astrocytomas from glioblastomas (4% error rate). The predictor distinguished oligodendrogliomas from astrocytomas and from normal tissue with 100% accuracy (0% error rate). In addition, the predictor distinguished short surviving glioblastoma patients from longer surviving patients (22% error rate). A final gene list was then constructed by pooling the most differentially expressed genes from these individual comparisons, and redundant genes were eliminated (supplementary figure x). We used this final gene list (170 genes – available as supplementary information) to hierarchically cluster the tumor and normal brain samples. This analysis identified four molecular subsets: (1) normal brain samples; (2) grade II and grade III astrocytomas (which were distinguishable from each other as demonstrated by a further branch in the dendrogram); (3) oligodendrogliomas and (4) glioblastomas ($P<0.000001$) (Figure 3). These data indicate that a relatively small number of genes can be used to characterize the key molecular distinctions between gliomas of different pathologic subtype and grade.

Table 1 Leave one out crossvalidation error rates of the gene voting predictor

Comparisons	*Naïve error rate*[a]	*GV error rate*
Normals ($n=7$) vs glial tumors ($n=23$)	0.23	0.07
Astrocytomas ($n=5$) vs primary GBM ($n=18$)	0.22	0.04
Astrocytomas ($n=5$) vs oligodendrogliomas ($n=3$)	0.38	0.00
Normal ($n=7$) vs oligodendrogliomas ($n=3$)	0.3	0.00
Short ($n=10$) vs long survival primary GBM ($n=8$)	0.44	0.22

[a]Error rate that assigns the most frequent class to each observation

These comparative gene expression patterns may provide potentially important information about the underlying biology of these subtypes of gliomas. We found that the transcriptional profile of the astrocytomas (grade II, grade III and primary glioblastomas) was enriched for genes involved in cellular proliferation, RNA processing, signal transduction and proteosomal function. Representative genes are demonstrated in Table 2. As this comparison was based on a relatively large number of samples (five lower grade astrocytomas and 18 primary glioblastomas vs seven normal white mater samples) and because we have previously

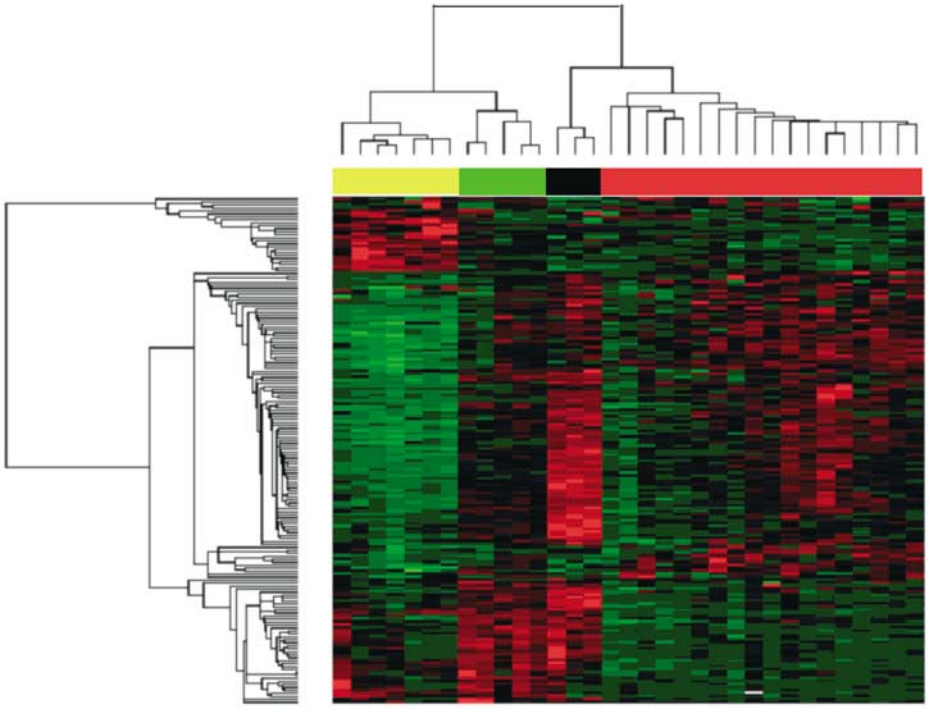

Figure 3 Hierarchical clustering of seven normal white matter tissue samples and 26 glial tumor samples using 170 filtered genes. We used dChip to perform hierarchical clustering of the samples using $1-r$ where r is Pearson's correlation coefficient as the distance measure (Kaufmann and Rousseeu, 1990; Li and Wong, 2001; Mischel *et al*., 2003). Samples are coded by color. Gene expression values are represented as expression relative to the mean of all samples; red is a relatively higher expression and green is a relatively lower expression

Table 2 Representative genes with increased expression in astrocytomas (I) and oligodendrogliomas (II) relative to normal white matter

	Gene symbol	*ProbeSet*	*Fold*	*P-value*	*Function*	*Refs.*
I						
	SAM68	39346_at	2.8	4.63E−07	Signal transduction-dependent regulator of RNA splicing	Coyle *et al*. (2003), Matter *et al*. (2002), Najib and Sanchez-Margalet (2002)
	nmt55/p54nrt	38527_at	3.5	1.90E−07	RNA splicing, cellular proliferation, and carbonic anhydrase activity	Shav-Tal and Zipori (2002)
	RNPS1	36186_at	2.0	1.28E−07	Regulator of RNA splicing	Lykke-Andersen *et al*. (2001)
	NPM	38542_at	2.1	3.38E−07	Cellular proliferation and cell cycle control	Okuda (2002), Okuda *et al*. (2000)
	RAN	38708_at	3.3	7.05E−07	Cellular proliferation	De Luca *et al*. (2003), Yano (2002)
	PSMB7	1313_at	2.7	1.72E−07	Component of 20S proteosome	Nandi *et al*. (1997)
	PSMB1	1447_at	2.6	1.06E−06	Component of 20S proteosome	Nandi *et al*. (1997)
	CTNND2 (delta catenin)	40444_s_at	2.8	2.32E−07	Motility and signaling	Burger *et al*. (2002), Lu *et al*. (1999)
	ARPC3	35810_at	2.5	1.16E−07	Motility and signaling	Robinson *et al*. (2001)
II						
	PSMB3	1309_at	4.4	2.04E−08	Component of 20S proteosome	Nandi *et al*. (1997)
	PSMB4	1311_at	3.4	7.99E−08	Component of 26S proteosome	McCusker *et al*. (1997)
	PSMA2	1446_at	9.1	8.21E−08	Component of 20S proteosome	Nandi *et al*. (1997)
	PSMA3	1448_at	9.9	1.57E−08	Component of 20S proteosome	Nandi *et al*. (1997)
	PSMC1	688_at	3.4	1.59E−07	Component of 26S proteosome	Tanahashi *et al*. (1998)
	UBC13	1660_at	7.1	1.77E−11	DNA repair-RAD 6 pathway	Hoege *et al*. (2002)
	VBP1	171_at	5.2	3.92E−08	VBP1-DNA mismatch repair	Her *et al*. (2003)
	NDUFA5	38462_at	10.5	8.61E−09	Mitochondrial respiratory chain complex	Zhang *et al*. (2002)
	NDUFB1	38605_at	3.7	3.9E−08	Mitochondrial 38 espiratory chain complex	Zhang *et al*. (2002)
	NDUFS4	38695_at	4.2	1.62E−07	Mitochondrial respiratory chain complex	Zhang *et al*. (2002)
	NDUFB3	38981_at	7.9	1.05E−07	Mitochondrial respiratory chain complex	Zhang *et al*. (2002)

demonstrated a high level of correlation between mRNA expression level as detected by the microarray assay and by RT–PCR (average correlation coefficient 0.84) (Mischel *et al*., 2003), these data may suggest important alterations in these cellular processes in astrocytoma that may be clinically exploitable. The list

of genes whose expression levels distinguish normal white matter from astrocytomas, appears robust. Using U133A arrays, 149 of the 170 genes could be reliably mapped and support the distinction of normal white matter from astrocytomas (Fisher's exact test *P*-value = 4.144e−07, data not shown). Further, although it was based on a much smaller number of comparisons, we also found that the oligodendroglioma gene expression profile was highly enriched for proteosomal subunits and genes involved in DNA repair and energy metabolism, potentially suggesting the importance of these alterations in oligodendrogliomas (Table 2).

In summary, we have used gene expression profiling and unsupervised learning methods to demonstrate the presence of three distinct gene expression signatures of gliomas, which correspond to glioblastomas, lower grade astrocytomas and oligodendrogliomas. We also show that primary glioblastomas, which arise *de novo*, are molecularly distinct from secondary glioblastomas that develop from lower grade gliomas and likely constitute a highly heterogeneous group. We demonstrate that a relatively small number of genes characterize the distinction between these pathologic/molecular subsets, and suggest that these gene expression changes reflect potentially important alterations in such critical processes as cellular proliferation, proteosomal function, energy metabolism and signal transduction. These results further define molecular subtypes of gliomas and may potentially be used to define potential targets and further refine stratification approaches for therapy.

Acknowledgements
This work was supported by U01 CA88127 from the National Cancer Institute (SFN) and K08NS43147 from the National Institute of Neurological Disorders and Stroke (PSM). PSM was also supported by an Accelerate Brain Cancer Cure Award, a Henry E Singleton Brain Tumor Fellowship, a generous donation from the Kevin Riley family to UCLA Comprehensive Brain Tumor Program, and the Harry Allgauer Foundation through The Doris R Ullmann Fund for Brain Tumor Research Technologies. Tao Shi is a predoctoral trainee supported by the UCLA IGERT Bioinformatics Program funded by NSF DGE 9987641.

References

Alizadeh AA, Ross DT, Perou CM and van de Rijn M. (2001). *J. Pathol.,* **195,** 41–52.

Burger MJ, Tebay MA, Keith PA, Samaratunga HM, Clements J, Lavin MF and Gardiner RA. (2002). *Int. J. Cancer,* **100,** 228–237.

Coyle JH, Guzik BW, Bor YC, Jin L, Eisner-Smerage L, Taylor SJ, Rekosh D and Hammarskjold ML. (2003). *Mol. Cell. Biol.,* **23,** 92–103.

De Luca A, Mangiacasale R, Severino A, Malquori L, Baldi A, Palena A, Mileo AM, Lavia P and Paggi MG. (2003). *Cancer Res.,* **63,** 1430–1437.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES. (1999). *Science,* **286,** 531–537.

Hastie T, Tibshirani R, Botstein D and Brown P. (2001). *Genome Biol.,* **2**(1), Research 003.

Hastie T, Tibshirani R and Friedman J. (2001). *The Elements of Statistical Learning: Data Mining Inference Prediction*. Springer: New York.

Her C, Wu X, Griswold MD and Zhou F. (2003). *Cancer Res.,* **63,** 865–872.

Hoege C, Pfander B, Moldovan GL, Pyrowolakis G and Jentsch S. (2002). *Nature,* **419,** 135–141.

Kaufmann LAR and Rousseeu PJ. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, Inc.: New York.

Kleihues P, Louis DN, Scheithauer BW, Rorke LB, Reifenberger G, Burger PC and Cavenee WK. (2002). *J. Neuropathol. Exp. Neurol.,* **61,** 215–225 discussion 226–229.

Lal A, Glazer CA, Martinson HM, Friedman HS, Archer GE, Sampson JH and Riggins GJ. (2002). *Cancer Res.,* **62,** 3335–3339.

Li C and Wong W. (2001). *Proc. Nat. Acad. Inc. USA,* **98**(1), 31–36.

Ljubimova JY, Lakhter AJ, Loksh A, Yong WH, Riedinger MS, Miner JH, Sorokin LM, Ljubimov AV and Black KL. (2001). *Cancer Res.,* **61,** 5601–5610.

Lu Q, Paredes M, Medina M, Zhou J, Cavallo R, Peifer M, Orecchio L and Kosik KS. (1999). *J. Cell Biol.,* **144,** 519–532.

Lykke-Andersen J, Shu MD and Steitz JA. (2001). *Science,* **293,** 1836–1839.

MacDonald TJ, Brown KM, LaFleur B, Peterson K, Lawlor C, Chen Y, Packer RJ, Cogen P and Stephan DA. (2001). *Nat. Genet.,* **29,** 143–152.

Matter N, Herrlich P and Konig H. (2002). *Nature,* **420,** 691–695.

McCusker D, Jones T, Sheer D and Trowsdale J. (1997). *Genomics,* **45,** 362–367.

Mischel PS and Cloughesy TF. (2003). *Brain Pathol.,* **13,** 52–61.

Mischel PS, Shai R, Shi T, Choe GC, Horvath S, Seligson D, Kremen TJ, Palotie A, Liau LM, Cloughesy TF and Nelson SF. (2003). *Oncogene,* **22**(15), 8361–8373.

Najib S and Sanchez-Margalet V. (2002). *J. Cell Biochem.,* **86,** 99–106.

Nandi D, Woodward E, Ginsburg DB and Monaco JJ. (1997). *EMBO J.,* **16,** 5363–5375.

Okuda M. (2002). *Oncogene,* **21,** 6170–6174.

Okuda M, Horn HF, Tarapore P, Tokuyama Y, Smulian AG, Chan PK, Knudsen ES, Hofmann IA, Snyder JD, Bove KE and Fukasawa K. (2000). *Cell,* **103,** 127–140.

Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lonning PE, Borresen-Dale AL, Brown PO and Botstein D. (2000). *Nature,* **406,** 747–752.

Pomeroy SL, Tamayo P, Gaasenbeek M, Sturla LM, Angelo M, McLaughlin ME, Kim JY, Goumnerova LC, Black PM, Lau C, Allen JC, Zagzag D, Olson JM, Curran T, Wetmore C, Biegel JA, Poggio T, Mukherjee S, Rifkin R, Califano A, Stolovitzky G, Louis DN, Mesirov JP, Lander ES and Golub TR. (2002). *Nature,* **415,** 436–442.

Rickman DS, Bobek MP, Misek DE, Kuick R, Blaivas M, Kurnit DM, Taylor J and Hanash SM. (2001). *Cancer Res.,* **61,** 6885–6891.

Robinson RC, Turbedsky K, Kaiser DA, Marchand JB, Higgs HN, Choe S and Pollard TD. (2001). *Science,* **294,** 1679–1684.

Sallinen SL, Sallinen PK, Haapasalo HK, Helin HJ, Helen PT, Schraml P, Kallioniemi OP and Kononen J. (2000). *Cancer Res.,* **60,** 6617–6622.

Shav-Tal Y and Zipori D. (2002). *FEBS Lett.,* **531,** 109–114.

Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, Gaasenbeek M, Angelo M, Reich M, Pinkus GS, Ray TS, Koval MA, Last KW, Norton A, Lister TA, Mesirov J, Neuberg DS, Lander ES, Aster JC and Golub TR. (2002). *Nat. Med.,* **8,** 68–74.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P and Borresen-Dale AL. (2001). *Proc. Natl. Acad. Sci. USA,* **98,** 10869–10874.

Tanahashi N, Suzuki M, Fujiwara T, Takahashi E, Shimbara N, Chung CH and Tanaka K. (1998). *Biochem. Biophys. Res. Commun.,* **243,** 229–232.

Venables WN and Ripley BD. (1999). *Modern Applied Statistic with S-Plus*. Springer: New York.

Yano T. (2002). *Mol. Aspects Med.,* **23,** 345–368.

Zhang W, Wang H, Song SW and Fuller GN. (2002). *Brain Pathol.,* **12,** 87–94.