Serial analysis of gene expression: from gene discovery to target identification

Stephen L. Madden, Clarence J. Wang and Greg Landes

Serial Analysis of Gene Expression (SAGE) is a sequence-based genomics tool that features comprehensive gene discovery and quantitative gene expression
capabilities. As an 'open' system, SAGE can reveal which genes are expressed and their level of expression rather than merely quantifying the expression level
of a predetermined, and presently incomplete, set of genes as carried out by 'closed' system gene expression profiling platforms such as microarrays. These distinguishing attributes enable SAGE to be used as a primary discovery engine that can characterize human disease at the molecular level while illuminating
potential targets and markers for therapeutic and diagnostic development, respectively.

Improvements in the development process for the next generation of therapeutic products requires a strategy to overcome the 96% attrition rate currently observed between drug discovery projects and new drugs in the marketplace[1,2]. This strategy must accept that the classical development mode of screening chemical compounds for potential therapeutic effects on unknown targets is now ineffective and, therefore, costly in today's marketplace. What is required is a 'contemporary' strategy, that is, one directed towards identifiable therapeutic targets while addressing drug development attrition by yielding not only more targets but biologically relevant targets.

Meeting this aim requires an improved understanding of the pathophysiology of human disease at the molecular level to elucidate alterations in biochemical pathways associated with disease phenotypes. These pathway changes reflect the programmatic alterations in expression resulting in the disease phenotype. Elucidating these changes can reveal disease-associated processes and focus diagnostic and therapeutic development efforts on relevant disease markers and/or targets, respectively. Both gene and protein expression profiling methodologies have emerged to monitor and inventory changes in the expression of genes and gene products. For the purpose of this article, the discussion will be limited to gene expression profiling in general and, specifically, Serial Analysis of Gene Expression (SAGE).

Development of SAGE

SAGE, developed in 1995 by Velculescu, Vogelstein and Kinzler at Johns Hopkins University (Baltimore, MD, USA), is a sequence-based approach that identifies which genes are expressed and quantifies their level of expression[3]. This catalog of gene expression for a given cell type or tissue is defined as the 'transcriptome'. Generation of comprehensive and representative transcriptomes requires exacting transcript identification and accurate quantification using unbiased and highly efficient molecular processes. SAGE was developed to satisfy these functional requirements using three founding principles.

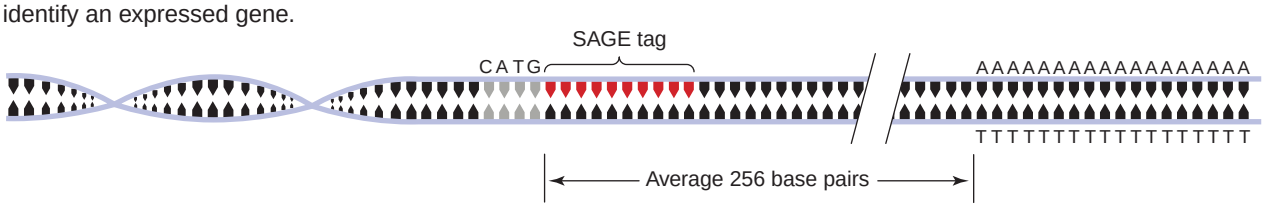
Firstly, SAGE uses a short contiguous sequence of 10–11 base pairs (bp), derived from a defined location within each transcript and unique to each tag, to identify individ-

*Stephen L. Madden, Clarence J. Wang and Greg Landes, Genzyme Molecular Oncology and Genzyme Corporation, PO Box 9322, Framingham, MA 01701-9322, USA. *tel: +1 508 270 2175, fax: +1 508 620 1203, e-mail: steve.madden@genzyme.com

 1359-6446/00/$ – see front matter PII: S1359-6446(00)01544-0

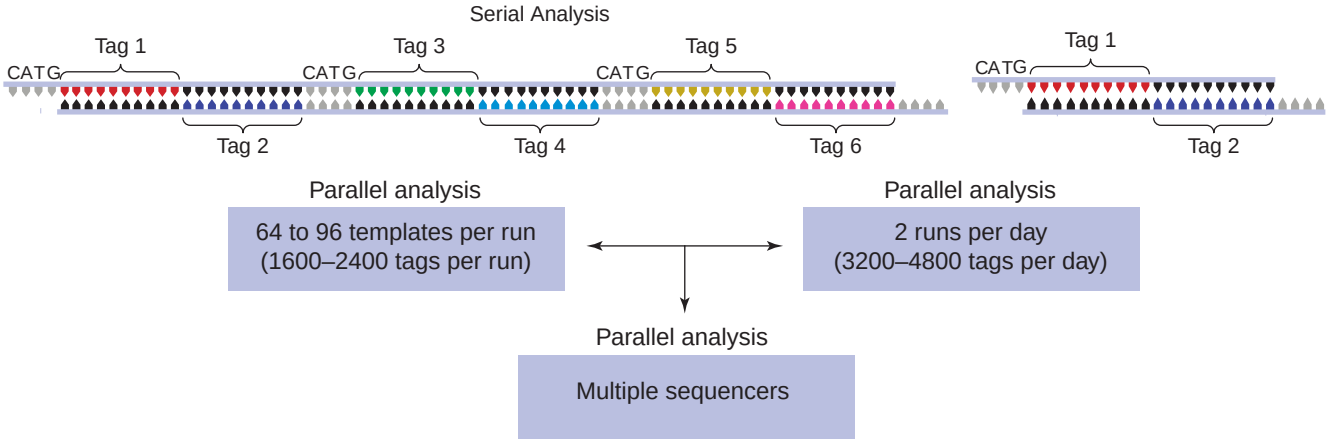
(a) SAGE principle 1

A short oligonucleotide sequence from a defined location within a transcript, a 'tag', encodes sufficient complexity to identify an expressed gene.



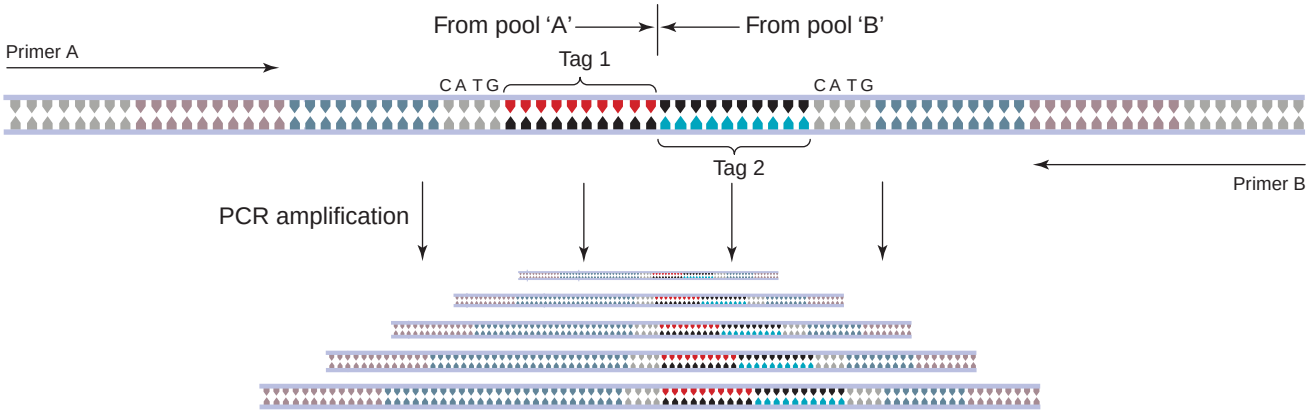
(b) SAGE principle 2

A combination of serial and parallel analysis maximizes throughput.



(c) SAGE principle 3

Minimize, recognize and eliminate PCR-mediated amplification bias.
Defer PCR until amplicons are almost equivalent in size (100 bp) and composition (~70% adaptors).



Recognize and eliminate residual amplification bias by using each ditag sequence once per project.

Drug Discovery Today

Figure 1. *(facing page) (a) Serial Analysis of Gene Expression (SAGE) Principle 1: Representation of a typical 3′-end of a cDNA molecule denoting the ultimate NlaIII site (CATG) and the adjacent SAGE tag. (b) SAGE Principle 2: Concatenation of multiple gene-specific tags within individual clones enables serial detection of tags. Multiple lanes and sequencers enable parallel processing of serially aligned SAGE tags. (c) SAGE Principle 3: Random association of any given tags to form ditag PCR templates.*

ual mRNAs (Fig. 1a). SAGE tags are not only used for gene identification, but are also used to measure the relative abundance of their cognate transcripts within the mRNA population based on the number of occurrences of a given SAGE tag within a SAGE sequencing project. Secondly, to overcome throughput limitations associated with sequencing-based approaches, SAGE uses serial processing such that 25–50 transcripts (or SAGE tags) are analyzed on each lane of an automated DNA sequencer while continuing to use parallel processing with multiple sequencing lanes and sequencers operating simultaneously (Fig. 1b). Finally, as with many contemporary expression analysis methods, SAGE uses polymerase chain reaction (PCR) amplification, but is unique among these in having a mechanism for recognizing and eliminating amplification bias from the expression profile (Fig. 1c). Consequently, individual transcript representation is maintained when analyzing complex natural mixtures of mRNA.

SAGE: the method

Each mRNA population to be analyzed by SAGE requires the construction of a library of clones containing concatenated SAGE tags. Library generation has been described previously[3], but will be briefly reviewed here (Fig. 2). As in all gene expression profiling methods, mRNA is converted to double-stranded cDNA using oligo(dT) to prime first-strand synthesis (Fig. 3a). In this instance, the oligo(dT) primer contains a 5′-biotin moiety to enable recovery of 3′-cDNA fragments. The resulting double-stranded cDNA is digested with the restriction enzyme NlaIII (anchoring enzyme), which recognizes and cleaves DNA immediately 3′ of the sequence CATG. This digestion step creates the defined location within each cDNA for subsequent excision of the adjoining SAGE tag. Biotinylated 3′-cDNAs are affinity-purified using streptavidin-coated magnetic particles (Fig. 3b). The 5′-termini of the captured cDNAs are halved then linker-adapted at their 5′-ends using oligo duplexes encoding a NlaIII 4-nucleotide (nt) cohesive overhang, a Type IIS recognition sequence (BsmFI), and a PCR primer sequence (primer A or B). The adapted cDNAs are digested with BsmFI (tagging enzyme), which cleaves 14–15 bp 3′ of its recognition sequence, releasing the linker-adapted SAGE tag from each cDNA (Fig. 3c).

It is worth mentioning that ≈20% of the cleavage sites for BsmFI are 14 bp downstream of its recognition sequence while the majority of the cleavage sites are 15 bp downstream of its recognition sequence. The linker-adapted SAGE tags from each pool are repaired using DNA polymerase (Klenow), mixed together and then ligated using T4 DNA ligase. The resulting linker-adapted ditags are amplified by PCR using primers A and B, digested with NlaIII to release the primer-adapters, and the SAGE ditags purified (Fig. 4a). The SAGE ditags are then polymerized using T4 DNA ligase, size-selected and cloned into a high-copy plasmid vector. Each cloned insert is organized as a concatenated series of ditags of 20–22 bp in length, separated by the 4-bp recognition sequence for the anchoring enzyme NlaIII (Fig. 4b).

Each SAGE library contains ≈2×10^6 SAGE tags (or transcripts) based on ≈5×10^4 colony forming units per

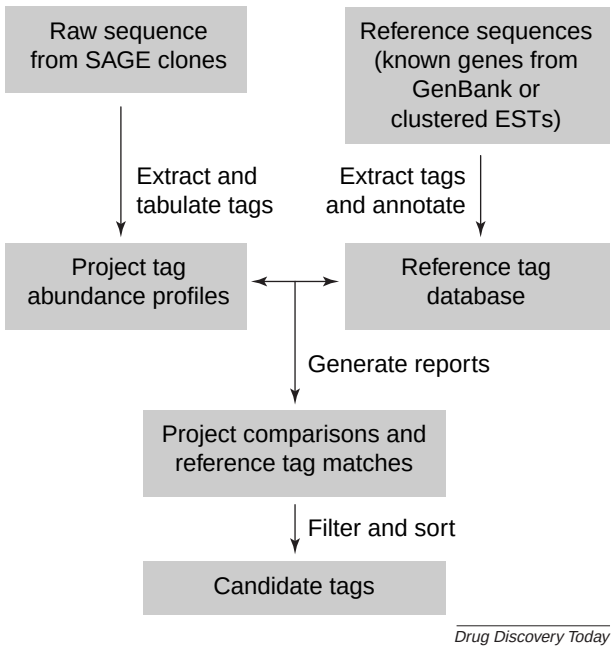


Figure 2. *Serial Analysis of Gene Expression (SAGE) data analysis. Transcript abundance profiles are generated from sequencing of tags in SAGE clones. Observed tags are matched to reference sequences from various sources. Sorting and filtering of tags based on differential expression reveals candidates for further investigation. Abbreviation: EST, expressed sequence tag.*

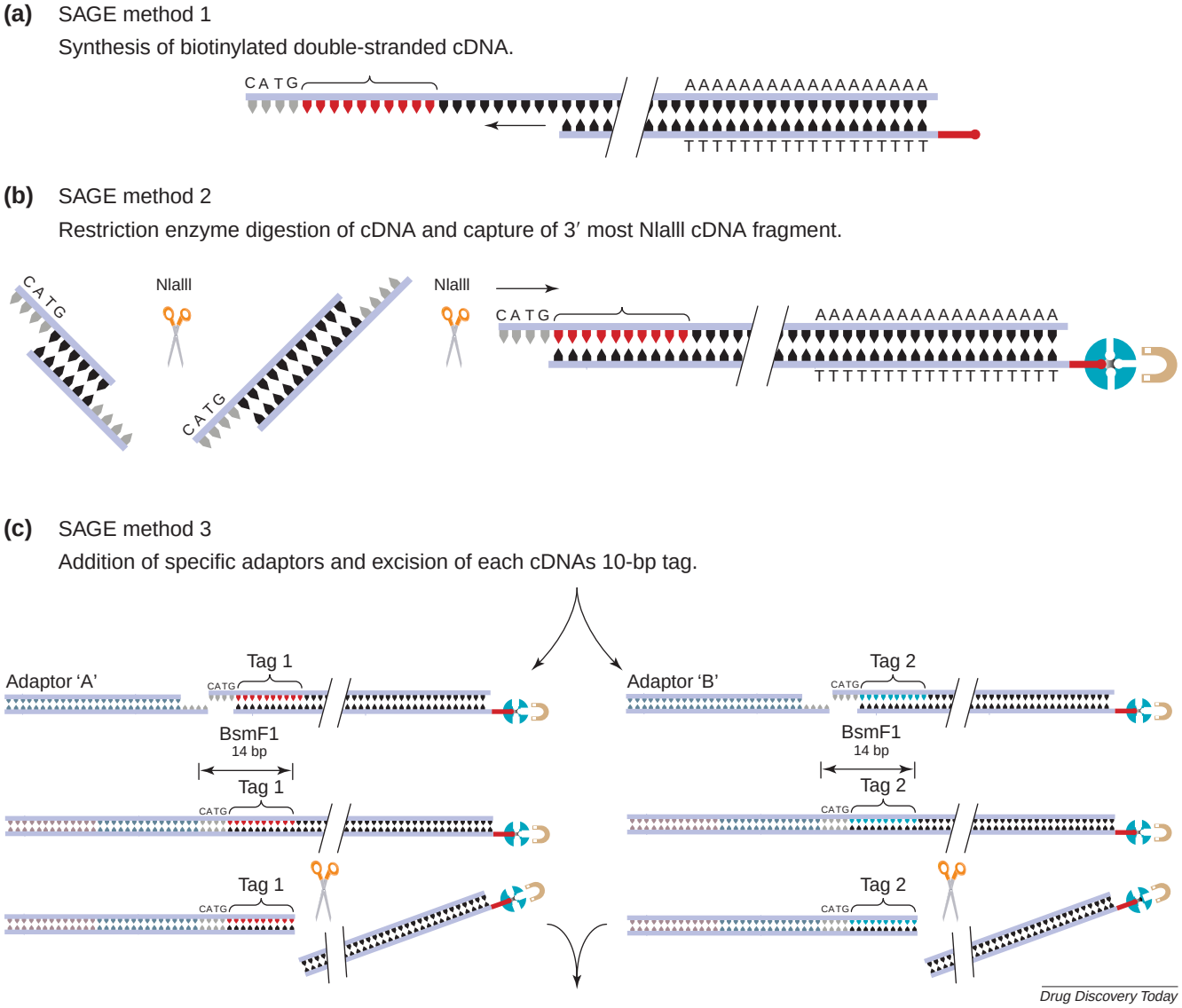


Figure 3. *The Serial Analysis of Gene Expression (SAGE) method Part I. (a) Synthesis of double-stranded cDNA. (b) Restriction enzyme digestion of cDNA and capture of 3′-cDNA fragments. (c) Adaptor modification and excision of tags.*

library, with cloned insert sizes of ≈500 bp (40 SAGE tags per clone). For most projects, ≈2000 individual SAGE clones are sequenced to yield ≈50 000 SAGE tags. Standard sequencing chemistries and platforms are used with the resulting sequence outputs analyzed by custom software.

SAGE bioinformatics

Data analysis software

The power of SAGE as a comprehensive and quantitative transcript profiling method relies on efficient computational tools for data generation, management and analysis. The custom tool for these tasks is the SAGE Software Suite (Genzyme Molecular Oncology, Framington, MA, USA), a PC-based application designed to handle processing and analysis of tag data from SAGE experiments or projects[3].

The core functions of the SAGE software are:

- Extraction and tabulation of tag sequences and counts from raw sequence files
- Comparison of tag abundances between projects
- Matching tags to reference sequences in other databases.

The tag extraction function takes input from concatenated ditag sequences in text files generated from sequencing of

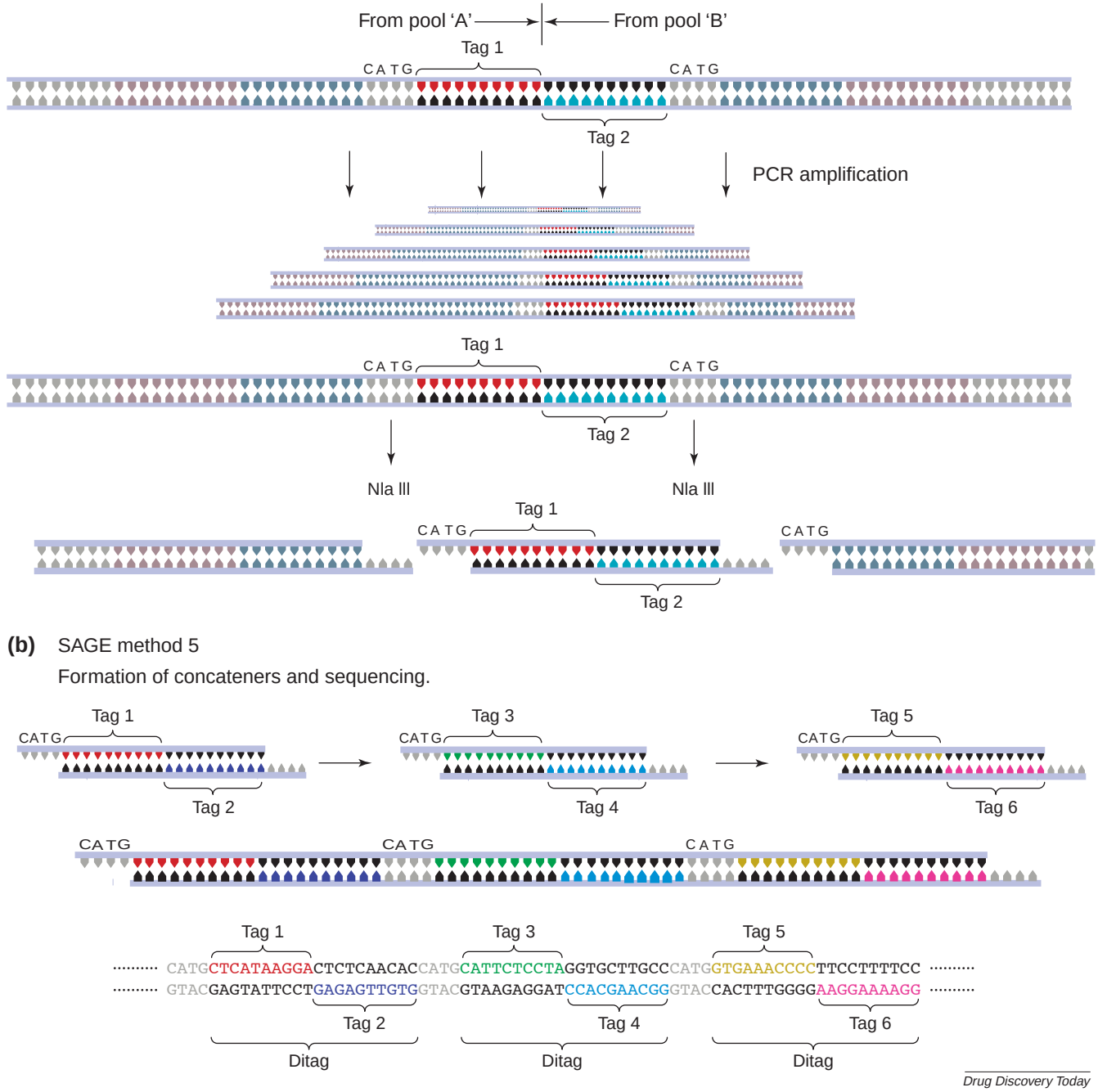


Figure 4. *The Serial Analysis of Gene Expression (SAGE) method Part II. (a) Ditag formation, amplification and adaptor removal. (b) Formation of concateners and sequencing.*

library clones. The sequence is analyzed base by base for ditags punctuated by the anchoring enzyme recognition site. Ditags that pass minimum and maximum length criteria are checked against a running list of ditags, and discarded if previously observed in that project. This constraint of allowing only unique ditags within a SAGE project minimizes PCR-induced bias in the SAGE analysis. The allowed ditags are dissected for their component tags, which are tabulated in separate lists for each project. These lists are used for the second core software function, generation of a report com-

paring the expression profile of a project against others. The report provides normalized or raw tag counts for each project, sorted in order of abundance.

The third key function of the SAGE software is the matching of experimentally observed tags to a species-specific database of reference sequences. The software constructs this database in advance by parsing a text file containing multiple sequences in a standard format. Each sequence is scanned for the 3′-most anchoring enzyme site and its adjacent SAGE tag, which is collected and stored together with abbreviated annotation in a tag-indexed database file. Matching a list of observed tags against this file reveals the corresponding known genes.

Figure 2 summarizes the data analysis procedure for a typical SAGE experiment. As sequencing of SAGE clones is initiated and progresses, the newly obtained sequence is processed with the software to confirm the tag content of the project clones, and to monitor the depth of the tag library. Even at this early stage, matching of the most abundant observed tags to an appropriate reference sequence database could suggest potentially informative functional relationships. When sequencing is complete and the final tag processing is done, various reports can be generated. In a simple study, for example, there might be two samples, representing a test condition and a control. A report comparing the two projects would contain a table of all the observed tags with their normalized abundances in each sample. The tags are then matched to a reference sequence database containing tag entries from known transcripts for the source species.

SAGE tag mapping

SAGE tags that are not readily ascribed to a known, characterized transcript might be further investigated using several gene expression databases accessible via the Internet. These databases have been generated by similarity clustering algorithms applied to large collections of transcript fragments called expressed sequence tags (ESTs). The resulting subsets of ESTs, or clusters, each represent a single transcript or a family of related transcripts. Querying with a SAGE tag sequence (4-nt NlaIII site + 10-nt tag) in the UniGene database (National Center for Biotechnology Information, NCBI; Bethesda, MD, USA)[4–6], or the Sequence Tag Alignment and Consensus Knowledgebase (South African National Bioinformatics Institute, SANBI; Bellville, South Africa)[7] might return EST hits constituting one or more such clusters of homologous sequences. Alternatively, a SAGE tag query in the Human Gene Index (The Institute for Genome Research, TIGR; Rockville, MD, USA)[8], or Sanigene database (SANBI)[7] might return a hit of a consensus sequence representing an entire cluster. As such searches do not take into account the number or location of NlaIII sites in the clusters, hits from these databases must be screened to ascertain which of the represented transcripts contain the query tag at its 3′-most anchoring enzyme site. Further examination of the resulting candidate sequences might provide clues to the function of a potentially novel transcript, extended sequence information, as well as a physical source for the actual clone.

SAGEmap, a public repository of human SAGE data, has recently been made available via the Internet (http://www.ncbi.nlm.nih.gov/SAGE)[9]. Part of the Cancer Genome Anatomy Project at the National Cancer Institute (NCI; Bethesda, MD, USA), this online resource provides tools for viewing and analyzing tag data from a compilation of >40 SAGE libraries. These data have been provided by several laboratories investigating gene expression in human cancer. Web query forms enable the search for expression levels of individual genes or tags, and to view comprehensive comparisons between multiple SAGE libraries.

A key feature of SAGEmap is its systematic approach to mapping SAGE tags to UniGene EST clusters. Rather than simply matching SAGE tags across all NlaIII sites in the sequence clusters, SAGEmap incorporates EST frequency, location, and orientation information, as well as a correction for sequencing error to refine the mapping of tags to genes. This mapping represents a ‘best guess’ for the matching of observed tags to expressed sequences in the public database.

SAGE database

Workers at Genzyme have expanded on the basic functions of the SAGE software in an attempt to streamline the production and analysis of SAGE data, enhance the sharing of project data and annotation between researchers, and link the tags to several external data sources. Tag abundance data and annotations from all the internal SAGE projects (comprising over 3 million tag counts) have therefore been compiled into a centralized relational database. With a common, shared storage location, researchers throughout the organization have access to consistent, up-to-date SAGE data at all times.

The interface to the SAGE database is a custom PC application with features that support flexible, yet detailed, analysis of SAGE data. Researchers can compare transcript profiles from different projects in multiple combinations. In a simple example, expression profiles from several samples of normal tissue can be merged into a ‘virtual profile’, which can then be compared with that of disease tissue samples. Such a comparison highlights any disease-specific differences in gene expression, while reducing sample-specific variations. The investigator can construct additional comparison scenarios by combining and comparing experiments, with the option of saving both analysis schemes

and results for later review.

The database interface provides built-in sorting and filtering functions, permitting the user to apply relative abundance, ratio, or significance cut-offs to the data. In this manner, a full set of tens of thousands of observed tags can be reduced to a more manageable set of several hundred of the most interesting differentially expressed tags from a given group of experiments.

In addition to storing experimental tag data, the SAGE database also serves as the repository for reference sequence information extracted from external sources, such as the GenBank, UniGene and SAGEmap public databases at NCBI. The interface application supports searching and retrieval of this pre-filtered information, and also provides an integrated web-browser for interactively linking tags or reference sequences to the original databases.

Researchers can also add their own custom tag-specific annotations to the database. Thus, investigators conducting independent analyses can conveniently share observations and insights regarding tags of mutual interest. The opportunity for such 'cross-fertilization' maximizes the integration of SAGE data into a research knowledge base.

Additional informatics tools

Finally, Genzyme is implementing a SAGE pipeline that automates the process of producing tag data. The pipeline is a series of utility programs run on a network server to collect raw sequence files, analyze them for tag abundance data, and load the new data into the SAGE database. Automation of these steps increases not only the efficiency of generating new SAGE data, but also the consistency and integrity of the database itself. Continued development of the SAGE database and interface application is focused on enhancing the links between tag data and diverse downstream data sources, such as HTS databases and signaling pathway schemes.

SAGE applications

Pathway elucidation

The applications for high-throughput transcript profiling technologies continue to evolve[9–34]. Underlying these applications is the unprecedented opportunity to view phenotypic responses within the context of global transcriptional changes ongoing in cell populations. Within the foundation of these comprehensive views of transcriptional responses will be distinct pathways and cascades previously unrecognized. It is the identification of these distinct pathways, ultimately forming a network correlating phenotypic response with transcriptional changes, that will enable functional assignment to specific gene products involved in these pathways.

SAGE is heralded as an unbiased and comprehensive high-throughput transcript profiling technology and is therefore ideally suited for novel pathway elucidation. As with any attempt at pathway elucidation utilizing transcript profiling, transcriptional changes caused by a single or a few gene changes in an otherwise isogenic background potentially provides few transcriptional differences and therefore the data output is fairly straightforward to interpret. SAGE has been applied in this way towards elucidating transcriptional manifestations within cells harboring functional or non-functional p53.

Table 1. Genes induced by p53 in the rat fibroblast growth arrest model

Gene	Fold difference (+p53/−p53)[a]
MDM2	3-fold induced
HSP70	5-fold repressed
WAF1	⩾5-fold induced
Cyclin G	25-fold induced
BAX	2-fold induced
CGR11	⩾12-fold induced
CGR19	⩾5-fold induced
EGR1	30-fold induced
Shabin 80	⩾20-fold induced
Shabin 123	⩾5-fold induced

[a]SAGE analysis was used to measure the expression level of each gene in the presence of an active (+) or inactive (−) form of p53. The fold difference in expression represents the ratio of expression of a given gene when p53 is active versus in-

An example of this application was the use of SAGE to elucidate the p53 pathway in a well-characterized rat model system where the end-point phenotype compares growth with growth arrest[19]. Similarly, SAGE was applied in a human system where cells expressing functional p53 resulted in an apoptotic end-point[24]. The growth arrest rat p53-model system demonstrated the utility of SAGE for identifying a large number of previously identified p53-regulated genes (Table 1). In one experiment, SAGE also enabled identification of previously uncharacterized transcripts potentially important in mediating downstream p53 effects.

The human p53 SAGE study was more rewarding in that a large fraction of the p53-induced genes revealed by SAGE could be categorized into a functional family of genes that helped to explain the ultimate apoptotic phenotype exhibited. Eight of the top 14 genes differentially induced in the presence of p53 were previously implicated in either generating or responding to oxidative stress (Table 2), a pathway documented to play a role in apoptosis[24]. As in the rat p53 SAGE study, the human apoptotic system also revealed novel genes highly differentially expressed with respect to p53 status. Having this SAGE-derived immortal p53 database as an archive proved to be extremely useful in subsequent SAGE experiments on γ-

Table 2. Genes induced by p53 in a human epithelial cell apoptosis model

Name	Function or homology	[a]ROS effect
p21	CDK inhibitor	Induced by ROS
[b]*PIG1*	Galectin 7	Enhancer of superoxide production
PIG3	Quinone oxidoreductase homologue	ROS generator
PIG4	Serum amyloid A	Induced by ROS
PIG6	Proline oxidase homologue	Glutathione biosynthesis
PIG7	TNFα-induced mRNA	ROS induces TNFα
PIG8	Etoposide-induced mRNA	A quinone that leads to ROS
PIG12	GST homologue	Induced by ROS

[a]The relationship between the genes induced by p53 and the generation of, or response to, reactive oxygen species (ROS).

[b]p53-induced genes (PIG).

irradiated colorectal cancer cells. SAGE studies on these colorectal cancer cells revealed overlaps with pre-existing p53 differential SAGE data[15]. By pooling SAGE data sets and determining commonalities from different experiments, the *14-3-3 σ* gene was shown to be important in responding to γ-irradiation in a p53-dependent manner.

The utility of applying global transcript profiling for pathway elucidation was further demonstrated by SAGE transcript profiling of cells expressing or not expressing the tumor suppressor gene *APC* (adenomatous polyposis coli), which functions to monitor cell growth in the intestinal mucosa[35]. It was demonstrated prior to this study that *APC* affects cell growth, in part, by altering the activity of the β-catenin/Tcf-4 transcription factor complex[36]. SAGE transcript profiles could define a subsequent step in the *APC* cascade by delineating a transcriptional effect on the *c-Myc* oncogene which, as it turns out, is directly regulated by β-catenin/Tcf-4. Thus, this novel pathway elucidation enabled the intriguing functional convergence of the tumor suppressor *APC* and the oncogene *c-Myc*[37]. Numerous other SAGE studies are currently being performed comparing nearly isogenic systems.

The relatively simple transcript profiles resulting from plus/minus treatment regimens, as just described, have great value in identifying biological function. However, it is a more difficult challenge to use transcript profiling to reveal previously unrecognized or novel pathways important in complex diseases such as cancer. To begin to unravel complex disease pathways, SAGE has been used to elucidate normal versus disease transcriptomes for gastrointestinal cancers[33], non-small cell lung cancer (NSCLC)[16] and breast cancer[21]. Relative to the isogenic SAGE comparisons already described, the disease transcriptomes identified many more differentially regulated genes (Fig. 5). With these complex disease transcriptomes, revelations concerning functional pathways involved in disease progression are not immediately apparent, although several individual genes previously implicated in cancer progres-sion were observed.

The dilemma that has evolved from many large complex disease transcriptome analyses is that further verification efforts are necessary to limit the pool of potential candidate genes. The wealth of possible disease-specific genes that have been revealed by SAGE (Ref. 17) in the cancer arena necessitates a verification platform to aid in limiting the time-consuming functional work downstream from data generation. cDNA microarrays are currently being formulated by utilizing SAGE data to program the target array, which will ultimately be interrogated by additional samples. This approach is more advantageous than using prefabricated cDNA or oligomer arrays where only previously defined cDNAs can be interrogated, omitting novel genes that could be relevant to the specific disease being investigated. SAGE tags that do not correspond to known genes or ESTs, and which therefore could represent novel genes, can be used effectively to amplify additional sequences from cognate cDNAs (Ref. 14 describes this approach). This facilitates the integration of potentially novel cDNAs into the design of SAGE-defined microarrays. This approach of SAGE-based custom micro-arrays has been used to identify genes that were differentially expressed in primary and metastatic breast cancer[21].

The Genzyme SAGE cancer database currently contains profiling data encompassing >3 million normal and disease-specific tags derived from the colon, lung, pancreas, breast, melanoma and prostate (Table 3). While the data set is predominantly biased towards transcriptomes of normal and cancer cells, the database does include transcriptomes of cells and tissues from non-cancerous diseases of the kidney, liver and chondrocytes. The tag collection represents >100 000 tags expressed two or more times in the collection. At least two SAGE libraries were constructed for each normal or disease indication, with each library sequenced to a level of >20 000 SAGE tags (with the exception of the normal pancreas). This database is being used for a variety of applications including defining the minimum transcriptome in normal or tumor-derived human cells, identifying genes that are differentially expressed in a specific tumor type compared with its normal tissue counterpart, and identifying genes that are differentially expressed in a tissue-specific manner[29]. Furthermore,

analysis of SAGE-based transcriptomes can be used to design custom microarrays whose elements represent tissue-, disease- or pathway-specific cDNAs. This application combines the ability of SAGE to comprehensively define the complexity of expression with the higher-throughput, semi-quantitative nature of microarray hybridization. It is clear from sifting through this cancer-specific SAGE database that valuable information can be garnered from a comprehensive data set where patterns of expression can be more readily associated with specific cancer phenotypes. The continuing evolution and compilation of cancer-specific SAGE data will enable further refinements in the design of these cancer-specific microarrays.

Diagnostic applications

The functional elucidation of crucial pathways involved in disease progression often requires large resource commitments of time and money. Relatively rapid utilization of the SAGE archived database without necessarily integrating functional characterization has been demonstrated for the prognostic and diagnostic evaluation of cancer. In the case of squamous NSCLC, several genes observed with SAGE to be differentially induced in squamous cell carcinoma with respect to normal cells were evaluated for their expression in additional cell lines and clinical tumor samples[38]. This verification yielded a gene subset that showed consistent overexpression in squamous lung cancer cells. PGP9.5, a ubiquitin carboxy-terminal hydrolase, was particularly valuable, in that expression of this gene product not only correlates with the presence of lung cancer, but has prognostic value in showing ever-increasing expression as the disease progresses to more advanced stages.

Through the use of SAGE and other transcript profiling technologies, expanded pools of genes or gene products for diagnostic purposes are likely to be identified. SAGE has already shown utility in adding an additional marker to a previous diagnostic screen for pancreatic cancer[34]. In this case, the tissue inhibitor of metalloproteinase (TIMP-1) gene product was discovered by SAGE to be specifically overexpressed in pancreatic tumor samples but not in normal or cancerous colon samples. This TIMP-1 marker, by itself, was inadequate for diagnosis of pancreatic cancer. However, TIMP-1 in conjunction with the previously characterized pancreatic cancer markers, CA-19 and carcinoembryonic antigen (CEA), increased the sensitivity of pancreatic cancer detection by >10% (Ref. 34).

Toxicological profiling

One of the most significant perceived added values of transcription profiling is the enormous potential to reduce costly and time-consuming drug development by assessing drug toxicity early in the evaluation process. It is well established that correlations exist between drug toxicity and the expression of certain genes or gene families (e.g. peroxisome proliferation[39]). Certainly, SAGE and other transcript profiling platforms can decipher gene expression changes associated with these previously defined toxicity-related genes. Importantly, however, many drugs fail in clinical trials because of both the inability to evaluate toxicity early in drug development and simply because the drug might alter gene expression directly or indirectly within biological pathways.

The affected biological pathways might be unknown or

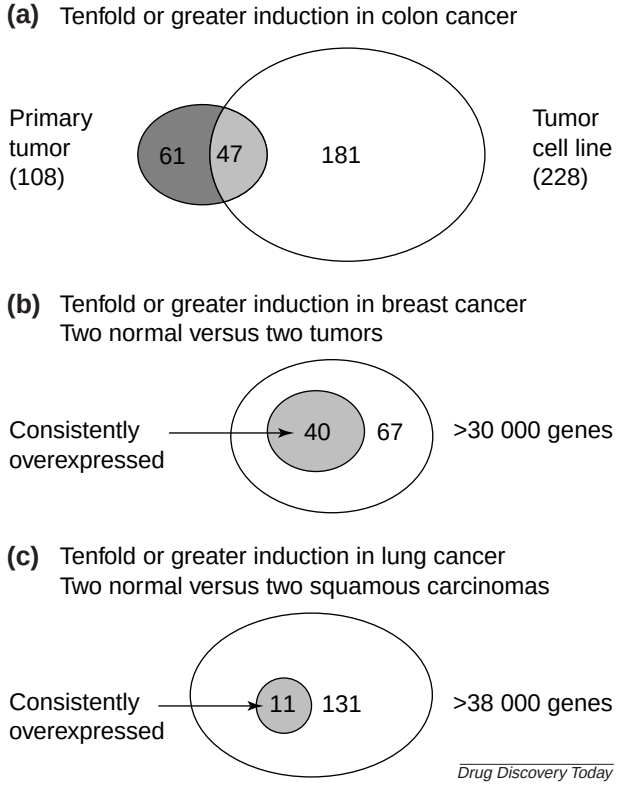


Figure 5. *Genes induced in cancer. (a) A comparison of genes induced at least tenfold above normal in colon cancer as measured using primary tumors and tumor-derived cell lines against normal colon epithelium. (b) Genes induced at least tenfold above normal in breast cancer as measured using cell lines derived from both primary and metastic tumors and compared with normal mammary epithelium. Forty genes were induced in both of the tumor lines and 27 genes were only induced in one of the two tumor cell lines. Over 30 000 unique genes contributed to this analysis. (c) Genes induced at least tenfold above normal in breast cancer as measured using cell lines derived from non-small cell lung cancer tumors and compared with normal lung epithelium. Over 38 000 genes were included in this analysis.*

Table 3. Human transcriptomes

Tissue/cells	[a]Normal tags	[b]Disease tags	Total tags
Breast	108 000 [c](2)	322 000 (8)	430 000
Colon	194 000 (3)	1 176 000 (21)	1 370 000
Lung	113 000 (2)	294 000 (8)	40 000
Pancreas	840	127 000 (4)	128 000
Prostate	142 000 (3)	98 000 (3)	240 000
Melanoma	111 000 (2)	271 000 (10)	382 000
Hemangiopericytomas	–	203 000 (5)	203 000
Kidney	51 000	192 000 (6)	243 000
Liver	53 000	61 000	114 000
Chondrocytes	67 000 (3)	22 000	89 000
Myocytes	98 000 (4)	–	98 000
Totals	938 000 (22)	2 766 000 (67)	3 704 000 (89)

[a]The number of tags sequenced from SAGE libraries made from cells or tissues of normal phenotype.

[b]The number of tags sequenced from SAGE libraries made from cells or tissues of disease phenotype.

[c]The number of independent SAGE libraries sequenced. For those not specified, only one library was made and sequenced.

known but not anticipated for certain clinical indications. Thus, even though a drug might not be toxic *per se*, its effects on certain biological pathways could prevent or limit its therapeutic use. Awareness of the effects early in the drug development process can provide a more rigorous basis for advancing candidates that do not elicit these effects or for optimizing the chemistry of lead compounds to eliminate or minimize these effects. In either case, gene expression profiling can enable the monitoring of direct and indirect effects of a drug candidate on biochemical pathways. Fulfilling the enormous potential for the molecular classification of drug effects relies on accurate and comprehensive analytical approaches. This is a challenge with certain transcript profiling technologies but particularly difficult when evaluating model organism transcriptomes. As drug toxicity and biological pathway elucidation is usually assessed initially in animals such as the mouse and rat, the ability to profile gene expression changes in these model systems is vital for the application of transcript profiling to drug toxicology assessment. The limited genomic information currently available for mouse and rat model organisms compared with humans restricts the utility of technologies that rely exclusively on prior transcript identification. As SAGE is aided by these data sets but does not require this prior knowledge, it has the power to extract much more information from these model organisms for toxicological studies.

Therapeutic development

Disease management facilitated by comprehensive and sensitive diagnostic tools and early drug evaluation utilizing toxicological transcript profiling certainly are ventures that ultimately reduce cost and potentially increase efficacy of lead drug candidates. A seemingly greater challenge is to use SAGE transcript profiles to discover or facilitate the discovery of novel therapeutics. This challenge is necessarily greater because of the wealth of information obtained and the functional characterization required in evaluating a good therapeutic lead. Thus, the real challenge is to design clever data analysis strategies in which the results reveal valid therapeutic candidates based on functional characterization.

In this regard, SAGE has been employed to profile cells and tissues that either exhibit or lack a specific phenotype of therapeutic interest. By initially stratifying samples based on phenotype, expression profiles can be performed on the most appropriate samples (extreme phenotypes, intermediate phenotypes and different genetic backgrounds). This approach is most powerful with a quantitative phenotypic assay and access to sufficient numbers of different cell and tissue samples to adequately cover the phenotypic spectrum. An expression comparison between transcriptomes derived from cells/tissues of known phenotypes rapidly reveals the strongest candidates from a large pool of differentially expressed genes, while enhancing the likelihood of being able to correlate phenotype to genotype at the cDNA level. This strategy is currently being employed to identify several expressed genes that can be tested for their ability to confer a phenotype of therapeutic value.

Finally, providing efficacious therapies requires the ability to administer the therapy to the appropriate location in the body. This attribute is particularly important in the case of gene therapy to maximize efficacy and safety by limiting expression to defined cell or tissue types. The ability of SAGE to define specific transcriptomes will aid in the development of gene therapies whereby cell or tissue-specific promoters and genes can be utilized to appropriately express and deliver a given therapy.

Conclusions

The identification of more therapeutic targets of high biological relevance is a necessary step in improving the drug development process. SAGE provides a robust platform for discerning relevant drug targets in a comprehensive fashion, based on differential gene expression in diseased cells or tissues. SAGE alone or in combination with proteomic

approaches, such as protein expression profiling and protein network screens, will accelerate the identification of high-quality drug targets and catalyze the development of high-throughput screens to reveal the next generation of therapeutic products.

REFERENCES

1 Drews, J. (1996) Genomic sciences and the medicine of tomorrow. *Nat. Biotechnol.* 14, 1516–1518

2 Drews, J. and Ryser, S. (1997) The role of innovation in drug development. *Nat. Biotechnol.* 15, 1318–1319

3 Velculescu, V.E. *et al.* (1995) Serial analysis of gene expression. *Science* 270, 484–487

4 Schuler, G.D. (1997) Pieces of the puzzle: expressed sequence tags and the catalog of human genes. *J. Mol. Med.* 75, 694–698

5 Schuler, G.D. *et al.* (1996) A gene map of the human genome. *Science* 274, 540–546

6 Boguski, M.S. and Schuler, G.D. (1995) ESTablishing a human transcript map. *Nat. Genet.* 10, 369–371

7 Miller, R.T. *et al.* (1999) A comprehensive approach to clustering of expressed human gene sequence: the sequence tag alignment and consensus knowledge base. *Genome Res.* 9, 1143–1155

8 Quackenbush, J. *et al.* (2000) The TIGR Gene Indices: reconstruction and representation of expressed gene sequences. *Nucleic Acids Res.* 28, 141–145

9 Lal, A. *et al.* (1999) A public database for gene expression in human cancers. *Cancer Res.* 59, 5403–5407

10 Boers, W. *et al.* (1999) Identification of expressed genes from transformed human stellate cells associated with liver fibrosis. *Cells Hepatic Sinusoid* 7, 171–172

11 Datson, N.A. *et al.* (1999) MicroSAGE: a modified procedure for serial analysis of gene expression in limited amounts of tissue. *Nucleic Acids Res.* 27, 1300–1307

12 de Waard, V. *et al.* (1999) Serial analysis of gene expression to assess the endothelial cell response to an atherogenic stimulus. *Gene* 226, 1–8

13 Hashimoto, S. *et al.* (1999) Serial analysis of gene expression in human monocyte-derived dendritic cells. *Blood* 94, 845–852

14 Hashimoto, S. *et al.* (1999) Serial analysis of gene expression in human monocytes and macrophages. *Blood* 94, 837–844

15 Hermeking, H. *et al.* (1997) *14-3-3* σ is a p53-regulated inhibitor of G2/M progression. *Mol. Cell* 1, 3–11

16 Hibi, K. *et al.* (1998) Serial analysis of gene expression in non-small cell lung cancer. *Cancer Res.* 58, 5690–5694

17 Inoue, H. *et al.* (1999) Serial analysis of gene expression in a microglial cell line. *Glia* 28, 265–271

18 Kal, A.J. *et al.* (1999) Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources. *Mol. Biol. Cell* 10, 1859–1872

19 Madden, S.L. *et al.* (1997) SAGE transcript profiles for p53-dependent growth regulation. *Oncogene* 15, 1079–1085

20 Matsumura, H. *et al.* (1999) Transcript profiling in rice (*Oryza sativa* L.) seedlings using serial analysis of gene expression (SAGE). *Plant J.* 20, 719–726

21 Nacht, M. *et al.* (1999) Combining serial analysis of gene expression and array technologies to identify genes differentially expressed in breast cancer. *Cancer Res.* 59, 5464–5470

22 Neilson, L. *et al.* (2000) Molecular phenotype of the human oocyte by PCR-SAGE. *Genomics* 63, 13–24

23 Peters, D.G. *et al.* (1999) Comprehensive transcript analysis in small quantities of mRNA by SAGE–lite. *Nucleic Acids Res.* 27, e39

24 Polyak, K. *et al.* (1997) A model for p53-induced apoptosis. *Nature* 389, 300–305

25 Rothstein, J.L. *et al.* (1992) Gene expression during preimplantation mouse development. *Genes Dev.* 6, 1190–1201

26 Ryo, A. *et al.* (1999) Serial analysis of gene expression in HIV-1-infected T cell lines. *FEBS Lett.* 462, 182–186

27 van den Berg, A. *et al.* (1999) High expression of the CC chemokine TARC in Reed–Sternberg cells. A possible explanation for the characteristic T-cell infiltratein Hodgkin's lymphoma. *Am. J. Pathol.* 154, 1685–1691

28 Velculescu, V.E. *et al.* (1997) Characterization of the yeast transcriptome. *Cell* 88, 243–251

29 Velculescu, V.E. *et al.* (1999) Analysis of human transcriptomes. *Nat. Genet.* 23, 387–388

30 Virlon, B. *et al.* (1999) Serial microanalysis of renal transcriptomes. *Proc. Natl. Acad. Sci. U. S. A.* 96, 15286–15291

31 Welle, S. *et al.* (1999) Inventory of high-abundance mRNAs in skeletal muscle of normal men. *Genome Res.* 9, 506–513

32 Yamashita, T. *et al.* (2000) Comprehensive gene expression profile of a normal human liver. *Biochem. Biophys. Res. Commun.* 269, 110–116

33 Zhang, L. *et al.* (1997) Gene expression profiles in normal and cancer cells. *Science* 276, 1268–1272

34 Zhou, W. *et al.* (1998) Identifying markers for pancreatic cancer by gene expression analysis. *Cancer Epidemiol. Biomarkers Prev.* 7, 109–112

35 Sparks, A.B. *et al.* (1998) Mutational analysis of the APC/β-catenin/Tcf pathway in colorectal cancer. *Cancer Res.* 58, 1130–1134

36 Morin, P.J. *et al.* (1997) Activation of the β-catenin–Tcf signaling in colon cancer by mutations in β-catenin or APC. *Science* 275, 1787–1790

37 He, T.C. *et al.* (1998) Identification of c-MYC as a target of the APC pathway. *Science* 281, 1509–1512

38 Hibi, K. *et al.* (1999) PGP9.5 as a candidate tumor marker for non-small-cell lung cancer. *Am. J. Pathol.* 155, 711–715

39 Masters, C.J. (1998) On the role of the peroxisome in the metabolism of drugs and xenobiotics. *Biochem. Pharmacol.* 56, 667–673