Commentary

SAGE Genie: A suite with panoramic view of gene expression

Peng Liang*

Department of Cancer Biology, 658 MRB II, Vanderbilt–Ingram Cancer Center, Nashville, TN 37232

If the achievement of complete sequencing of the one-dimensional genetic codes of the human genome can be compared with man landing on the moon, the interpretation of genomic instruction in a four-dimensional biological context, such as during development and diseases, will prove to be a much more challenging and daunting task than that of getting man back from the moon to the earth. One of the greatest mysteries of life has been how a fertilized egg, which contains all of the genetic information that defines a living organism, can give rise to so many different tissues, which organize into different organs, as it divides and differentiates. It is clear that to unravel life's mysteries, we will have to rely, at least in large part, on tools that can allow us to determine when and where a gene is to be turned on or off in a cell as it divides, differentiates, and ages. Obviously, such tools are also important for the detection of when and where a seemingly precise interpretation of genomic instruction goes awry, which underlies many disease states such as cancer.

Several technologies that show promises of high-throughput and potential for global analysis of gene expression were developed in the 1990s (1–3). However, the realization of these promises and potentials has been slow in coming, partly because of the lack of a unified standard for accurate data collection, analysis, and presentation for each methodology. As reported in a recent issue of PNAS (4), Boon and colleagues have made a major stride in this direction by developing a suite of bioinformatic tools that provides a single platform for compiling, annotating, and interpreting large sets of gene expression data collected by one of these technologies, serial analysis of gene expression, or more widely known as SAGE. SAGE technology, which was originally developed by Kinzler and Vogelsteins' group at Johns Hopkins University (2), is a clever high-throughput 3 expressed sequence tag (EST) counting methodology. Unlike the original brute-force EST sequencing strategy, where cDNA clones were randomly picked from cDNA libraries, SAGE technology measures the level of gene expression based on the frequency of occurrence of the 3′ signature SAGE tags of 10 bases unique to each transcript. Because of the minimal sequence information necessary to define an expressed gene, or messenger RNA (mRNA), many SAGE tags from different genes can be obtained and sequenced at a time, which greatly speeds up the EST counting process. The method has been used successfully and extensively in the past for comparison of gene expression between a pair of RNA samples to identify differentially expressed genes within a given biological context (5). Such horizontal comparisons mainly focus on SAGE tags corresponding to genes that are either up- or down-regulated, whereas the bulk of the gene expression information, which took a great deal of effort to collect, often sits untapped. SAGE Genie is a logistically laid out suite of bioinformatic tools that allow automatic and reliable matches of SAGE tags to known gene transcripts. This process was accomplished first by filtering out experimentally obtained SAGE tags that had incorrect linker sequences, appeared only once, or were generated by sequencing errors, from millions of tags collected from over 100 different human cell types as part of the National Institutes of Health Cancer Genome Anatomy Project (CGAP). The resulting confident SAGE tags (CSTs) then were used to evaluate and match the virtual SAGE tags predicted from known mRNA transcript (cDNA) sequences of different publicly available databases, including full-length cDNAs or 3′ ESTs. The virtual tags were divided into different groups based on the origin of the databases from which the tags were generated, the absence and presence of polyadenylation signals and poly(A) tails, and whether the tags represented differentially spliced or internal (non-3) transcript sequences. The match in percentage of virtual tags to CSTs allows ranking of available databases with known transcript sequences. Reciprocal cross-referencing between virtual tags and CSTs provides not only the best match of a CST to a known gene transcript sequence, but also confirmation that experimentally obtained SAGE tags indeed come from mostly 3′ ends of mRNA transcripts. The resulting bioinformatic interface allows automatic tag-to-gene identification, measurement of gene expression normalized to the occurrence of a tag per 200,000 tags collected from a SAGE experiment, and the origins from which a tag is counted. Thus, SAGE Genie provides a computational platform on which not only more than two horizontal comparisons (e.g., normal brain versus brain tumors; Fig. 1), but also a nearly infinite number of vertical comparisons (e.g., different tissue or organ types) in gene expression at a global scale can be conducted. The data output can be presented with interfaces such as the Anatomic Viewer, Digital Northern, and Digital Gene Expression Display for any given SAGE tag or gene transcript of interest, thus providing a quick glance at when and where a gene may be expressed. With SAGE Genie, experimentally collected SAGE tags for each biological system can be continuously annotated and inputted into the growing number of unique CSTs. With increasing collections of both CSTs and virtual tags, SAGE Genie could prove to be a very powerful tool for archiving and analyzing the expression profile for any given gene under any biological context.

SAGE Genie could prove to be a very powerful tool for archiving and analyzing the expression profile for any given gene under any biological context.

In contrast, DNA microarray methodology (3), which has received much attention recently in the field of gene expression analysis, is still lacking a unified standard for

See companion article on page 11287 in issue 17 of volume 99.

*E-mail: peng.liang@vanderbilt.edu.

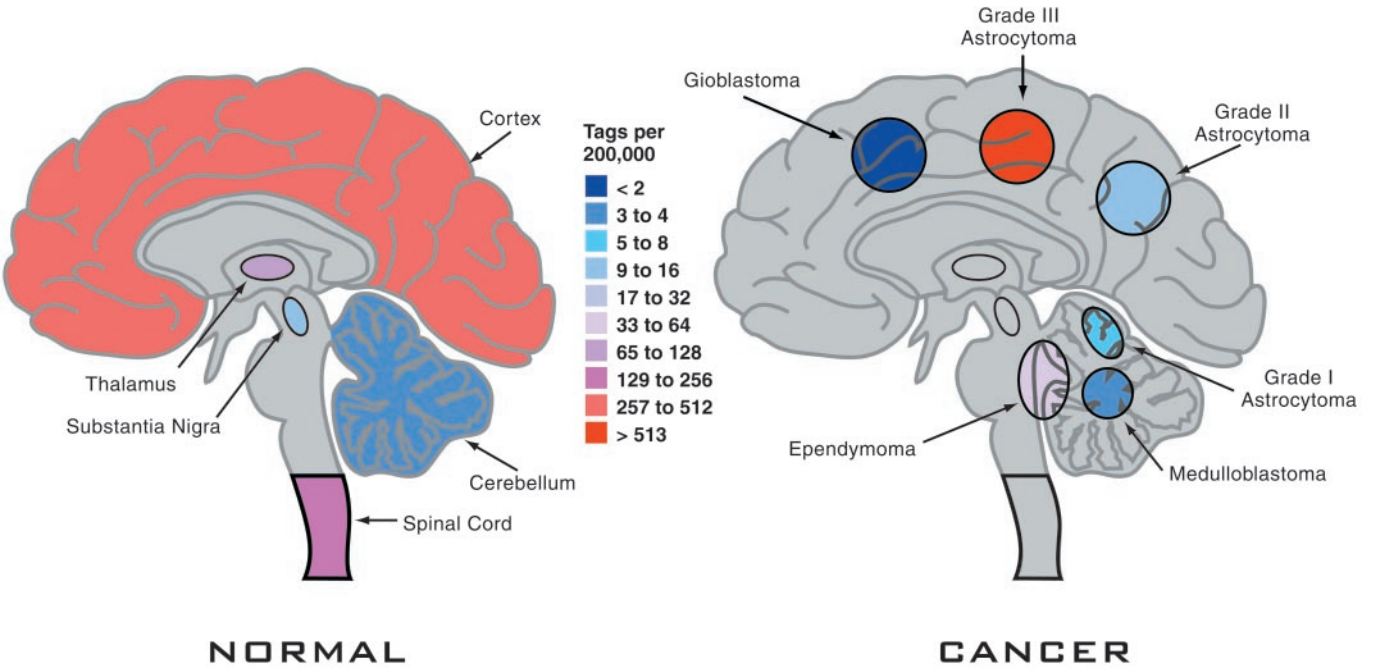

Fig. 1. Diagrams of the human brain showing gene expression levels in false-color (center bar) for a normal brain (*Left*) and or the expression levels for various brain tumors (*Right*). The expression levels are determined by counting mRNA transcripts from each of these tissues. Expression levels are archived for nearly any human gene, and are displayed on-line using CGAP's SAGE Genie.

both data collection and analysis (6, 7). This makes gene expression data archiving and comparison from different research groups difficult. One of the major challenges for microarrays has been determining that a hybridization signal is specific to a known sequence laid on a chip when a complex cDNA probe is used, whereas methodologies such as SAGE are sequence-dependent in gene identification, which is more accurate. In fact, a cDNA probe used for microarrays can be so complex that it consists of as many as 10,000 different species, ranging from a few to thousands of copies per cell. Further compounding the problem in signal specificity for microarrays has been the fact that eukaryotic genes often come in families with many conserved sequences among the family members. Also for microarrays, one is limited to the detection of whatever genes are spotted on a slide, making it a closed system for gene discovery, unlike SAGE and Differential Display (1), which are open system-based gene screening procedures capable of identifying both known and novel gene transcripts.

Although there is no doubt that SAGE Genie has greatly enhanced the utility of SAGE in global analysis of gene expression, challenges remain for the method with regard to the comprehensiveness in gene coverage as a function of the number of tags needed to be counted for each SAGE screen (8), and SAGE tags that either failed to match any known gene transcript sequences or matched more than one known transcript. Messages that failed to be represented because of the lack of anchoring restriction enzyme site are estimated to be as low as 1%. These, in time, may be overcome by increasing the number of SAGE tags collected for future SAGE screens, and the use of longer SAGE tags (9) or different anchoring enzymes. Such improvements may further increase the power of SAGE Genie in archiving and the analysis of gene expression in model tissues/organs or biological systems. But for biomedical and agricultural research, there seem to be an infinite number of comparisons in gene expression with different biological systems, disease states, developmental stages, drug treatment, and stress conditions, etc., which need to be conducted. Such efforts will still require the use of technologies such as arrays and Differential Display as well as SAGE for custom gene-expression analysis. Nonetheless, with an intuitive web-site-based interface, SAGE Genie offers one of the most comprehensive collections of gene expression data across many different cancer and tissue types, making it a valuable tool for a quick glimpse of expression patterns of any known human gene sequences with the need of only a few strokes on a computer keyboard. In contrast to commercial EST collections and microarray databases, one more added bonus for SAGE Genie is that it is accessible to all, free of charge.

1. Liang, P. & Pardee, A. B. (1992) *Science* **257,** 967–971.
2. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) *Science* **270,** 484–487.
3. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270,** 467–470.
4. Boon, K., Osório, E. C., Greenhut, S. F., Schaefer, C. F., Shoemaker, J., Polyak, K., Morin, P. J., Buetow, K. H., Strausberg, R. L., de Souza, S. J. & Riggins, G. J. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 11287–11292.
5. El-Deiry, W. S. (1998) *Semin. Cancer Biol.* **8,** 345–357.
6. King, H. C. & Sinha, A. A. (2001) *J. Am. Med. Assoc.* **286,** 2280–2288.
7. Mills, J. C., Roth, K. A., Cagan, R. L. & Gordon, J. I. (2001) *Nat. Cell Biol.* **8,** E175–E178.
8. Stollberg, J., Urschitz, J., Urban, Z. & Boyd, C. D. (2000) *Genome Res.* **10,** 1241–1248.
9. 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.