An anatomy of normal and malignant gene expression

Kathy Boon*, Elisson C. Osório†, Susan F. Greenhut‡, Carl F. Schaefer§, Jennifer Shoemaker*, Kornelia Polyak¶, Patrice J. Morin‖, Kenneth H. Buetow§, Robert L. Strausberg‡, Sandro J. de Souza†, and Gregory J. Riggins*,**

*Duke University Medical Center, Durham, NC 27710; †Ludwig Institute for Cancer Research, Sao Paulo 01509-010, SP, Brazil; ‡Office of Cancer Genomics and §Center for Bioinformatics, National Cancer Institute, Bethesda, MD 20892; ¶Dana–Farber Cancer Institute and Harvard Medical School, Boston, MA 02115; and ‖Laboratory of Biological Chemistry, National Institute on Aging, 5600 Nathan Shock Drive, Baltimore, MD 21224

Edited by Bert Vogelstein, The Sydney Kimmel Comprehensive Cancer Center at John Hopkins, Baltimore, MD, and approved June 3, 2002 (received for review May 28, 2002)

A gene's expression pattern provides clues to its role in normal physiology and disease. To provide quantitative expression levels on a genome-wide scale, the Cancer Genome Anatomy Project (CGAP) uses serial analysis of gene expression (SAGE). Over 5 million transcript tags from more than 100 human cell types have been assembled. To enhance the utility of this data, the CGAP SAGE project created SAGE Genie, a web site for the analysis and presentation of SAGE data (http://cgap.nci.nih.gov/SAGE). SAGE Genie provides an automatic link between gene names and SAGE transcript levels, accounting for alternative transcription and many potential errors. These informatics advances provide a rapid and intuitive view of transcript expression in the human body or brain, displayed on the SAGE Anatomic Viewer. We report here an easily accessible view of nearly any gene's expression in a wide variety of malignant and normal tissues.

In which cells of our body is a human gene expressed? Biologists gather this information routinely when investigating a gene. Having a complete picture of gene expression is useful, but relying on *de novo* experiments or deciphering the literature to determine expression levels is slow and inefficient. Electronic databases that catalog gene expression information are beginning to help biologists rapidly access gene expression levels. However, the task of methodically cataloguing all transcripts observed in different human cells is a vast challenge, as is making the information accurate, accessible, and easy to interpret.

Recent technological advances have made large-scale gene expression measurements routine. One of these technologies, serial analysis of gene expression (SAGE) (1), yields transcript counts independent of an arbitrary measure and is well suited to forming a digital gene expression database. SAGE counts polyadenylated transcripts by sequencing a short 14-bp tag at the gene's 3′end, adjacent to the last restriction site, normally *Nla*III. All expressed transcripts with a *Nla*III site can be "tagged" and counted efficiently in large numbers (typically >50,000 per RNA sample) by using automated sequencing. The tag counts are then archived electronically for future analysis and digital comparisons.

The Cancer Genome Anatomy Project (CGAP) SAGE Project is the largest supplier of public gene expression data, and has sponsored a SAGE public database for over 4 years (2–4). These data are posted at the National Center for Biotechnology Information's SAGEmap web site (http://www.ncbi.nlm.nih.gov/SAGE), where SAGE tags are assigned to UniGene clusters, differentially expressed tags can be identified, and the expression level of a particular tag can be displayed (3, 5). SAGEmap is powerful, but there are additional ways of processing and presenting this valuable data, many of which have been requested by the scientific community. There are, however, significant challenges. To create new user-friendly tools to analyze large numbers of genes first requires that gene names be automatically and accurately linked to a SAGE tag. One use of such informatics improvements would be to produce an easy means to view, for the various cells in a human body, the expression of any gene.

We report here SAGE Genie, a set of tools for processing SAGE data. Foremost of these tools is the SAGE Anatomic Viewer, which allows nearly any gene's transcript levels to be easily viewed in normal and malignant tissues. The anatomic view is based on a growing set of over 5.2 million SAGE tags assembled from 114 cell types, plus new web tools to view this data. An enhanced link between SAGE tag and gene is based on an experimental sample of 6.8 million SAGE tags, which was used to evaluate public transcript sequence databases. These informatics allow SAGE Genie to automatically identify SAGE tags from a gene's primary or alternatively polyadenylated transcript while screening for experimental artifacts. A large archive of SAGE data are now more accurately and easily viewed by using SAGE Genie, including a means to see anatomical-based gene expression.

Materials and Methods

Experimental SAGE Data. Data for SAGE Genie were collected as part of the CGAP SAGE Project's effort to create a comprehensive database of human gene expression (3). A network of collaborators supplied SAGE libraries, data, and corresponding information. Libraries were constructed by using *Nla*III as the anchoring enzyme and *Bsm*FI as the tagging enzyme as originally described (1). SAGE 2000 software version 4.12 (http://www.sagenet.org) was used to extract SAGE tags, remove duplicate ditags, and tabulate tag counts. Linker sequences used in library construction and 1-bp variations sequences were also removed from each SAGE library.

Confident SAGE Tag (CST) List. To determine which SAGE tags have been reliably observed in human mRNA, we first assembled 6,800,316 SAGE tags from 171 SAGE libraries. The libraries were derived from both cultured and bulk tissue samples. Tags were obtained from the published human transcriptome (6), CGAP data (ftp://ftp.ncbi.nih.gov/pub/sage/seq/) (3), normal muscle data (7) posted at http://www.urmc.rochester.edu/smd/crc/swindex.html, and brain cancer libraries from the National Cancer Institute Director's Challenge. A table of the various libraries used to derive the CST list is posted online at http://cgap.nci.nih.gov/SAGE/Download. From this compilation of 6.8 million tags, 464,825 were unique. Linker sequences, 1-bp variations, and tag sequences occurring only once were removed, yielding 267,677 unique tags. Sequencing error rates of 6.0%, 4.5%, and 4.5% were estimated for single base pair substitutions, deletions, and insertions, respectively, and errors were filtered as described (6). This procedure removed 73,549 potentially erroneous tags, leaving 194,126 unique tags that represented the corrected tag list of CST. After this processing, 6,319,109 of the original 6,800,316 tag counts remained. The CST list of 194,126

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: SAGE, serial analysis of gene expression; CGAP; Cancer Genome Anatomy Project; CST, confident SAGE tags; EST, expressed sequence tag; MGC, Mammalian Gene Collection; RefSeq, reference sequence.

**To whom reprint requests should be addressed. E-mail: greg.riggins@duke.edu.

tags and counts are available at SAGE Genie (http://cgap.nci.nih.gov/SAGE/Download).

Transcript Sequence Sources. To provide tag to gene links, the following seven sources of cDNA sequences were assembled: (*i*) The October 2001 release of the Mammalian Gene Collection (MGC) (8) of completed full-length cDNA sequences (http://mgc.nci.nih.gov/). (*ii*) The December 2001 update of the National Center for Biotechnology Information (NCBI) Reference Sequence (RefSeq) Project (9) (ftp://ftp.ncbi.nih.gov/refseq/H_sapiens/mRNA_Prot/). (*iii*) Predicted transcripts from chromosome 22 (Sanger Center release 2, May 2001) were used to evaluate genomic predictions. (*iv*) The human mitochondrial genome (GenBank accession no. X93334). (*v*) The "20K set" transcript database was generated by taking the longest non-EST (expressed sequence tag) cDNA GenBank entry for each UniGene cluster (10). (*vi*) "Consensus sequences" databases (Hs_est) were derived from a UniGene Cluster (December 2001 version) assembly and consensus extraction (11). (*vii*) Those ESTs not clustered by UniGene formed the "unclustered EST" databases (Nu).

Virtual Tag Databases. Virtual tags are extracted from transcript sequences and are predictions of the 10 bp regions that might be produced by a SAGE experiment. To form virtual tag databases, the MGC, RefSeq, 20K set, consensus, and unclustered EST databases were divided into subsets based on the presence of a poly(A) tail of at least 5 adenosines (databases ending with "P5R"), either a AAUAAA or AUUAAA poly(A) signal (SR), both signal and tail (P5S) and the remaining entries with neither signal nor tail (R). Four virtual tag databases were formed for each of the above options, extracting a virtual tag adjacent to the last four (3′-most) *Nla*III sites.

Databases that provide internally primed polyadenylated transcript sequences were constructed to provide a means to identify internal SAGE tags that resulted from cDNA synthesis priming from a poly(A) stretch other than the poly(A) tail. When transcript sequences from the MGC, RefSeq, 20K set and Consensus databases were used, any internal stretch (5′ to the last tag) of at least 8 adenosines in a 10-bp region was first identified. Virtual tags upstream of this possible internal priming site were entered into internal primed databases if at least two internally aligning cDNA or EST sequences were found that ended within ±15 bp of the internal poly(A) stretch and had a poly(A) tail of at least 5 adenosines. Virtual tag databases for alternative polyadenylation were constructed in a similar way by locating shorter transcript sequences with a poly(A) signal and a poly(A) tail of 5 adenosines that aligned within the longer entry. The last four virtual tags upstream from the internal poly(A) stretch were extracted from the longer transcript and placed in separate databases.

All of the above processing of seven original sources of cDNA sequence resulted in 105 different virtual tag databases. These databases were used to provide an association not only between a specific tag sequence and a transcript accession number, but also relative tag position and alternative transcripts from the same gene. The 105 databases were ranked by the percent representation of the virtual tags in the CST list (http://cgap.nci.nih.gov/SAGE/DataSets?RANK = 0) to provide a relative measure of reliability for each database.

SAGE Genie Tag Selection. SAGE Genie produces an association between cDNA sequence accession number and SAGE tag. The percent representation of each of the 105 databases in the CST list was used to rank each database, relative to each other. A set of rules were assembled for automated processes where the SAGE Genie provides the best match between gene name and tag, though there are options for manually viewing an alternate

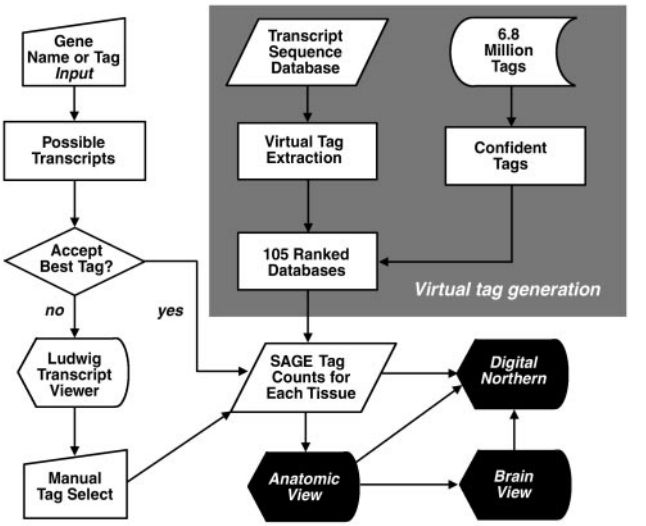

Fig. 1. SAGE Genie design. To form SAGE Genie, virtual tags were extracted from transcript sequence databases and ranked by their level of representation in the CST list of 194,126 experimentally observed tags. A gene name, keyword, or 10-bp tag sequence is entered producing a single best tag and a list of alternate tag to gene associations. Users can also select the Ludwig Transcript Viewer to see a diagram of alternate transcripts from the same gene. For any tag, expression can be graphically displayed in an anatomical context (Anatomic Viewer or Brain View) or in false-color in relation to other libraries (Digital Northern).

tag's expression data. "Best tag" selection is based on the ranking of the database where the virtual tag is observed, if the virtual tag is "internally primed" (yields a lower score), and the expression level of the various tags within a gene (higher expressed genes score higher). Details can be obtained at http://cgap.nci.nih.gov/SAGE/SAGEHelp.

Results and Discussion

A major goal of CGAP is to assemble and distribute quantitative expression profiles for representative normal and malignant cells. In particular, we sought to create a user-friendly method to view gene expression levels in an anatomical context.

To meet this objective, new bioinformatics were required to match SAGE tags to gene transcripts automatically and with a measurable reliability. This was accomplished in three major steps. First a CST list was distilled from 6.8 million experimentally observed SAGE tags. Second, virtual SAGE tags (predicted from cDNA transcript sequences) were obtained online and parsed into 105 databases. These databases reflect the origin of the transcript sequence, the presence of a poly(A) tail, and other features described below, and were ranked based on their CST list representation. In the third major step, customs programs were created to sift through the 105 databases, chose the best tag to gene match, and present the results online (Fig. 1). Alternative transcripts, redundant tags, and internal priming were also considered for tag selection. This CGAP interface, SAGE Genie, has an easy means to view quantitative gene expression by tissue type using the SAGE Anatomic Viewer, as well as providing other SAGE analysis tools (Table 1).

CST. To have a means to evaluate databases of predicted SAGE tags extracted from cDNA sequence (virtual tag databases), we first determined which tag sequences have been reliably observed by SAGE experiments in human cells. A total of 6.8 million SAGE tags, from 171 human SAGE libraries were compiled, covering a wide range of normal and malignant cells. These libraries were all constructed by using the same anchoring

Table 1. SAGE Genie Components posted by CGAP

Name	Purpose
Ludwig Transcript Viewer	For a given transcript sequence, shows alternative polyadenylated or internally primed transcripts that cluster to the same gene.
SAGE Anatomic Viewer	Shades organs or tissues in colors based on levels of a particular gene transcript.
SAGE DGED	Digital Gene Expression Display allows comparison between groups of SAGE libraries to find differentially expressed genes.
SAGE Absolute Level Lister	Lists all the various SAGE Genie libraries. Provides for each library the expression level of every expressed gene, grouped in order.
SAGE Digital Northern	For a specific transcript, provides for each library the normalized expression level and a comparison in color.

enzyme, thus all yielding tags likely to be 10 bp downstream from the 3′-most *Nla*III site in the transcript. This allowed pooling of the tags to create a broad and in-depth sampling of human transcripts.

From the 6.8 million tags, 267,677 different tag sequences were observed more than once. Sequencing errors may have generated some of these tag sequences and we removed potential errors as described (6). This process removes any tag that is not observed 5-fold greater than the expected occurrence by chance sequencing errors. The remaining 194,126 unique SAGE tags formed the CST list. The CST list was used to evaluate virtual tag databases because the list represents transcript tags independently observed from a large range of different human cell types.

Virtual Tag Databases. A series of 105 virtual tag databases were assembled with the purpose of creating an accurate link between a SAGE tag and a transcript sequence. A virtual SAGE tag is the 10-bp sequence adjacent to a *Nla*III site derived from a transcript sequence that corresponds to what might be observed in a SAGE experiment. Virtual tags were derived from a variety of sources, including MGC, Refseq, EST databases, and GenBank. Tags from these sources were further parsed depending on the presence of a poly(A) tail and/or a polyadenylation signal in the transcript sequence, so that tags were grouped based on how accurately the 3′ end of the transcript was defined.

We calculated the percent of cDNA database's virtual tags matching the CST list and used this percent to rank the databases. MGC derived databases predicted the best-represented SAGE tags. Because MGC generates accurate sequence from those cDNA clones first found to be full-length, the commonly occurring 3′ ends appear to be best represented. The percent match is also a reflection of the sequence quality, because errors at the tag site can randomly generate tags not on the CST list. Table 2 summarizes the 3′-most virtual tag databases assembled for SAGE Genie, and their relative ranking. A complete list is available at http://cgap.nci.nih.gov/SAGE/Download.

Missing Genes. The next step was to identify potential errors in our expression database, starting with genes that might be missed by SAGE technology. Less than 1% of the 6,474 sequences from the MGC full-length cDNA database lacked an *Nla*III site. This finding suggests that SAGE misses less than 1% of transcripts because of lack of an anchoring enzyme restriction cleavage site (Table 2). Accordingly, full-length transcripts lacking an *Nla*III site are identified by SAGE Genie and produce an error warning when a query is attempted.

Redundant Genes. Redundancy was also studied by seeing how frequently virtual SAGE tags from the transcript databases matched more than one entry. Some redundancy is genuine because of different transcripts having by chance the same tag, but most databases also have multiple entries for the same transcript. A-10 bp SAGE tag was sufficient to uniquely identify over 95% of the transcripts by using RefSeq data, and more than 98% when transcripts were first clustered (20K set

Table 2. Representation of different virtual tags databases when compared to a list of 194,126 confident human SAGE tags

Database origin	Abbreviation	Total entries	Poly(A) signal	Poly(A) tail	% no *Nla*III sites	% Representation
MGC	MgcSR	10	yes	no	0	100.0
	MgcP5S	3,827	yes	yes	0.8	96.8
	MgcP5R	2,532	no	yes	0.6	95.4
	MgcR	105	no	no	0	89.5
RefSeq	RefSeqP5S	3,135	yes	yes	0.8	90.9
	RefSeqP5R	1,760	no	yes	0.3	88.4
	RefSeqSR	4,480	yes	no	0.5	87.0
	RefSeqR	4,663	no	no	2	68.3
20K set (longest entry in a cluster)	20KP5S	5,040	yes	yes	0.7	90.0
	20KSR	5,433	yes	no	0.6	86.0
	20KP5R	3,629	no	yes	0.5	85.9
	20KR	5,430	no	no	2.5	69.7
Consensus sequences	Hs_estP5S	9,793	yes	yes	10.7	67.0
	Hs_estP5R	7,616	no	yes	11.6	62.3
	Hs_estSR	4,181	yes	no	11.4	54.1
	Hs_estR	29,374	no	no	13.3	42.0
Unclustered ESTs	NuP5S	20,037	yes	yes	34.7	77.1
	NuR	906,237	nd	nd	28.9	43.8

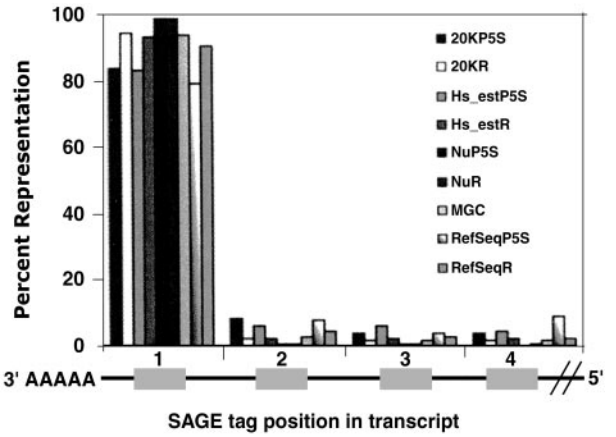

Fig. 2. Expression of SAGE tags by transcript position. The percentage of tags by position is displayed for the last four tag positions from nine virtual tag databases. Regardless of the virtual tag database analyzed, the 3′-most (position 1) virtual tag has the highest average representation. The databases shown are the longest entry in a UniGene cluster (20K), a database of consensus sequences (Hs_est), nonclustering EST sequences (Nu), the MGC, and the Reference sequence (RefSeq) database. The databases were parsed in a poly(A) signal and poly(A) tail containing part (ending in ''P5S'') and a remaining part containing all other entries (ending in ''R'').

of longest UniGene Entry). Therefore, it is likely that a 10-bp SAGE tag in combination with accurate transcript databases can distinguish 98% of the transcripts from different genes.

To reduce the risk of reporting combined expression data from a repetitive tag, we identified 50 different repetitive tags that were present in 20 or more gene clusters. These tags were programmed to produce an error warning on SAGE Genie. For these tags it would be difficult to determine which gene was actually contributing to its expression.

Expression of Internal Tags. There are both biological and experimental reasons why SAGE tags other than the most 3′ tag are observed from the predicted 3′ end of a gene. These ''internal'' tags are genuine when alternative polyadenylation signal usage produces a shorter transcript. Alternative splicing near the 3′ end and polymorphisms in the *Nla*III can also produce shorter transcripts and bona fide ''internal'' tags. However, internal tags could be experimental artifacts. For example, if cDNA synthesis is primed from somewhere other than the poly(A) tail, or if *Nla*III enzyme digestion is incomplete during library construction, then a tag will be produced that is not the 3′-most in that transcript.

Independent of the various ways in which internal tags are generated, we evaluated virtual tag counts for the 3′-most predicted tag and the next three internal tags (Fig. 2). For these four tags, tag counts from the CST list were compared. Regardless of the virtual tag database tested, tag counts were overwhelmingly from the 3′-most tag for all databases (Fig. 2). Therefore, the 3′-most tag will, on average, produce the most highly expressed virtual tag. However, even the small expression of internal tags suggests that they should be considered when linking all tag possibilities to a gene. Internal tags could also be

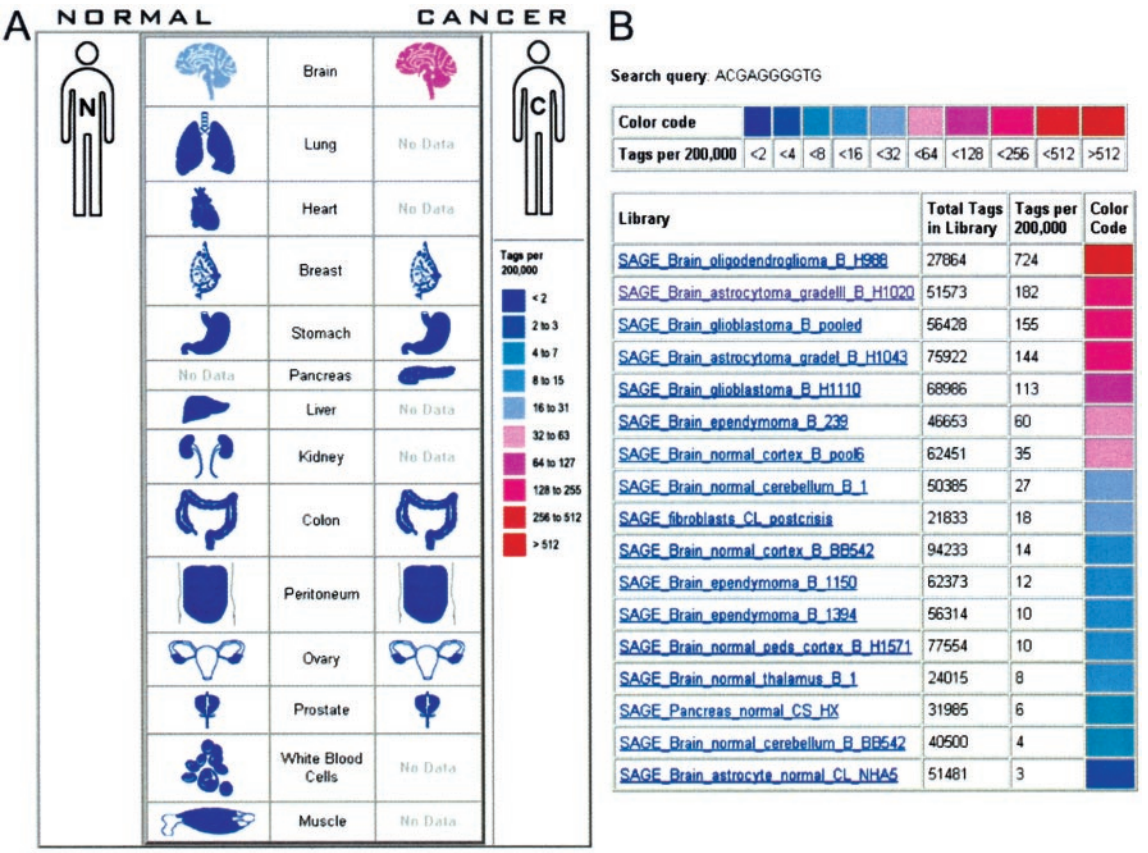

Fig. 3. The SAGE Anatomic Viewer. (*A*) Expression profile for chondroitin sulfate proteoglycan (BC010571) found by the SAGE DGED to be highly expressed in brain malignancies when compared with normal tissues. The brain cancer group appears in pink for an average expression of 32–64 tags per 200,000, compared with all other classes of tissue with average expression of <8 tags per 200,000 (shades of blue). (*B*) Digital Northern of the same highly expressed gene based on all libraries archived at SAGE Genie.

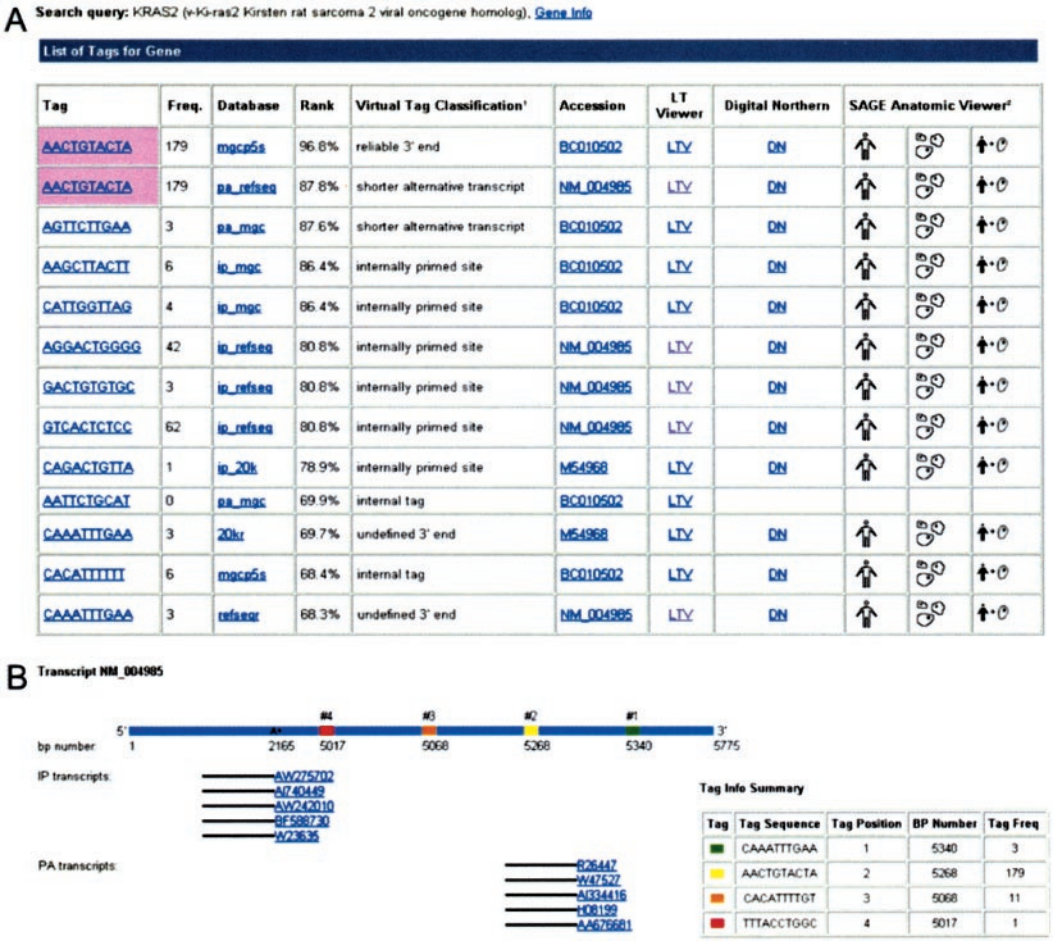

Fig. 4. (*A*) Detailed tag to gene associations obtained for the v-Ki-ras2 rat sarcoma 2 viral oncogene homolog. The best tag selected by SAGE Genie is highlighted. Other tags for the same gene are less abundant based on their tag frequency (expression level in dataset of 5.2 million tags) and are generated as a result of alternative polyadenylation (PA) or internal priming (IP). (*B*) The Ludwig Transcript Viewer shows the transcript sequence encoding v-Ki-ras2 as a blue line. The colored boxes represent the last four virtual tags and show position relative to the 5′ end of the gene, with expression levels in the database provided in the key to the right. An internal stretch of at least 8 adenosines is found at position 2165 and is marked as A+ on the blue line. The accession numbers of sequences that confirm IP transcripts or PA transcripts are given below the corresponding position in the blue line representing the v-Ki-ras2 transcript.

deliberately considered to investigate differential expression of transcripts from a particular gene. We therefore developed additional informatics described below to identify internal tags based on how they might be produced.

Alternatively Polyadenylated Transcripts. Genes that produce alternate 3′ ends can influence SAGE tag usage. To specifically address this influence, alternative polyadenylated virtual tag databases were generated. The databases with the longer transcript sequence entries were blasted against EST and full-length databases to determine whether a shorter entry existed that had a poly(A) tail and signal. If these shorter transcripts predicted an internal tag in the longer transcript, the tag was archived as an alternatively polyadenylated (PA) virtual tag. Experimental evidence was therefore used to support a means by which alternative polyadenylated transcripts could be identified, and the SAGE expression levels for that transcript could be accessed.

Internally Primed Sequences. Evidence was found in the EST databases that cDNA synthesis priming could occur from internal stretches of adenosines, rather than just from the poly(A) tail added by polyadenylase. ESTs with a poly(A) tail of at least 5 adenosines were observed that ended at a transcribed internal stretch of adenosines from a longer transcript sequence. Tag sequences upstream from the confirmed internal poly(A) stretch were extracted from the various entries to form databases of possible internally primed (IP) virtual tags. These databases are used to identify the artifactual generation of SAGE tags from internal priming, and warn SAGE Genie users that a tag is caused by internal priming.

Mitochondrial Tags. Transcripts encoded by the mitochondrial genome, as well as contamination with mitochondrial genomic DNA, can be found by SAGE. All possible 49 tags from the mitochondrial genome were used to form a virtual tag database. Many mitochondrial tags are highly expressed in human cells. SAGE Genie alerts the user if a tag maps to the mitochondrial genome, but still displays its expression level.

SAGE Genie Informatics. The informatics tools that are available on SAGE Genie are shown in Table 1, and Fig. 1 shows the overall design of SAGE Genie. Expression data in the form of SAGE tag counts flow into SAGE Genie from the various labs and sequencing centers. Public transcript data were downloaded, virtual tags were extracted, and 105 databases were formed and ranked for reliability. SAGE Genie informatics allow a single best tag to be selected based on the database ranking, filtering for artifacts and a partial preference for highly expressed tags.

This automatic tag selection allows the association of SAGE data for large numbers of genes, starting with gene names or accession numbers, and can be accessed through the Anatomic Viewer (Fig. 3*A*). Advanced users are still provided with the opportunity to view alternative tags, using tag-to-gene associations provided by SAGE Genie (Fig. 4*A*) and the Ludwig Transcript Viewer (Fig. 4*B*).

The SAGE Anatomic Viewer enables researchers to determine transcript expression in various tissues, for example brain cortex, cerebellum, muscle, blood, heart, liver, kidney, lung, colon, and breast. The database will expand as tags from tissue types are continuously being added. Fig. 3*A* displays the expression profile of a chondroitin sulfate proteoglycan (BCAN) that was found by using the SAGE Digital Gene Expression Display to have high expression mainly in brain cancers. The expression level for a particular gene can also be viewed for all of the libraries by using the SAGE Digital Northern (Fig. 3*B*). Expression levels are color-coded using the same color scheme as the SAGE Anatomic Viewer. SAGE library names for SAGE Genie have been labeled by using a consistent naming convention that contains organ site of tissue origin, tissue histology or pathology, a code for type of tissue purification (or culturing), and a unique identifier.

Differentially expressed genes can be found by comparison for various SAGE libraries using the SAGE Digital Gene Expression Display (SAGE DGED). Highly expressed genes in a given tissue can be listed by using the SAGE Absolute Level Lister (SALL). This list can be downloaded and used for those groups wishing to perform analysis on the genes from a particular tissue. One potential application for SAGE DGED and SALL is to help design custom DNA arrays that target highly or differentially expressed genes for a particular tumor or tissue.

SAGE Genie Testing. SAGE Genie was tested against published tag to gene associations (3, 12–15) by using those tags that were experimentally verified by Northern blotting, real-time PCR, *in situ* hybridization, and/or immunohistochemistry. In total 70 tags of the 77 yielded the reported gene description. The 7 tags that were "mismatched" by SAGE genie were analyzed in detail, and 5 of these tags appeared to be incorrectly assigned in the various publications, perhaps by errors in the reported tag sequence. In one case, a new gene name had been added to a former ORF. In the last case, the authors had chosen a tag from a less abundant cDNA, but still from the same gene. Starting first with gene name and obtaining a tag sequence on SAGE Genie produced the converse: the reported tag in 70 of the 77 test cases. Therefore, SAGE Genie was able to correctly assign a SAGE tag to a gene in nearly all cases, but it is, of course, dependent on the accuracy of the transcript databases from which it is derived. Updated versions of the virtual tag databases for SAGE Genie will continue to improve its accuracy.

Future Directions. SAGE is a powerful technology that has been adopted as a standard for gene expression by CGAP and others. The over 5 million (and growing) SAGE tags make this set of data the most in-depth sequence-based human expression database. This large amount of data are now more easily interpreted and exploited by using SAGE Genie, and its utility will grow with each new tissue added.

SAGE Genie's enhanced tag predictions need to be continually improved. Planned updates with the sequence data from transcript finishing projects will further increase accuracy. Although chromosome 22 genomic sequence predictions were not as reliable as cDNA predictions (data not shown), genomic sequence might still be used. The accuracy of genomic sequence might be combined with cDNA defined 3′ ends to improve virtual tag prediction. Alternatively, another method for generating 21-bp SAGE tags (16) allows transcript mapping directly to genomic sequence.

We have considered both alternative polyadenylation and internal primed generation of transcripts. This provides a rationale to group tags for a given gene. It also allows investigators to observe how alternative transcripts might be differentially expressed between tissues or between tumor and normal. These modular databases provide a foundation for easy addition of new alternative tag databases. Databases that identify alternative tags because of splicing or single nucleotide polymorphisms could be added next.

To help decipher how the human genome is used differently between normal and malignant tissues, an accurate and comprehensive archive of transcript counts is needed. The CGAP SAGE project is characterizing normal and malignant transcriptomes by setting a goal of over 100,000 transcript tag counts from a series of representative tissues. SAGE Genie was created for better analysis and dissemination of these digital gene expression profiles.

We thank CGAP collaborators for SAGE libraries and data (N. Papadopoulos, S. Kern, M. Aldaz, K. Kinzler, B. Vogelstein, A. Lal, I. Siu, J. B. Edwards, B. Fee, R. Beaty, D. Olschner, C. Turner, and S. Powell), clone arraying (C. Prange), sequencing (BC Genome Sequencing Center, National Institutes of Health Intramural Sequencing Core, Agencourt Bioscience), and informatics (A. Lash, G. Bouffard). We also thank V. Velculescu for helpful advice and A. Fabri for Ludwig Transcript Viewer programming. Funding was provided by CGAP and the National Cancer Institue Director's Challenge (U01 CA88128).

1. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) *Science* **270,** 484–487.
2. Strausberg, R. L., Buetow, K. H., Emmert-Buck, M. R. & Klausner, R. D. (2000) *Trends Genet.* **16,** 103–106.
3. Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., *et al.* (1999) *Cancer Res.* **59,** 5403–5407.
4. Riggins, G. J. & Strausberg, R. L. (2001) *Hum. Mol. Genet.* **10,** 663–667.
5. Lash, A. E., Tolstoshev, C. M., Wagner, L., Schuler, G. D., Strausberg, R. L., Riggins, G. J. & Altschul, S. F. (2000) *Genome Res.* **10,** 1051–1060.
6. Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J., Rago, C., Lal, A., Wang, C. J., Beaudry, G. A., Ciriello, K. M., *et al.* (1999) *Nat. Genet.* **23,** 387–388.
7. Welle, S., Bhatt, K. & Thornton, C. A. (1999) *Genome Res.* **9,** 506–513.
8. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. (1999) *Science* **286,** 455–457.
9. Pruitt, K. D. & Maglott, D. R. (2001) *Nucleic Acids Res.* **29,** 137–140.
10. Camargo, A. A., Samaia, H. P., Dias-Neto, E., Simao, D. F., Migotto, I. A., Briones, M. R., Costa, F. F., Nagai, M. A., Verjovski-Almeida, S., Zago, M. A., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 12103–12108.
11. Pagni, M., Iseli, C., Junier, T., Falquet, L., Jongeneel, V. & Bucher, P. (2001) *Nucleic Acids Res.* **29,** 148–151.
12. Lal, A., Peters, H., St. Croix, B., Haroon, Z. A., Dewhirst, M. W., Strausberg, R. L., Kaanders, J. H., van der Kogel, A. J. & Riggins, G. J. (2001) *J. Natl. Cancer Inst.* **93,** 1337–1343.
13. Boon, K., Caron, H. N., van Asperen, R., Valentijn, L., Hermus, M. C., van Sluis, P., Roobeek, I., Weis, I., Voute, P. A., Schwab, M., *et al.* (2001) *EMBO J.* **20,** 1383–1393.
14. Hough, C. D., Sherman-Baust, C. A., Pizer, E. S., Montz, F. J., Im, D. D., Rosenshein, N. B., Cho, K. R., Riggins, G. J. & Morin, P. J. (2000) *Cancer Res.* **60,** 6281–6287.
15. Saha, S., Bardelli, A., Buckhaults, P., Velculescu, V. E., Rago, C., St. Croix, B., Romans, K. E., Choti, M. A., Lengauer, C., Kinzler, K. W., *et al.* (2001) *Science* **294,** 1343–1346.
16. Saha, S., Sparks, A. B., Rago, C., Akmaev, V., Wang, C. J., Vogelstein, B., Kinzler, K. W. & Velculescu, V. E. (2002) *Nat. Biotechnol.* **20,** 508–512.