[CANCER RESEARCH 59, 5403–5407, November 1, 1999]

Advances in Brief

A Public Database for Gene Expression in Human Cancers[1]

Anita Lal,[2] Alex E. Lash,[2] Stephen F. Altschul, Victor Velculescu, Lin Zhang, Roger E. McLendon, Marco A. Marra, Christa Prange, Patrice J. Morin, Kornelia Polyak, Nickolas Papadopoulos, Bert Vogelstein, Kenneth W. Kinzler, Robert L. Strausberg, and Gregory J. Riggins[3]

Department of Pathology, Duke University Medical Center, Durham, North Carolina 27710 [A. L., R. E. M., G. J. R.]; National Center for Biotechnology Information, National Library of Medicine, NIH, Bethesda, Maryland 20894 [A. E. L., S. F. A.]; The Johns Hopkins Oncology Center [V. V., K. W. K.], The Howard Hughes Medical Institute [L. Z., B. V.], and Department of Pathology [P. J. M.], Johns Hopkins University School of Medicine, Baltimore, Maryland 21231; Washington University Genome Sequencing Center, St. Louis, Missouri 63108 [M. A. M.]; The I.M.A.G.E. Consortium, Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, California 94550 [C. P.]; Laboratory of Biological Chemistry, Gerontology Research Center, National Institute on Aging, Baltimore, Maryland 21224 [P. J. M.]; Department of Adult Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115 [K. P.]; Department of Pathology, Columbia University, New York, New York 10032 [N. P.]; and Cancer Genome Anatomy Project, Office of the Director, National Cancer Institute, Bethesda, Maryland 20892 [R. L. S.]

Abstract

A public database, SAGEmap, was created as a component of the Cancer Genome Anatomy Project to provide a central location for depositing, retrieving, and analyzing human gene expression data. This database uses serial analysis of gene expression to quantify transcript levels in both malignant and normal human tissues. By accessing SAGEmap (http://www.ncbi.nlm.nih-.gov/SAGE) the user can compare transcript populations between any of the posted libraries. As an initial demonstration of the database's utility, gene expression in human glioblastomas was compared with that of normal brain white matter. Of the 47,174 unique transcripts expressed in these two tissues, 471 (1.0%) were differentially expressed by more than 5-fold ($P < 0.001$). Classification of these genes revealed functions consistent with the biological properties of glioblastomas, in particular: angiogenesis, transcription, and cell cycle related genes.

Introduction

The advent of new technologies has made analysis of overall gene expression patterns a practical reality (1–3). A major goal of CGAP[4] has been the identification of genes that undergo alteration in expression during malignant transformation, in part by capitalizing on these recent technical advances.[5] However, the means to efficiently generate, publicly distribute, and analyze comprehensive gene expression data from multiple laboratories did not exist previously. SAGE was chosen to create a comprehensive quantitative expression database for CGAP, because this assay provides absolute transcript numbers in a digital format (3, 4) and can be easily adapted to provide statistical comparisons of data from multiple laboratories. The basis of SAGE is to count expressed transcripts by sequencing a 14-bp 'tag' of the gene, which is normally sufficient for gene identification. The four bases of the 3′-most *Nla*III restriction site plus the following ten variable bases define the transcript tag. Past SAGE experiments have revealed that a small, but likely important, fraction of a tissue's expressed genes has sustained a substantial change in expression after malignant transformation (5, 6). Unfortunately, the large data sets from these studies are not easily accessible for most researchers, and the time and cost of generating new tumor specific data are generally prohibitive. The goal of this project was to create the means to disseminate to the research community, without restrictions, comprehensive gene expression data for the study of most common human cancers. To initiate this first public quantitative gene expression database, we cloned and sequenced SAGE tags from 25 different normal and tumor tissues and created a system for rapid transfer of the data to the SAGEmap web site. Custom tools were created for this web site to analyze the expression of a single gene in multiple tissues or to compare the full expression pattern between tissues. Furthermore, we demonstrate one use of this database by reporting differentially expressed genes found by comparing glioblastoma with normal brain tissue. This is one of many comparisons that may be performed on SAGEmap.

Materials and Methods

Tumor and Normal Samples. The following samples were used to construct the first 25 SAGEmap libraries: two colorectal carcinoma primary tumors, four colorectal cancer cell lines, two normal colon epithelium samples, four ovarian carcinoma cell lines, one normal ovary surface epithelium, two ovarian tumors, one prostate cell line, one GBM primary tumor, one pool of five GBM primary tumors, one fibrillary astrocytoma, one glioblastoma cell line, two medulloblastoma cell lines, one normal human astrocyte culture, and two normal brain cortical white matter samples. One normal brain white matter library was made from a rapid autopsy sample with a postmortem delay of 3 h and 15 min and confirmed normal histology (provided by the Bryan Alzheimer Disease Research Center, Duke University). The other normal brain library was constructed from normal surrounding brain removed during the surgical approach to remove a seizure focus or a glioblastoma. For each glioblastoma sample a neuropathologist (R. E. M.) confirmed the appearance of a GBM (Grade IV astrocytoma), without necrosis, by examining adjacent stained sections from the sample margins. Immunohistochemistry and slide histology were performed on fixed tissue sections embedded in paraffin. HAM56, a mouse monoclonal antibody (DAKO Corp., Carpinteria, CA), was used to identify macrophage specific markers in primary GBM tissue sections.

SAGE. Total RNA for use in the SAGE protocol was prepared from tissue or cell lines by ultracentrifugation over a cesium chloride gradient followed by mRNA selection by oligo(dT) cellulose (Life Technologies, Gaithersburg, MD). The SAGE libraries were constructed as described previously (3, 7). For the brain libraries, 24 colonies were screened from each library before large-scale sequencing to ensure that they had an insert containing SAGE ditags. Approximately 2500 bacterial colonies were randomly picked by a robot and arrayed in 384-well plates. Plasmids from each colony were purified and sequenced as described previously (8). Transcript tags were extracted from the sequence files, using the SAGE software, v3.03. Tags matching linker sequence (approximately 4%) and duplicate ditags were excluded before analysis or posting on the web site. To estimate the total number of expressed genes in each library, the unique tags were matched to a list of presumed possible 'real' tags.[6] This list compromises only those tags that are unlikely to occur by random sequencing errors, in part, due to the number of

Received 7/20/99; accepted 9/21/99.

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.

[1] Funding for this work was provided by Cancer Genome Anatomy Project (contract S98-146A).

[2] These authors contributed equally to this work.

[3] To whom requests for reprints should be addressed, at Duke University Medical Center, Box 3156, Durham, NC 27710. Phone: (919) 684-5343; Fax: (919) 681-2796; E-mail: greg.riggins@duke.edu.

[4] The abbreviations used are: CGAP, Cancer Genome Anatomy Project; SAGE, Serial Analysis of Gene Expression; EST, expressed sequence tag; GBM, glioblastoma multiforme; RT-PCR, reverse-transcriptase PCR; VEGF, vascular endothelial growth factor.

[5] See http://www.ncbi.nlm.nih.gov/ncicgap.

[6] V. E. Velculescu, S. L. Madden, L. Zhang, A. E. Lash, J. Yu, C. Rago, A. Lal, C. J. Wang, G. A. Beaudry, K. M. Ciriello, B. P. Cook, M. R. Dufault, A. T. Ferguson, Y. Gao, T. C. He, H. Hermeking, S. K. Hiraldo, P. M. Hwang, M. A. Lopez, H. F. Luderer, B. Mathews, J. M. Petroziello, K. Polyak, L. Zawel, W. Zhang, X. Zhang, W. Zhou, F. G. Haluska, J. Jen, S. Sukumar, G. M. Landes, G. J. Riggins, B. Vogelstein, and K. W. Kinzler. Analysis of human transcriptones, submitted for publication.

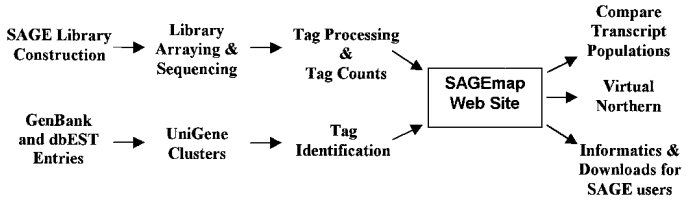

Fig. 1. Schematic of data flow to and from the SAGEmap web site. SAGE libraries from various tumors, normal tissue, and cell lines are constructed at the various laboratories, and the bacterial clones are plated and shipped for arraying and sequencing. From the raw sequencing data, the ten bp of unique SAGE tag sequence is extracted and used as a gene identifier. The National Center for Biotechnology Information manages the informatics and links a UniGene cluster to each tag. Web site users can define groups of libraries and statistical parameters for on-line comparisons. It is also possible to infer gene expression levels by determining a gene's SAGE tag and looking at its frequency in the different libraries, using a virtual Northern function. Mapping information is provided as a resource to SAGE users to enhance the ability to locate the correct gene from a SAGE tag. The data are fully downloadable, for local analysis.

times they were observed in a pool of 3.5 million transcripts. Statistical comparisons of transcript numbers and *P* values presented here were derived from the Monte-Carlo analysis in the SAGE software (5). SAGE tags from colon libraries were a combination of newly generated tags for this project or from a previous study (5).

Northern Blotting and RT-PCR. For northern analysis, total RNA was isolated by ultracentrifugation over a cesium chloride gradient from five normal brain samples, seven glioblastoma primary tumors, three cell lines, and six xenografts. Radioactive labeled PCR probes were made from the following genes (with GenBank accession number in parentheses):

(*a*) *α-glucosidase* (X87237);
(*b*) *p1-cdc46* (X74795);
(*c*) *cartilage-39 glycoprotein* (M80927);
(*d*) *nmb* (X76534);
(*e*) *nicotinamide N-methyltransferase* (U08021);
(*f*) *SPARC* (J03040);
(*g*) *insulin-like growth factor binding protein 2* (M35410);
(*h*) *CAPL* (M80563);
(*i*) *CSRP2* (U57646);
(*j*) *Eph-family protein* (D83492);
(*k*) *SH3GL2* (X99657);
(*l*) *hCDCrel-1* (U74628); and
(*m*) *EST* (AI214960).

Total RNA was electophoresed on an agarose gel and blotted overnight. Bands were normalized to β-actin and compared using autoradiography.

To confirm expression changes by PCR, total RNA was converted to random-primed cDNA using reverse-transcriptase for a panel of three normal brain samples, six glioblastoma primary tumors, two cell lines, and two xenografts. PCR of each cDNA was performed for 20, 25, 28, and 32 cycles using primers specific for *β-actin*, *lipocortin* (GenBank accession X05908), three ESTs (GenBank accessions AI302970, AI022985, and AA903288), and *dynamin* (GenBank accession L07807). The PCR reactions were normalized to the *β-actin* levels, and the products were compared on gel electrophoresis by band intensity to confirm large changes in gene expression between tumor and normal.

SAGEmap Informatics. Tag-to-gene mapping was accomplished by first orienting GenBank sequences using polyadenylation signal and tail and placing the sequences into five "confidence" classes. In each of these confidence classes, tags were extracted from the ten-base tag directly 3′-adjacent to the 3′-most *Nla*III site, and then linked to a UniGene cluster identifier, based on the sequence's UniGene cluster assignment. This full mapping was used to provide the tag-to-gene and gene-to-tag information on the web site. In addition, a 'reliable' tag-to-gene mapping was designed to reduce the effects of sequencing errors in EST GenBank entries. To correct for these errors, tag-gene pairings were first ordered by the frequency that a tag was observed in a particular cluster. Assuming the errors occur randomly and a sequencing error rate of 1% per base, 10 bases have roughly a 10% chance $[1-(0.99)^{10}]$ of having one or more errors. Therefore, the lowest 10% of the rank-ordered tag-gene pairings were eliminated. It was this mapping that is used to make tag-gene assignments for SAGE tag frequency on the web site. The above approach was written into an automatic algorithm, which is used to update SAGEmap weekly.

To avoid the random simulations required to assess statistical significance by the SAGE software (5), which would be impractical for an interactive web site, SAGEmap uses a previously described Bayesian approach (9). We have extended this method so that it can deal with unequal aggregate numbers of tags *A* and *B* from the two libraries being compared. For an mRNA species with concentrations y and z in these libraries, we study the quantity $x = y \div (y + z)$, which we assume has a *prior* probability density $f(x)$ over the interval zero to one. This function captures knowledge about the distribution of relative mRNA concentrations for the complete set of mRNA species. It is reasonable to assume that $f(x)$ will peak at, and be symmetric about, 0.5 (5, 10), and, for simplicity, $f(x)$ may be taken proportional to $x^c(1-x)^c$. Setting $c = 1$ has been proposed (9), but various data (5, 10) suggest a greater concentration of x near 0.5, implying $c = 3$ is a more appropriate value. If *a* and *b* copies of a tag corresponding to a specific mRNA are sequenced from the two libraries, the *posterior* probability density for x (*i.e.*, the density given these data) is proportional to $g(x)$, given by the following equation (derivation omitted).

$$g(x) = f(x)\,\frac{x^a(1-x)^b}{[1+(A/B-1)x]^{a+b}} \qquad \text{(A)}$$

One may calculate from $g(x)$ the posterior probability that the concentration y exceeds z by any given factor. Specifically, y exceeds z by at least the factor F when $x \geq L$, where $L = F \div (F + 1)$. The desired probability is simply the proportion of $g(x)$, viewed as a density, that falls greater than L. The posterior probability that this is the case is given by the following equation.

$$P(x \geq L) = \frac{\int_L^1 g(x)\,dx}{\int_0^1 g(x)\,dx} \qquad \text{(B)}$$

Results and Discussion

More than 1,000,000 transcript tags from 25 different tumor or normal tissue SAGE libraries were cloned, sequenced, and deposited

Table 1 *Comparison of transcript numbers from GBM and normal brain*

Library comparison	Total transcripts compared	Unique transcripts[a]	Elevated transcripts >5-fold[b]	Reduced transcripts >5-fold[b]	% differentially expressed
GBM *vs.* white matter (single cases, no pooling)	164,810	33,582	172	217	1.16
GBM (single case) *vs.* GBM Cell line	128,753	28,571	158	170	1.15
GBM *vs.* white matter (4 libraries, no cell lines)	282,760	47,174	194	277	1.00
All GBM *vs.* all normal brain (with cell lines)	389,881	56,625	175	320	0.87
GBM Cell line *vs.* cultured normal human astrocytes	108,569	23,975	79	97	0.73
GBM *vs.* white matter (pooled library only)	119,717	28,395	63	92	0.55
GBM (single case) *vs.* GBM (pooled case)	132,006	30,325	30	32	0.20
Autopsy white matter *vs.* surgical white matter	152,521	30,769	23	33	0.18

[a] Unique transcripts are the number of nonredundant tags observed one or more times and present in a list of likely tags[6].
[b] Elevated or reduced transcripts were derived by comparing different sets of human SAGE tag libraries and counting the number of transcripts induced or repressed more than 5-fold and were also statistically significant ($P < 0.001$).

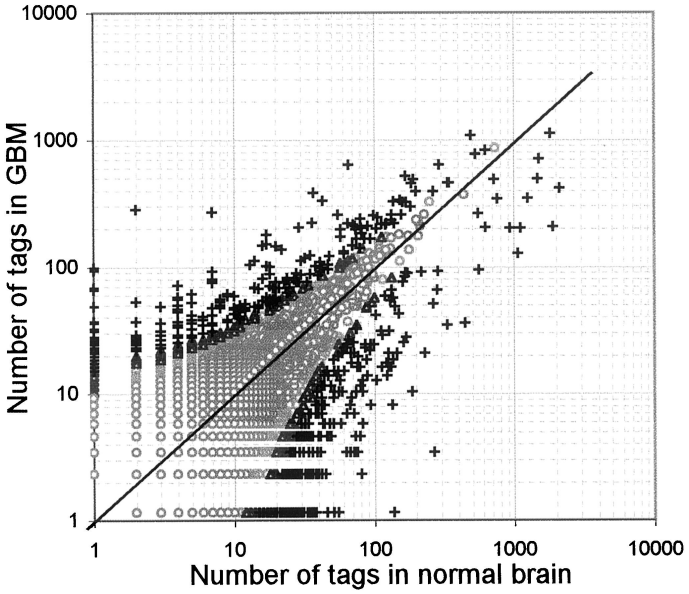

Fig. 2. Distribution of the SAGE tags from surgically resected glioblastomas and normal white-matter libraries. The number of times each unique SAGE tag was observed was plotted on a logarithmic scale, using a total of 107,349 tags from surgically resected glioblastomas (*y-axis*) *versus* 123,880 tags for primary brain white matter (*x-axis*). Tags with no expression in one of the two groups were set to a value of one; and the tag populations were normalized. The line with slope of one through the center predicts equal gene expression in the two tissues. The *P* that an observed difference was real was calculated for each gene. Those genes with a significant *P* less than 0.001 ('+' *symbol in black*) are furthest from the line, whereas more than 99% of the genes fell closer with a *P* between 0.001 and 0.01 (△) or *P* greater than 0.01 (○).

in the SAGEmap database using the project organization shown in Fig. 1. Six of these libraries were designed to study glioblastoma pathogenesis. Some of the possible transcript number comparisons from these libraries are shown in Table 1. From these comparisons we sought one that might provide insight into the pattern of expression necessary for glioblastoma maintenance or formation. Glioblastoma is a deadly, highly malignant form of brain cancer thought to arise mostly from astrocytes present in cortical white matter (11). Therefore, we chose for further analysis the expression comparison between surgically removed glioblastomas and normal brain white matter. Tag counts generated by sequencing the libraries were used to determine transcript expression levels and to find statistically significant differences between the expressed genes. This comparison yielded the expected distribution (5) of the 282,760 SAGE tags (Fig. 2) and showed that 1.0% (471) of the unique expressed genes had more than a 5-fold expression change between the tissues, and a $P < 0.001$.

By matching tag sequences to the predicted tag from GenBank human transcript entries, we identified well-characterized named GenBank entries to 160, or 34%, of the differentially expressed tags. Forty-three percent of these tags matched one or more ESTs. For the remaining 23%, there was no obvious matching GenBank entry, although one could retrieve the gene experimentally by using a PCR-based library screen (10) or by searching other sequence databases for a match. This latter group suggests that there are a significant number of expressed genes in these tissues without a GenBank sequence entry.

Proper tag-to-gene mapping was obvious when the tag identified an accurate full-length sequence, such as most of the named GenBank

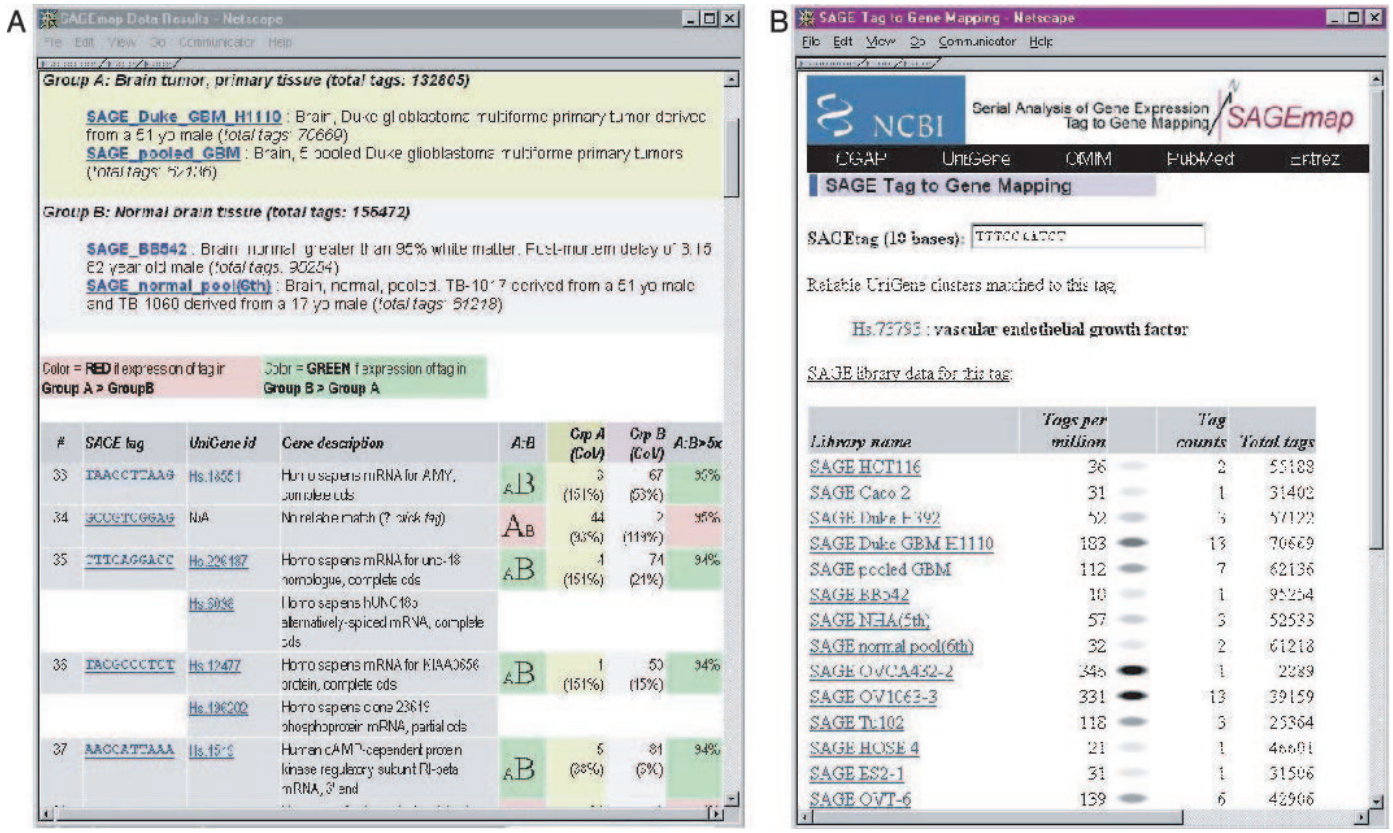

Fig. 3. Screen images from the SAGEmap web site. Users can select libraries or groups of libraries for comparison and view lists of differentially expressed tags (*A*). In this case, only those genes with greater than 5-fold difference were included in the search parameters. Over- or underexpressed genes are color-coded in *green* or *red*, the number of tags in each group listed, a measure provided of the variation of the tag number within a group (*CoV*), a level of statistical confidence that the difference is not due to random chance, and a hyperlink to the relevant tag containing UniGene cluster is provided. SAGEmap users can paste into the virtual Northern tool a sequence of interest (containing the 3′ end of the gene) and the web site will extract the likely tag and show its expression level in all of the available SAGE libraries. Tags produced from library comparisons are also linked to this function. The result of creating a virutual Northern for VEGF, from a known sequence pasted from GenBank, is shown in *B*.

Table 2 *Genes induced or repressed in glioblastomas compared with normal brain tissue*

SAGE tag sequence	Gene symbol (name)[a]	Accession No.[b]	PubMed ID[c]	Fold change[d]
Metaplasia[e]				
GTATGGGCCC	CHI3L1 (*cartilage glycoprotein-39*)	M80927	8245017	Inc 82×
TGGGATTCCC	CHIT1 (*YKL-39 precursor*)	U49835	8702629	Inc 17×
AGTGGTGGCT	FMOD (*fibromodulin*)	U05291	9480886	Inc 16×
GGAAGCTAAG	OSF2 (*osteoblast-specific factor* 2)	D13666	8363580	Inc 10×
Cell cycle regulators				
GAAGGAAGAA	CDK4 (*cyclin-dependent kinase* 4)	M14505	8259215	Inc 33×
ATCTGCTCGG	NMB (*neuromedin B*)	M21551	9178908	Inc 14×
GACTCGCCCA	MCM5 (*P1-cdc46*)	X74795	8265339	Inc 11×
ATCCCTTCCC	PNUTL1 (*peanut-like 1/CDCREL*)	Y11593	9385360	Red 9×
Transcription factors				
CCCAGTAAGA	CSRP2 (*cysteine and glycine-rich protein* 2)	U57646	9286703	Inc 10×
GTCCACACAG	ZNF230 (*Zinc finger protein/FDZF2*)	U95044	2499084	Red 8×
CTCACTTTTT	CEBPD (*nuclear factor IL6 β*)	M83667	1741402	Inc 6×
AACCACTGCT	ID3 (*HLH 1R21/heir*1)	X69111	8437843	Inc 5×
GGGCAGGCGT	(*Transcription factor ETR101*)	M62831	2061303	Inc 5×
ACCCACGTCA	JUNB (*jun B proto-oncogene*)	X51345	2513129	Inc 5×
Angiogenesis				
TTAAACTTAA	CXCR4 (*neuropeptide Y receptor Y3*)	L06797	9634237	Inc 15×
AGATGGTATA	TIMP4 (*tissue inhibitor of metalloproteinase* 4)	U76456	9791730	Inc 13×
AAAGAGAAAG	ADM (*adrenomedullin*)	D14874	9808778	Inc 11×
ATGTGAAGAG	SPARC (*osteonectin*)	J03040	8773299	Inc 10×
ATCTTGTTAC	FN1 (*fibronectin-1*)	X02761	6188316	Inc 8×
TTTCCAATCT	VEGF (*vascular endothelial growth factor*)	AF024710	2479986	Inc 5×
Neurotransmission				
TACCCTTCTG	SYT1 (*synaptotagmin I*)	U19921	7624059	Red 24×
TGGTTCACAT	NPTX1 (*neuronal pentraxin* 1)	U61849	8884281	Red 13×
GCCCCAGCTG	GRIN 1 (*NMDA receptor subunit*)	L13266	7685113	Red 13×
CTTCAGGACC	STXBP1 (*Unc-18 homolog*)	D63851	9545644	Red 13×
GTGCAGTGAA	SYP (*synaptophysin/p38*)	X06389	7903586	Red 12×
TGTCTGTTTG	ZNT3 (*zinc transporter* 3)	U76010	8962159	Red 11×
TCCTGGTGCG	SYNGR1 (*synaptogyrin 1a*)	AJ002305	9760194	Red 9×
TTCCGACTGC	SYNGR3 (*synaptogyrin* 3)	AJ002309	9760194	Red 7×
TCAATCAAGA	YWHAH (*14-3-3 eta*)	L20422	8561965	Red 7×
Immune related				
GCAAAACAAC	LTF (*lactoferrin*)	M93150	7672721	Inc 56×
TGCAGCACGA	HLA-C (*MHC Class I HLA C×52*)	M21963	2843461	Inc 43×
TTCTGTGCTG	C1R (*complement C1r*)	X04701	3030286	Inc 28×
AATAGAAATT	SPP1 (*osteopontin*)	J04765	9916729	Inc 12×
CATCTGTACT	HLA-DRB1 (*MHC Class II HLA DR-β*)	M20429	3259543	Inc 11×
GTTCACATTA	CD74 (*HLA-DR invariant chain p33*)	X00497	6586420	Inc 9×
CTCCCCTGCC	CAPG (*macrophage capping protein*)	M94345	1322908	Inc 8×
Metabolism				
GCCAACAACG	NNMT (*nicotinamide N-methyltransferase*)	U08021	8575745	Inc 24×
GTCAACAGTA	MRP3 (*multidrug resistance protein* 3)	AB010887	9738950	Inc 18×
AGCACTTACA	APOC2 (*apolipoprotein C-II*)	M29844	565877	Inc 16×
TTCTAACATA	ATP1B1 (Na^+K^+ *ATPase β-subunit*)	X03747	3008098	Red 14×
AAATAAAGAA	MGST1 (*microsomal glutathione S-transferase*)	J03746	3372534	Inc 12×
CACGGACACG	GOT1 (*cytosolic aspartate aminotransferase*)	M37400	1974457	Red 11×
TTTGAAATGA	SAT (*spermidine/spermine N1-acetyltransferase*)	M77693	8500690	Inc 9×
Previously reported as overexpressed in glioblastoma				
GAAGGAAGAA	CDK4 (*cyclin-dependent kinase* 4)	M14505	8044775	Inc 33×
AAAGAGAAAG	ADM (*adrenomedullin*)	D14874	9396051	Inc 11×
TCTTGATTTA	A2M (*α-2 macroglobulin*)	M36501	1371755	Inc 10×
TCCAAATCGA	VIM (*vimentin*)	M25246	3020864	Inc 9×
ATCTTGTTAC	FN1 (*fibronectin-1*)	X02761	3974826	Inc 8×
TTTCCAATCT	VEGF (*vascular endothelial growth factor*)	AF024710	1279432	Inc 5×
CGTGGGTGGG	HMOX1 (*heme oxygenase*)	X06985	8694803	Inc 5×

[a] HUGO/GDB nomenclature committee approved symbols for individual genes. Each gene sequence had a poly-A signal and/or poly-A tail and was matched to the SAGE tag.
[b] The GenBank accession number was used to identify the gene and contains the differentially expressed tag.
[c] The PubMed ID is the reference used to support the functional classification or a previous report of elevated expression in glioblastoma (bottom category). PubMed ID can be used to retrieve a reference supporting the functional classification by searching Medline with the indicated number at http://www.ncbi.nlm.nih.gov/PubMed/medline.html.
[d] The fold increase (Inc) or reduced (Red) gene expression was calculated by normalizing the total number of tags in the two SAGE libraries and taking the ratio of the tags in tumor:normal (Inc) or normal:tumor (Red).
[e] A review of publications and databases on each gene was used to group genes with potentially similar function.

entries. However, sequencing errors in ESTs can create ambiguity in the mapping process. To enhance tag-to-gene mapping for ESTs, an algorithm based on UniGene[7] was constructed to ascertain mapping reliability. UniGene has sequence similarity clusters built from overlapping sequences and may have different tag sequences within the same cluster. However, SAGEmap takes into account the number of times a particular tag sequence is observed within each cluster, the orientation of the sequence, and the presence of a polyadenylation signal and tail. Thus, the likelihood that a particular tag is real and is not generated by sequencing errors is assessed and used for proper mapping.

We also constructed a tool for on-line statistical comparisons of tag populations for SAGEmap (Fig. 3*A*). This function allows comparisons between any possible combination of libraries and provides a

[7] See http://www.ncbi.nlm.nih.gov/UniGene.

confidence level for each differentially expressed tag. Desired fold differences can be specified to select the desired tags and libraries for comparison to suit the particular needs of an investigator. A 'virtual northern' function is available that calculates the fractional representation of a particular gene in all of the posted libraries (Fig. 3*B*). In addition, tag frequencies and mapping information are downloadable for local analysis.

Our next step was to validate and analyze a sample of individual genes with altered expression in our glioblastoma *versus* normal brain comparison. At least seven of the overexpressed genes were previously reported as such (Table 2, *bottom*). Eighteen other differentially expressed genes were tested by Northern blot, or when the gene of interest had low expression levels, by RT-PCR. Differential expression was confirmed in the appropriate direction using the original tumors, suggesting that tag-to-gene mapping and the observed tag differences were correct. However, altered expression in the original samples did not normally predict altered expression, by Northern analysis, in more than one-third of other glioblastomas (not shown). This likely reflects the molecular heterogeneity of these tumors (12). Nevertheless, gene alteration in only a small fraction of cases may implicate a pathway involved in tumor pathogenesis.

Functional classification of these glioblastomas' differentially expressed genes was used to provide insight into the pattern of altered expression (Table 2). Although comparing tumor with normal tissues produces an average gene expression difference, the observed changes often reflected our present understanding of tumor biology. For example, there is an expected loss of gene expression related to neurotransmission. The expression comparison was consistent with bone or cartilage metaplasia, which can occasionally be observed by histology in GBMs (11, 13, 14). A group of genes with altered expression were related to macrophage function, which was supported by probing the original tumors with a macrophage-specific antibody. This experiment revealed positive-staining regions adjacent to necrotic regions of the tumor but no staining in normal brain areas (not shown). Also observed in the tumor sections was evidence of angiogenesis, a well-documented feature of glioblastoma (15). This phenomenon was reflected by the 5-fold overexpression of VEGF as determined by SAGE. It is likely that in the 471 differentially expressed genes, there will be others involved in the ability of the tumor to sustain adequate blood supply but not revealed by our functional classification. Experiments aimed at ferreting out these genes from the candidate pool may provide useful therapeutic targets.

SAGEmap is the first completely public database for quantitative gene expression comparisons and provides a central repository for SAGE data. Most major tumor types and normal tissues are planned for representation on this database as well as comparisons designed to provide insight into the major genetic pathways involved in malignant transformation (7, 16–18). Another practical advantage of this data is that they provide candidate tumor-specific genes for immune-based cancer therapy (19, 20). SAGEmap now makes some large-scale expression data and comparisons quickly accessible to any researcher with Internet access. These data demonstrate one aspect of the utility of a gene expression database that provides a prototype for future reporting and dissemination of quantitative gene expression data. It is hoped that the virtual experiments possible using this database will accelerate certain aspects of cancer research.

Acknowledgments

We thank D. Bigner, S. Bigner, H. Friedman, and the Duke Neuro-Oncology Program members for assistance in acquiring samples. G. J. R. is a James S. McDonnell Foundation Scholar and a Novartis Faculty Scholar. Work by C. P. was supported by the United States Department of Energy contract W-7405-Eng-48 to Lawrence Livermore National Laboratory.

References

1. Lockhart, D. J., Dong, H., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., Horton, H., and Brown, E. L. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat. Biotechnol., *14:* 1675–1680, 1996.
2. Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science (Washington DC), *270:* 467–470, 1995.
3. Velculescu, V. E., Zhang, L., Vogelstein, B., and Kinzler, K. W. Serial analysis of gene expression. Science (Washington DC), *270:* 484–487, 1995.
4. Velculescu, V. E., Zhang, L., Zhou, W., Vogelstein, J., Basrai, M. A., Bassett, D. E., Jr, Hieter, P., Vogelstein, B., and Kinzler, K. W. Characterization of the yeast transcriptome. Cell, *88:* 243–251, 1997.
5. Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., Vogelstein, B., and Kinzler, K. W. Gene expression profiles in normal and cancer cells. Science (Washington DC), *276:* 1268–1272, 1997.
6. Hibi, K., Liu, Q., Beaudry, G. A., Madden, S. L., Westra, W. H., Wehage, S. L., Yang, S. C., Heitmiller, R. F., Bertelsen, A. H., Sidransky, D., and Jen, J. Serial analysis of gene expression in non-small cell lung cancer. Cancer Res., *58:* 5690–5694, 1998.
7. Hermeking, H., Lengauer, C., Polyak, K., He, T. C., Zhang, L., Thiagalingam, S., Kinzler, K. W., and Vogelstein, B. *14–3-3* σ is a p53-regulated inhibitor of G_2-M progression. Mol. Cell, *1:* 3–11, 1997.
8. Marra, M. A., Kucaba, T. A., Hillier, L. W., and Waterston, R. H. High-throughput plasmid DNA purification for three cents per sample. Nucleic Acids Res., in press, 1999.
9. Chen, H., Centola, M., Altschul, S. F., and Metzger, H. Characterization of gene expression in resting and activated mast cells. J. Exp. Med., *188:* 1657–1668, 1998.
10. Polyak, K., Xia, Y., Zweier, J. L., Kinzler, K. W., and Vogelstein, B. A model for p53-induced apoptosis. Nature (Lond.), *389:* 300–305, 1997.
11. Kleihues, P., Burger, P. C., Plate, K. H., Ohgaki, H., and Cavenee, W. K. Glioblastoma. *In:* P. Kleihues and W. K. Cavenee (eds.), Pathology and Genetics: Tumors of the Nervous System. Lyon, France: IARC, 1997.
12. Kleihues, P., and Ohgaki, H. Primary and secondary glioblastomas: from concept to clinical diagnosis. Neuro-Oncology, *1:* 44–51, 1999.
13. Richman, A. V., Balis, G. A., and Maniscalco, J. E. Primary intracerebral tumor with mixed chondrosarcoma and glioblastoma–gliosarcoma or sarcoglioma? J. Neuropathol. & Exp. Neurol., *39:* 329–335, 1980.
14. Sarmiento, J., Ferrer, I., Pons, L., and Ferrer, E. Cerebral mixed tumour: osteochondrosarcoma–glioblastoma multiforme. Acta Neurochir., *50:* 335–341, 1979.
15. Jensen, R. L. Growth factor-mediated angiogenesis in the malignant progression of glial tumors: a review. Surg. Neurol., *49:* 189–195, 1998.
16. Polyak, K., Waldman, T., He, T-C., Kinzler, K. W., and Vogelstein, B. Genetic determinants of p53 induced apoptosis and growth arrest. Genes Dev., *10:* 1945–1952, 1996.
17. He, T-C., S. A., Rago, C., Hermeking, H., Zawel, L., da Costa, L. T., Morin, P. J., Vogelstein, B., and Kinzler, K. W. Identification of c-MYC as a target of the APC pathway. Science (Washington DC), *281:* 1509–1512, 1998.
18. Madden, S. L., Galella, E. A., Zhu, J., Bertelsen, A. H., and Beaudry, G. A. SAGE transcript profiles for p53-dependent growth regulation. Oncogene, *15:* 1079–1085, 1997.
19. von Mehren, M., and Weiner, L. M. Monoclonal antibody-based therapy. Curr. Opin. Oncol., *8:* 493–498, 1996.
20. Gilboa, E., Nair, S. K., and Lyerly, H. K. Immunotherapy of cancer with dendritic-cell-based vaccines. Cancer Immunol. Immunother., *46:* 82–87, 1998.