High-throughput gene expression analysis using SAGE

Arthur H. Bertelsen and Victor E. Velculescu

The pharmaceutical industry has long been in search of new targets for drug development. A rational approach to the identification of relevant drug targets involves the characterization of gene products that participate in disease processes. The wealth of DNA data generated by the Human Genome Project has identified a substantial fraction of human genes, but has done little to elucidate their role in normal and disease states. One powerful method to reveal insights into gene function and gene pathways is the systematic analysis of gene expression profiles. Serial analysis of gene expression (SAGE) offers an efficient and comprehensive approach to gene expression analysis. It has already been used to provide insights into the pathophysiology of cancer and to open up possibilities for useful diagnostic and therapeutic interventions.

The generation of large quantities of DNA sequence information over the past several years, mostly in the form of completed microbial genomes and expressed sequence tags (ESTs), has led to the notion that we are now entering the 'post-genome era'[1]. Although the sequencing of many genomes, including human, remains unfinished, there already exists a widespread need for analytical tools to take full advantage of this new era. As sequences for most or all of the genes in an organism become available, the challenge of discovery has begun to shift from the identification of genes to the elucidation of their function.

Traditionally, the functional assessment of gene products has often been aided by the methods used to identify genes. Functional clues were often present for genes cloned by linkage to genetic phenotypes or by activity in specific biochemical assays. Recently, however, many genes or gene fragments have been identified through massive sequencing of cDNA clones[2] or genomic regions[3]. Such approaches have made assignment of function more difficult. Homology searches between newly identified genes and existing nucleic acid or protein databases offer some information for the assessment of function, but often this is inconclusive or may lead to misassignment through imperfect homology to relatively short segments of characterized genes. Analyses of protein–protein interactions have been useful in placing gene products in the context of known molecular pathways[4], and phenotype–genotype relationships can be revealed by overexpression or inactivation of genes in model organisms[5] or human somatic cells[6].

Another approach to the determination of gene function involves the analysis of gene expression within an organism. The several thousand different cell types that comprise the human body are each thought to have unique patterns of gene expression specifically 'designed' for their particular physiologic functions. Numerous external or internal agents can modulate these expression patterns, leading to altered physiologic or disease states. The ability to obtain 'global snapshots' of gene expression, both among different cell types and among different states of a particular cell type, can identify candidate genes that may be involved in a variety of

Arthur H. Bertelsen, Schering Plough Research Institute, 2015 Galloping Hill Road, Kenilworth, NJ 07033, USA. **Victor E. Velculescu***, The Johns Hopkins Oncology Center, 424 N. Bond Street, Baltimore, MD 21231, USA. *tel: +1 410 955 8886, fax: +1 410 955 0548, e-mail: velculescu@welchlink.welch.jhu.edu

  1359-6446/98/$19.00. PII: S1359-6446(97)01160-4

normal or disease processes. Additionally, characterization of genes abnormally expressed in diseased tissues may lead to the discovery of genes that can serve as diagnostic markers, prognostic indicators or targets for therapeutic intervention.

A variety of procedures have been advanced for the global evaluation of gene-expression patterns. The promise of many of these methods is that in the future they will be able to characterize fully the set of expressed genes in any cell under study. Two recently developed approaches are nucleic-acid-fragment differential display[7] and hybridization-based analyses using immobilized cDNAs (Refs 8–10) or oligonucleotides[11]. Currently, both of these methods are able to provide gene expression analyses for a few genes in a fairly short period of time. However, they are limited in their ability to compare the abundance of any particular transcript in a population with other transcripts in the same sample. Hybridization-based methods are further restricted in evaluating gene expression because they can only be used to analyze the expression of previously isolated genes. EST sequencing approaches[2,12], although useful for gene identification, are generally unsuitable for quantitative gene-expression monitoring because of the impracticality of sequencing sufficient numbers of cDNA clones from any particular cell or tissue[13].

The principles of SAGE

Recently, a sequencing-based approach has been developed that allows for quantitative analysis of gene expression. This method, termed serial analysis of gene expression (SAGE)[14], is based on two basic principles:

- A short sequence tag contains sufficient information to uniquely identify a transcript;
- Concatenation of tags in a serial fashion allows for increased efficiency in a sequence-based analysis.

The integration of these principles in SAGE is explained below.

Current estimates suggest that there are ~100,000 genes in the human genome[15]. As a 10 bp oligonucleotide sequence has a complexity of >1,000,000 different combinations (4^{10}), a 10 bp sequence tag should be sufficient to uniquely identify a specific transcript, provided that the tag is obtained from a defined position within each transcript. Practical experience has shown that such 10 bp sequences represent a vast excess in complexity over both the number of genes in the human genome and the subset of genes expressed in any mammalian cell, and that >95% of observed tags can be uniquely matched to specific genes.

The most common sequencing methods employ a cloned or PCR-amplified DNA template for the determination of a nucleotide sequence derived from a single gene[15,17]. Multiple genes are sequenced in parallel using multiwell gels or parallel bundles of capillaries into each of which a single sample is loaded.. SAGE exploits its second principle by replacing the conventional DNA sequencing template with one that consists of a tandem linear array of the short nucleic acid fragments derived from different transcripts. The single sequence that is determined from this concatemer thus provides information on multiple genes for the same effort usually required for a single gene. The concatemer contains sequence-specific 'punctuation marks', which delineate the beginning and end of each sequence tag. The increase in efficiency that this serial analysis provides is limited only by the size of high-quality read-length of the sequencing method employed. For conventional automated sequencing technology, SAGE can routinely provide about a 30-fold improvement in efficiency and in the best case more than a 50-fold efficiency increase. In other words, a routine 36-lane sequencing gel evaluates ~1,000 transcripts and the best 36-lane gels identify >1,500 transcripts.

Quantitation of the number of times a particular tag is observed in a population of SAGE tags provides direct information regarding the expression level of the corresponding transcript. As SAGE is a counting-based (i.e. digital) technology, its sensitivity and quantitative accuracy are theoretically unlimited. Practically, these properties are determined by the number of times each tag is observed, as well as the total number of tags identified in a particular tag library. Significant differences in abundances of specific tags among different transcript populations are readily determined using simulations[18] or statistical analyses[19].

For any method that proposes to assess gene expression accurately, it is important to use techniques of analysis that minimize experimental bias. It is already appreciated that methods using processive enzymes (e.g. DNA or RNA polymerases) to copy nucleic acids are prone to sequence-dependent bias. Furthermore, the simple process of cloning a cDNA population is known to affect the relative representation of individual cDNAs in a library. While SAGE uses both amplification and cloning steps, it avoids these pitfalls. The nucleic acid fragments amplified in the SAGE method

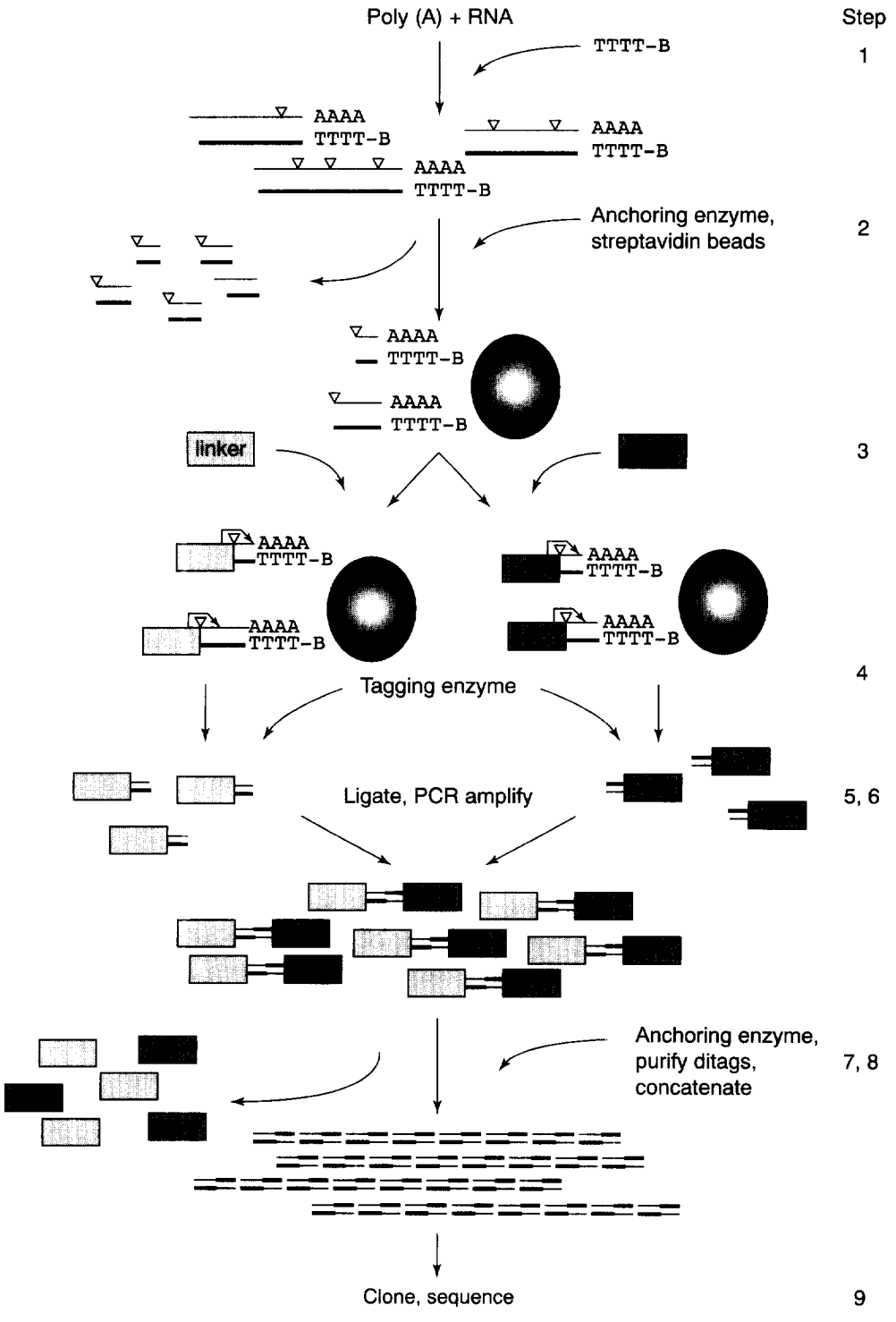

Figure 1. *Schematic of the serial analysis of gene expression (SAGE) method. Poly(A) RNA is converted to cDNA by use of biotin–oligo-dT (step 1). The cDNA is then cleaved by a frequently cutting restriction enzyme (anchoring enzyme) and the 3′-most portion of the cleaved cDNA is captured by streptavidin-coated magnetic beads (step 2). To the captured cDNA a linker is ligated (step 3) that contains a site for a type IIs restriction enzyme (tagging enzyme). The tagging enzyme is then used to release a short tag from the cDNA along with the linker (step 4). The linker–tag molecules are then ligated tail-to-tail to form ditags (step 5) which can be amplified by PCR (step 6). The ditags are then released from the linker molecules by use of the anchoring enzyme (step 7) and subsequently ligated together to form concatemers of many ditags (step 8). These concatemers can be cloned and sequenced (step 9) to reveal the identity of each tag. See text for additional details. B, biotin.*

(SAGE tags) are of short, uniform size, minimizing the propensity for amplification bias. Additionally, before amplification, SAGE tags are ligated to form heterodimers (ditags). This step further reduces bias created by uneven amplification. Finally, SAGE incorporates features that effectively eliminate cloning bias. Because individual clones contain short sequences from many transcripts, each clone in a SAGE library is unique. As a result, the SAGE tags for

any transcript are found in many clones of different composition. If selection against any given clone occurs, it only reduces the tag count for transcripts contained within it by a small fraction, because it will not affect the other SAGE tags for those transcripts that are present in many other clones of entirely different composition.

SAGE procedure

Operationally, the SAGE method is accomplished by a multistep procedure outlined in Figure 1. The position in the transcripts from which sequence tags are derived is defined using a sequence-specific restriction enzyme (anchoring enzyme) to digest the double-stranded cDNA. The 3′-terminal fragments of the cDNA are then captured using streptavidin-coated magnetic beads (steps 1 and 2). After ligating adapter-linkers to the cDNA fragments (step 3), the SAGE tags are liberated from the bound cDNA by cleavage with a type IIs restriction enzyme (tagging enzyme) (step 4), collected and dimerized by blunt-end ligation (step 5). The ditags can be amplified as needed without concern about affecting the quantitative aspects of the subsequent analysis.

Following amplification (step 6), the ditag mixture is digested with the anchoring enzyme that was used to cleave the original cDNA sample, and the ditag fragments are purified away from the linker-adapter fragments (step 7). The ditag fragments are a mixture of cohesive-ended molecules comprised entirely of sequences derived from transcript cDNAs. Ligation of these molecules results in the formation of linear concatemers of ditags (step 8). The size of the concatemers is dependent on the conditions of ligation and can be controlled to yield populations of molecules with lengths that are longer than can be accurately sequenced. Cloning of these concatemers yields a library of SAGE clones for sequencing (step 9). The sequence of each clone consists of multiple ditags, each comprising the sense strand of one tag followed by the antisense strand of its ditag partner, the result of the tail-to-tail blunt-end ligation that was used to form the ditags. The tag sequences in the concatemer can be precisely identified because each ditag is discretely bounded by the recognition sites for the anchoring enzyme used to create the library.

Data evaluation

The SAGE method is dependent on high-quality sequencing procedures. Unlike EST sequencing, which tolerates variation in sequencing read-length and a degree of sequencing errors in the generation of useful data, SAGE requires long-length, high-quality sequencing. Single-base sequencing errors result in the loss of useful information for that tag and can create a new tag entry where one should not exist. Improvements of automated sequencing technologies, that both reduce error rates and increase the sequence read-length, have greatly facilitated the generation of large quantities of SAGE expression information.

The primary sequence data can be evaluated using the SAGE software package[14]. The software extracts the SAGE tag sequences from the raw concatemer sequences, counts the occurrence of each tag, and provides a report containing each SAGE tag and its expression level. Individual SAGE tags can be used to search reference DNA databases to match the tags to the transcripts from which they were derived. For most investigators this sequence reference database will consist of the latest update of the GenBank database, which includes both characterized gene entries and the large amount of ESTs now publicly available. SAGE tags for uncharacterized genes with no entry in GenBank can be used as oligonucleotide probes for the isolation of the cDNAs from which they were derived. This has allowed the rapid identification of many previously undiscovered genes.

Method validation

The description of the SAGE method and its first application documented the presence of 428 different transcripts derived from pancreatic mRNA (Ref. 14). Several of the most abundant transcripts were evaluated to determine the correlation between the abundance predicted by the SAGE analysis and the abundance of the transcript in a cDNA library constructed from the same RNA source. The results indicated that there was good agreement between the two methods. Subsequently, this conclusion has been consistently confirmed for transcripts at lower levels of abundance using SAGE libraries from numerous sources. For example, SAGE analysis of yeast[20] has provided independent confirmation of the abundance of SUP44/RPS4 previously determined by quantitative hybridization analysis[21]. Using rat embryo fibroblast cell lines, a number of genes (e.g. cyclin G, CGR11 and SH80) of known abundance have been identified at expected levels using SAGE (Ref. 22). Thus, the transcript abundance of genes ranging from <0.01% to >1.0% has been accurately corroborated using the SAGE methodology. Moreover, when cDNA libraries have been probed using SAGE

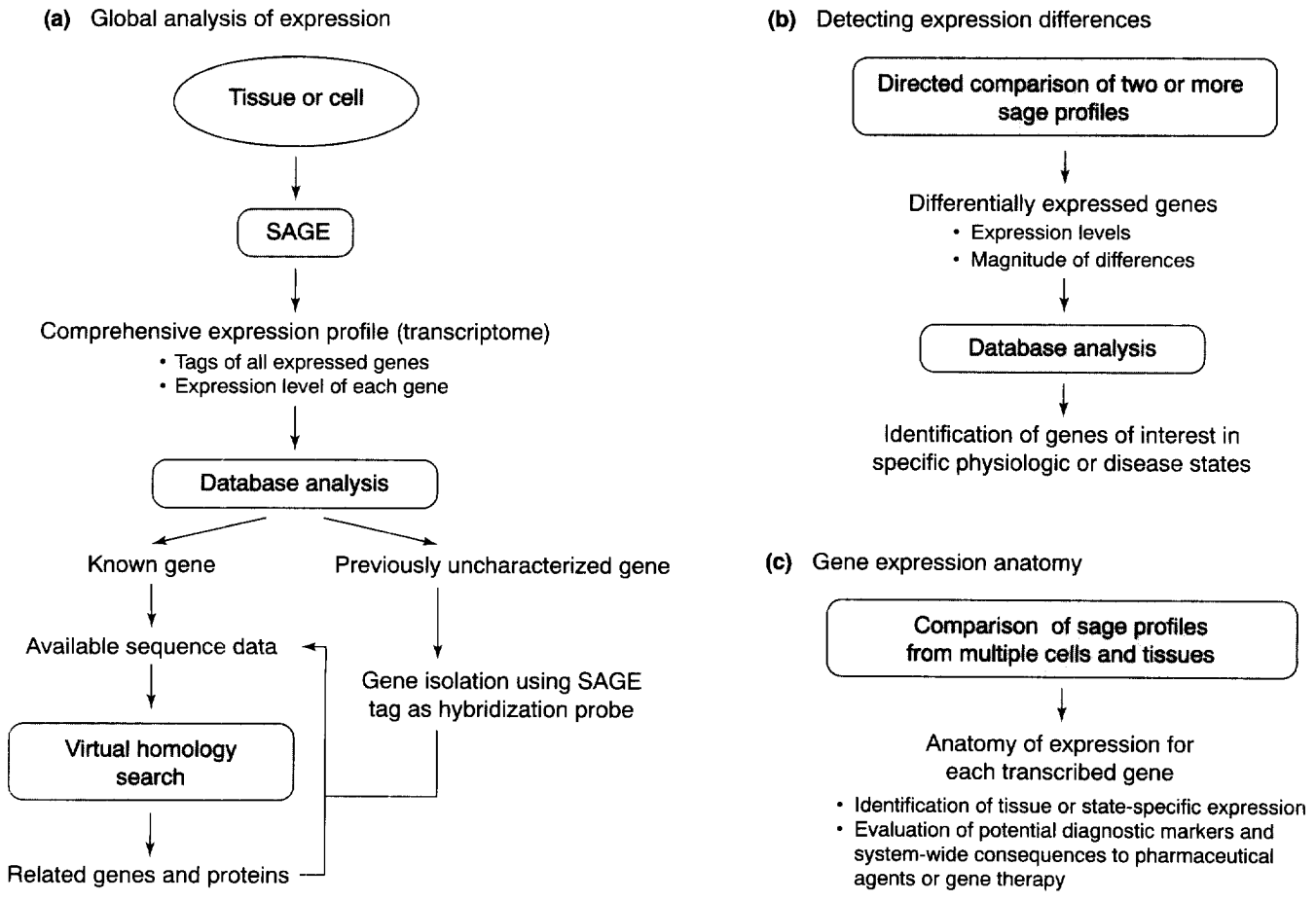

Figure 2. *Applications of serial analysis of gene expression (SAGE). (a) SAGE can provide a comprehensive profile of gene expression for a particular tissue or cell type. The number of times each tag is observed in a SAGE profile provides the expression level of each transcript, and analyses of DNA databases can identify the gene corresponding to each tag. For known genes, the sequence information for the identified gene can be used to perform homology searches to elucidate related genes that may provide clues to function. With SAGE tags that have no matches to the current set of known genes, the tag can be used to generate a probe for hybridization analysis to identify the uncharacterized gene. (b) In order to characterize genes that are differentially expressed in various physiologic or disease states, directed comparisons can be performed between SAGE profiles. The SAGE data can be used to establish expression levels of the differentially expressed genes, as well as the magnitude of expression differences. This information, in conjunction with DNA database analyses, can identify genes for further study in a variety of normal, developmental or pathologic states. (c) Comparison of SAGE profiles from multiple cells and tissues can be used to elucidate the anatomy of gene expression. Individual genes that are of interest in a particular condition can be evaluated across all available SAGE profiles to identify spatial and temporal patterns of expression. Genes that have restricted patterns of expression could be useful for tissue-specific gene therapy approaches. Anatomic gene expression information might be useful in evaluating diagnostic markers or the potential consequences of various therapeutic agents.*

oligonucleotides to isolate cDNAs not present in databases, the number of hybridization signals has consistently agreed with SAGE predictions[14,18]. As a result of this consistency, it has been possible to expedite the isolation of multiple unknown genes in a single hybridization by using mixtures of probes for genes of similar abundance and several cDNA plaques that would give at most a few signals for each different probe.

The reproducibility of SAGE has been demonstrated both by preparing multiple SAGE libraries from one RNA

preparation and also by using independent preparations of RNA from the same cells to construct SAGE libraries and then comparing the abundance of individual transcripts in the different libraries (W. Zhou, pers. commun.). While some differences between samples are observed as a result of sampling variation, these differences disappear when additional tags are analyzed, and therefore sampling variation can be distinguished from true differences in abundance, which persist throughout the analysis. The SAGE software package includes a Monte Carlo simulation feature that can be used to ascertain the probability that a difference in tag number between two samples reflects a true difference in the transcript abundances. This approach minimizes the number of statistical assumptions that must be made in assessing multiple samples.

SAGE applications

Although the potential usefulness of SAGE was presented in the original description of the technology, that report contained a sample of 1,000 transcripts identified using manual sequencing methods[14]. More recent studies describe results obtained using automated sequencing and evaluations of 20,000–50,000 transcripts per SAGE library[18,20,22,23]. The number of tags that need to be analyzed from any library depends on the particular application. In some studies, such as those for identification of potential disease markers, only moderate to highly expressed genes may need to be analyzed; in others, a more comprehensive analysis of the expressed genes, including the low-abundance transcripts, might be warranted. The sensitivity of SAGE increases with accumulation of tags and therefore is directly related to sequencing read-length and time. To identify transcripts present at a level >0.01% in a mammalian cell mRNA population (>30 transcript copies per cell) currently requires sequencing 1,000–2,000 concatemer clones, whereas sequencing 10,000–15,000 clones would identify nearly all transcripts present at as low as about three or more copies per cell.

The SAGE technology is therefore capable of developing comprehensive and quantitative gene expression profiles. Data obtained from individual samples provide expression information that is essentially immortal. Initially, they reveal a picture of global transcript levels within a particular cell or tissue type (Figure 2a). This information is essential in evaluating the expression level of any particular transcript with respect to the many other transcripts inside the cell. As more samples are analyzed, comparisons of profiles from multiple samples can be completed (Figure 2b, c). Comparisons can be performed among tissue or cell types in any organism, regardless of the extent to which the genetic content of the organism has been characterized. This may be particularly useful in a laboratory setting where animal models or cell lines, whose genomes often have not been well described, are frequently the system of choice for a particular experiment. Ultimately the identified differences in gene expression can serve to focus attention on the few genes that should gain high priority for further study from the thousands of genes that are expressed at equivalent levels.

Using SAGE to characterize transcriptomes

A 'transcriptome' has recently been defined as the identity and expression level of the full complement of expressed genes in a given population of cells[20]. In order to characterize transcriptomes fully, information on all of the genes in an organism must be available. The 12 Mb genome of the yeast *Saccharomyces cerevisiae* has been entirely sequenced and determined to contain approximately 6,000 genes[3], making it a convenient experimental model for genome-wide expression analysis.

SAGE was used to characterize yeast gene expression under three different growth conditions: log phase growth, arrest in the S phase of the cell cycle, and arrest in the G2–M phase[20]. Analysis of more than one cell complement of transcripts from each condition led to a number of important observations. Firstly, about three-quarters of genes contained within the yeast genome were found to be transcribed, at levels ranging from less than one copy per cell to several hundred copies per cell. For more than half of these genes this observation provided the first evidence for their expression. Secondly, approximately 160 genes were identified that were essentially undetected by sequence analysis alone, and these included genes that are among the most highly expressed in yeast. Thirdly, there were relatively few dramatic differences in gene expression among the three growth states examined (only 29 genes were more than tenfold differentially expressed among any of the states). Finally, the SAGE analysis confirmed that previous studies using RNA–DNA-reassociation kinetics provided a reasonable estimate of the classes of gene abundance in yeast, although such studies underestimated the contribution of the low-abundance (1–2 copies per cell) class[24].

The information provided by this work indicates that, for a modest effort, a virtually complete analysis of gene expression can be achieved for a variety of experimental

conditions. Given the physical mapping of nearly all the gene sequences in yeast, it is possible to create a graphical representation of the expression pattern by chromosomal location. Such maps may provide a convenient method of rapidly identifying expression patterns that might otherwise go undetected. The consistency of gene expression patterns that was observed among overtly different growth conditions highlights the need for global analyses of gene expression patterns in order to identify significant differences between physiologic states.

SAGE analysis in animal cells

Monitoring gene expression cannot always be performed in systems where the gene content is as well characterized as in yeast or humans. This may pose difficulties for certain technologies that are dependent on the a priori availability of sequence information, but is less of a concern when using SAGE. This is well illustrated in a study recently reported by Madden *et al.*[22] using a rat embryo fibroblast (REF) cell line to identify genes regulated by p53.

The tumor suppressor p53 is an important regulatory protein and the gene encoding it is the most highly mutated gene in human cancers[25]. It acts as a tumor suppressor in at least two ways: by causing growth arrest in some cell types[26] and by activating programmed cell death (apoptosis) in others[27]. There is evidence that at least a part of both activities is regulated by p53 transactivation of gene expression[28]; however, only some of the genes whose expression is affected by p53 have been identified. One model that has been investigated extensively in the study of p53 regulation of gene expression uses REF cells transformed with Ha-ras and a temperature-sensitive mutant of mouse p53 (Ref. 29). Previous studies using differential display, subtractive hybridization and other techniques with these cells have identified a number of p53-regulated genes[30].

To identify new p53-regulated genes, SAGE was used to evaluate gene expression in cells with or without functional p53 expression. The study examined a total of about 60,000 transcripts representing more than 15,000 different genes. Eight genes previously identified in this system were detected as p53-induced, and an additional 24 genes were identified that were not known to be p53-regulated in these cells. When cDNAs for some of these genes were isolated, the level of expression and the degree of induction predicted by the SAGE analysis were confirmed. Unknown rat genes were isolated by hybridization using SAGE tag sequences as probes.

In a similar line of investigation, SAGE has recently been used to directly identify human p53-regulated genes with a role in apoptosis[23]. Analysis of over 100,000 transcripts identified several known and previously uncharacterized targets of p53 that are specifically induced in cells undergoing apoptosis following stimulation of p53 activity. These observations have led the authors to propose a mechanism for p53-induced apoptosis in which p53 activates genes involved in the generation of reactive oxygen species, thereby perturbing the redox status of a cell and ultimately leading to cell death.

Identifying differences between diseased and normal tissue

Perhaps the most attractive aspect of transcript profiling is the use of such information to define the subset of genes that are potential diagnostic markers or therapeutic targets because they differ in expression between normal and diseased tissues. To begin to address the question of expression differences between cancer cells and their normal cellular counterparts, SAGE has recently been used to perform a study analyzing transcript levels in gastrointestinal tumors[18]. This work, which encompassed ~180,000 transcripts from colon samples and 120,000 transcripts from pancreatic cancers, provides the first comprehensive analysis of gene expression in human cancer. This analysis of both primary cancers and cancer cell lines has identified the expression of nearly 50,000 genes, expressed at levels ranging from 1 to >5,000 transcript copies per cell.

Surprisingly, the large majority of genes were expressed at similar levels between cancer cells and normal tissue. Only about 500 genes were significantly up- or downregulated in cancer cells. Many changes in gene expression were retained between primary cancers and cancer cell lines grown *in vitro*, but some differences were observed, suggesting that tumor cell microenvironments play a role in modulating gene expression. Additionally, there was overlap between genes highly expressed in colon and pancreatic cancers, suggesting that some genes with elevated expression in cancer cells may be specific to the neoplastic process, while others may be unique to cancers of certain cellular origins. Analysis of the differentially expressed genes revealed that some of the identified genes had been previously reported as being highly elevated in cancer cells, but that the vast majority had not been previously described in the literature as having any expression changes associated with neoplasia.

As in previous SAGE studies, this work includes abundant evidence that SAGE differences translate directly into RNA differences, as assessed by Northern blot analysis, and, more importantly, shows that many of the alterations identified in just a few samples are consistent with data from a larger sample of primary tumor isolates. These studies suggest that gene-expression monitoring will fulfill the promise of reducing the set of genes that are candidates for functional studies from the tens of thousands of genes that are expressed in cancers and normal cells to a few hundred or less that show wide and consistent disparity in the comparative conditions.

Future perspectives

Analyses of gene expression patterns may soon result in direct improvement in disease diagnosis and treatment. Availability of early diagnostic markers could greatly reduce morbidity and mortality for many illnesses, including cancer. Genes that are highly expressed in disease tissues, especially if they encode secreted proteins, could be measured in the blood or other body fluids. Detection of circulating cancer cells could be facilitated by reverse transcription (RT)-PCR specific for genes expressed in cancer cells but not in normal blood cells. The hope is that early diagnosis, especially of diseases with insidious onsets, would lead to specific interventions and better outcomes. Improvements in staging of a variety of pathological states might be realized by the identification of new tissue- or state-specific markers. Analysis of gene expression differences in treatment responders vs nonresponders could delineate differences between various patient populations and provide insight into the mechanism of action of different treatments. Finally, gene expression patterns could be useful in identifying new targets for therapeutic agents. For cancer cells, highly expressed gene products could be specifically targeted, either because their inactivation would abolish the aberrant growth of the cells or because the gene products could serve as triggers for various cytotoxic compounds or gene therapy strategies[31]. In other diseases, observed lack of expression of certain genes, if pathophysiologically important, could be replaced and normal gene activity restored.

As we enter the post-genome era we face the challenge of elucidating the function of large numbers of newly discovered genes, and using this information to better understand and intervene in various human diseases. While traditional approaches can provide clues to the activity of many gene products, the assignment of physiologic roles for most gene products will require considerable experimental effort and ingenuity. Comprehensive gene expression approaches like SAGE will have a fundamental role in providing basic information integral to biologic and clinical investigation for years to come.

Acknowledgement

Under a licensing agreement between the Johns Hopkins University and Genzyme Molecular Oncology (Genzyme), the SAGE technology described in this article was licensed to Genzyme. The SAGE technology is freely available to academia for research purposes. V.E.V. is entitled to a share of royalty received by the University from sales of the licensed technology, and the University and V.E.V. own Genzyme stock, which is subject to certain restrictions under University policy. V.E.V. is also a consultant to Genzyme. The terms of this arrangement are being managed by the University in accordance with its conflict of interest policies.

REFERENCES

1 Nowak, R. (1995) *Science* 270, 368–371
2 Adams, M.D. *et al.* (1995) *Nature* 377 (6547 Suppl.), 3–174
3 Goffeau, A. *et al.* (1996) *Science* 274, 546–567
4 Fields, S. and Sternglanz, R. (1994) *Trends Genet.* 10, 286–292
5 Capecchi, M.R. (1994) *Sci. Am.* 270, 52–59
6 Shirasawa, S. *et al.* (1993) *Science* 260, 85–88
7 Liang, P. and Pardee, A.B. (1992) *Science* 257, 967–971
8 Derisi, J. *et al.* (1996) *Nat. Genet.* 14, 457–460
9 Schena, M. *et al.* (1995) *Science* 270, 467–470
10 Schena, M. *et al.* (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 10614–10619
11 Lockhart, D.J. *et al.* (1996) *Nat. Biotechnol.* 14, 1675–1680
12 Okubo, K. *et al.* (1994) *DNA Res.* 1, 37–45
13 Bains, W. (1996) *Nat. Biotechnol.* 14, 711–713
14 Velculescu, V.E. *et al.* (1995) *Science* 270, 484–487
15 Fields, C. *et al.* (1994) *Nat. Genet.* 7, 345–346
16 Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. U. S. A.* 74, 5463–5467
17 Du, Z., Hood, L. and Wilson, R.K. (1993) *Methods Enzymol.* 218, 104–121
18 Zhang, L. *et al.* (1997) *Science* 276, 1268–1272
19 Audic, S. and Claverie, J. (1997) *Genome Res.* 7, 986–995
20 Velculescu, V.E. *et al.* (1997) *Cell* 88, 243–251
21 Iyer, V. and Struhl, K. (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 5208–5212
22 Madden, S.L. *et al.* (1997) *Oncogene* 15, 1079–1085
23 Polyak, K. *et al.* (1997) *Nature* 389, 300–304
24 Hereford, L.M. and Rosbash, M. (1977) *Cell* 10, 453–462
25 Harris, C.C. and Hollstein, M. (1993) *New Engl. J. Med.* 329, 1318–1327
26 Kastan, M.B. *et al.* (1991) *Cancer Res.* 51, 6304–6311
27 Shaw, P. *et al.* (1992) *Proc. Natl. Acad. Sci. U. S. A.* 89, 4495–4499
28 Vogelstein, B. and Kinzler, K.W. (1992) *Cell* 70, 523–526
29 Michalovitz, D., Halevy, O. and Oren, M. (1990) *Cell* 62, 671–680
30 Madden, S.L. *et al.* (1996) *Cancer Res.* 56, 5384–5390
31 Da Costa, L.T. *et al.* (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 4192–4196