

What can digital transcript profiling reveal about human cancers?

J.M. Cerutti¹,
G.J. Riggins² and
S.J. de Souza³

¹Laboratório de Endocrinologia Molecular, Divisão de Endocrinologia, Departamento de Medicina, Universidade Federal de São Paulo, São Paulo, SP, Brasil

²Duke University Medical Center, Durham, NC, USA

³Instituto Ludwig de Pesquisa sobre o Câncer, São Paulo, SP, Brasil

Abstract

Important biological and clinical features of malignancy are reflected in its transcript pattern. Recent advances in gene expression technology and informatics have provided a powerful new means to obtain and interpret these expression patterns. A comprehensive approach to expression profiling is serial analysis of gene expression (SAGE), which provides digital information on transcript levels. SAGE works by counting transcripts and storing these digital values electronically, providing absolute gene expression levels that make historical comparisons possible. SAGE produces a comprehensive profile of gene expression and can be used to search for candidate tumor markers or antigens in a limited number of samples. The Cancer Genome Anatomy Project has created a SAGE database of human gene expression levels for many different tumors and normal reference tissues and provides online tools for viewing, comparing, and downloading expression profiles. Digital expression profiling using SAGE and informatics have been useful for identifying genes that have a role in tumor invasion and other aspects of tumor progression.

Key words

- Serial analysis of gene expression
- SAGE
- Tumor antigens
- Gene expression profiling
- Cancer

Correspondence

S.J. de Souza
Instituto Ludwig de Pesquisa
sobre o Câncer
01509-010 São Paulo, SP
Brasil
E-mail:
sandro@compbio.ludwig.org.br

Presented at SIMEC 2002
(International Symposium
on Extracellular Matrix),
Angra dos Reis, RJ, Brazil,
October 7-10, 2002.

Publication supported by FAPESP.

Received December 2, 2002

Accepted February 20, 2003

Introduction

We are now very familiar with the premise that cancer is characterized by acquired genetic alterations that progressively alter genes involved in growth, apoptosis and/or DNA stability (1). It is also known that some of these genetic alterations seem to be specific to one or a very few subsets of cancer cell types, whereas others are frequently present in several types of cancer cells.

In the course of the last decade, there have been many efforts to determine if the genes altered in cancer are also useful as diagnosis and classification markers. Many of these cancer-causing genes have been exploited for improved clinical diagnosis and

to define potential therapeutic targets. Even though the concept of acquired genomic DNA alterations in malignancy has revolutionized cancer research, the translation of this information to clinically useful targets has been largely based on evaluating one candidate gene at a time. Advances in the fields of genomics and biotechnology, such as high-throughput mutational analysis, large-scale expression analysis and DNA sequencing, have considerably improved the quantity, quality and accessibility of the molecular information. This technical revolution is accelerating the process of identifying the genes of diagnostic, prognostic or therapeutic significance in cancer.

Cancer is indeed a complex, dynamic and

progressive process that is initiated by genetic alterations. It also leads to even more complex changes in the pattern of gene expression as the altered cells progress from normal to malignant. To further increase the challenges faced when interpreting these transcript patterns, expressed genes found within these various malignant "transcriptomes" are further modulated by genetic background, environment and host physiology. However, the pattern of expressed genes can still provide useful clues to better understand tumor development and progression, if experiments are carefully conceived and controlled. In addition to the biological insight that might be derived from this pattern, there are also practical applications for simply identifying which genes are expressed in cancer but not in the corresponding normal tissue. This has the practical advantage of identifying candidate tumor antigens. Gene expression profiling also has the potential to better define the molecular classes of tumors, previously unrecognized by histology alone. The RNA expression pattern may predict the phenotypic behavior of such cells more accurately than traditional histological approaches and possibly help to further differentiate between aggressive and nonaggressive tumors. Gene expression profiling has also been useful to identify genes that have a central role in the response to specific environmental conditions such as hypoxia, hormonal stimuli and drugs and may perhaps predict response to therapy. This approach may also be helpful to define genes in a specific pathway and to elucidate functions of characterized and uncharacterized genes.

The development of technologies that allow a large number of transcripts to be analyzed simultaneously has made it possible to determine the molecular profile of normal and disease cells in a quantitative fashion. In addition to the comprehensive functional investigation of one gene, cancer research can now identify and quantify the complex expression patterns that occur dur-

ing tumor development and progression.

However, the analysis of large-scale expression profiling is not a trivial task and requires the ability to compile data in a database and effectively interpret this information. Sophisticated computational and statistical approaches, either new or derived from approaches formerly applied to the physical sciences, are now required to interpret complex datasets. Other bioinformatics approaches are necessary to draw on the vast and growing archive of information available through public databases or the biomedical literature. Correlating expression levels in malignant cells with the information derived from the recently sequenced human genome is a particularly important example. Finally, this information has to be made readily available in a user-friendly format, so that scientists can concentrate on making progress on developing better insights and treatments.

Although such techniques are far from reaching their full potential, some important applications of this technology are already transforming cancer research. The ability to assay gene expression levels on a large scale holds the promise of revealing a much more complete picture of the molecular interactions within the malignant cell. However, these data are only a first step towards achieving a better understanding, diagnosis and treatment of cancer.

Transcript profiling

Several methods have been developed to monitor gene expression differences between two samples. The first techniques widely used to find differentially expressed transcripts were subtractive hybridization and differential display (2,3). Both techniques identify transcripts but they do not have the capacity to assay multiple samples as is possible with oligonucleotide arrays and cDNA arrays (4,5), nor do they provide an in-depth transcriptome characterization of sequenc-

ing-based techniques such as cDNA library sequencing and serial analysis of gene expression (SAGE). For this reason, DNA arrays and SAGE are currently the techniques most widely used for determining the relative and absolute abundance of transcript levels through several different stages of cancer and under cell response to different physiological conditions or environmental stimuli. This review focuses on the lesser known SAGE technology, on how it has been used to better understand cancer, and on its potential for future applications.

Expressed sequenced tags

Large-scale sequencing of cDNA libraries was first proposed as a rapid means to access transcribed regions from the human genome (6). Random transcribed sequences generated by cDNA library sequencing are known as expressed sequence tags (ESTs). The Merck/Washington University EST Project made one of the first large-scale efforts to disseminate EST sequence data (7). The Cancer Genome Anatomy Project (8) succeeded this effort with its Tumor Gene Index, contributing over one million ESTs from normal, premalignant and malignant cells. The Brazilian Genome Project has also made a major contribution by sequencing internal cDNA fragments that are complementary to existing data through its human cancer Orestes project (9). The data from these projects have revealed which tissues express which transcripts and have greatly reduced the time and effort necessary for many gene-cloning projects, but represent a laborious approach when simply used to define gene expression levels. A key advantage is that these data are free and easily accessed. Cancer Genome Anatomy Project-generated sequence data are made immediately available through the Cancer Genome Anatomy Project web site or through the National Center of Biotechnological Information's (NCBI) sequence resources such as

GenBank's dbEST (<http://www.ncbi.nlm.nih.gov/dbEST/index.html>) database or as part of UniGene sequence clusters (<http://www.ncbi.nlm.nih.gov/UniGene>). The main disadvantages are that the individual experimenter cannot practically generate his own EST data and that the level of detection is low, since often only a few thousand transcripts are assayed for each tissue or cell type, out of the tens of thousands expressed.

The EST data serve a dual purpose of determining coding nucleic acid sequences and revealing the presence of the sequenced transcripts in the RNA used for library construction. Although the presence of a transcript in a particular library can be revealing, the absolute level of a gene expression is lost when cDNA libraries are normalized or subtracted.

Serial analysis of gene expression

SAGE (10) uses automated DNA sequencing to efficiently count large numbers of mRNA transcripts from a small population of cells (Figure 1). SAGE increases the number of genes that can be counted per sequencing reaction, compared to cDNA library sequencing, by minimizing the portion of the transcript sequenced. The method works by cloning and sequencing a 10-base pair (bp) portion of the cDNA at a defined position near the 3' end of the transcript. This 10-bp portion, normally next to the last *Nla*III restriction site, is known as the transcript's 'tag'. SAGE tags are ligated and cloned end-to-end in a sequencing vector, allowing the 'serial' analysis of multiple transcripts. The number of times a particular tag is observed in a tag population made from one mRNA sample (SAGE library) is used to determine transcript abundance. The SAGE transcript profile from various types of cells can be archived on a computer database and electronically compared to find statistically significant differences in gene expression between cell types. To provide tag-to-gene links,

seven sources of cDNA were assembled (11).

Since SAGE counts transcripts by sequencing and avoids the errors inherent in hybridization-based assays, it is often regarded as a very accurate means for expression profiling. SAGE transcript levels are expressed as a fraction of the total transcripts counted, not relative to another experiment or a housekeeping gene, avoiding error-prone normalization between experiments. In addition, SAGE determines expression levels directly from an RNA sample and it is not necessary to have a gene-specific fragment of DNA arrayed to assay each gene. This allows SAGE to identify genes that are not included in an array and avoids the infrastructure necessary to create and read large DNA arrays.

The number of samples that can be processed using SAGE is small compared to DNA arrays; it takes two weeks or more of skilled labor to construct a SAGE library. However, when an in-depth and quantitative profile is desired for a small number of samples the extra work involved in creating a SAGE library can be justified. The strength of SAGE is its use in determining differentially expressed transcripts in well-controlled experimental systems (12,13).

A detailed protocol can be obtained through the SAGE home page from the Johns Hopkins Oncology Center (<http://www.sagenet.org>). The technology is patented by Johns Hopkins University and licensed to Genzyme Molecular Oncology (Framingham, MA, USA) but freely available to academia and nonprofit organizations for research purposes.

One advantage of SAGE is that public reference data are available. In order to provide a more efficient means for archiving quantitative expression profiles, the Cancer Genome Anatomy Project adopted SAGE and has sponsored the Cancer Genome Anatomy Project SAGE Project since 1998 (14-16). Over 5 million transcript tags from more than 100 human cell types are posted at

NCBI SAGEmap web site (<http://www.ncbi.nlm.nih.gov/SAGE>). Recently, the Cancer Genome Anatomy Project SAGE Project created a web site for analysis and presentation of SAGE data, including new informatics tools for the analysis of data. This large archive of SAGE data is viewed in an anatomical context by gene or by comparing profiles online at SAGE Genie (<http://cgap.nci.nih.gov/SAGE>) (11) (Figure 2).

Confirmation approaches

After a gene expression profile has been obtained on a set of RNA samples the expression differences need to be confirmed and it is often useful to determine if the observation is repeatable in independent samples. Normally a small set of interesting genes has been identified using DNA arrays or SAGE, but several different techniques are more efficient for assaying this smaller set of interesting genes. In addition, each gene expression technique has inherent errors and an independent method is required for validating the original expression levels.

Real-time polymerase chain reaction

Even though Northern blotting has been the gold standard for gene expression analysis for many years, real-time PCR, also called "quantitative" or "fluorescent" PCR or "kinetic RT-PCR", has gained popularity for rapid follow-up and confirmation of profiling data. Expression determination by real-time PCR is based on continuous fluorescent monitoring of PCR products from a cDNA template (17,18). Under the right conditions, the number of cycles required to PCR amplify a product to a certain level is directly proportional to the amount as starting template. There are a variety of methods for detecting the accumulation of PCR products during real-time PCR. A fluorescent DNA indicator, such as SYBR green or ethidium bromide, is included in each PCR, so that the

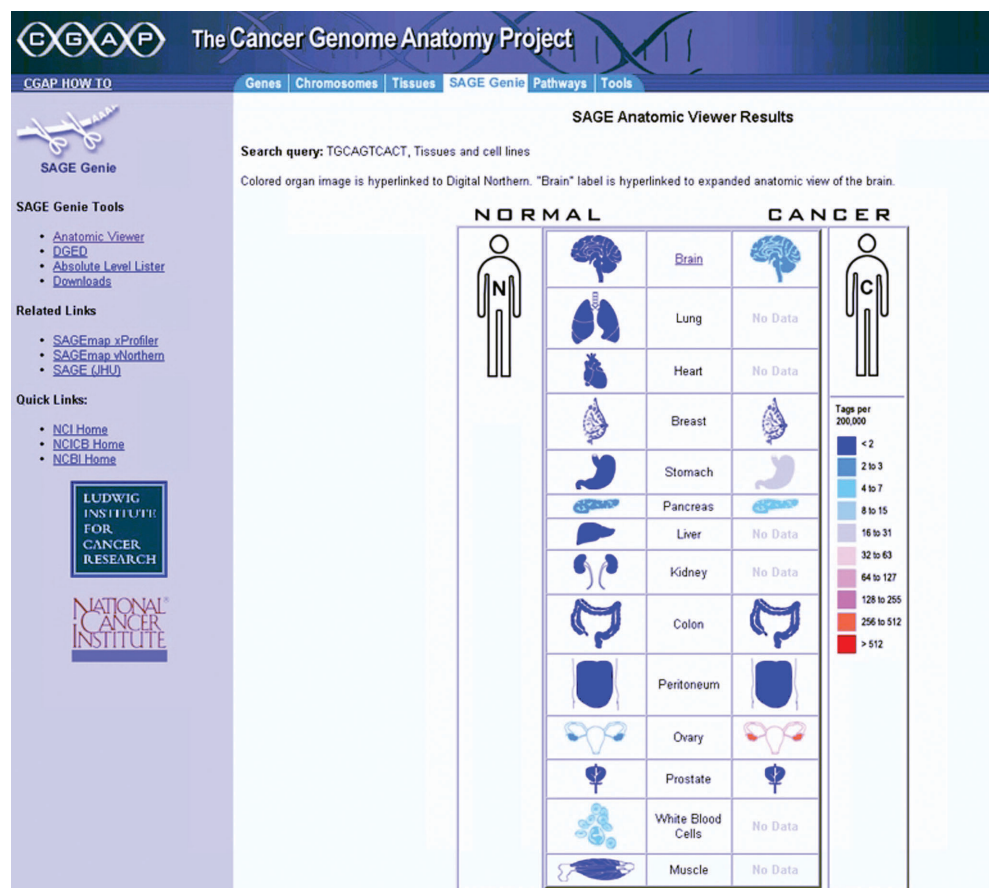
product accumulation can be monitored at each amplification cycle by a kinetic thermal cycler. Alternatively, to increase sensitivity and specificity of PCR product detection, additional oligonucleotide can be employed in the assays that hybridize to an internal portion of the PCR product (TaqMan Assay, PE Biosystems; Hybridization Probes, Roche and Molecular Beacons, Stratagene). Real-time PCR allows for a quick and low-cost assessment of the expression pattern of several genes in many tumors and can be automated. However, real-time PCR data must be interpreted with extreme caution since there are several sources of error inherent in any PCR-based technique.

Immunohistochemistry

To look for protein levels, a Western blot or immunohistochemistry are reliable meth-

ods for confirming expression changes. This approach is advantageous, in particular when the endpoint is knowledge of protein levels rather than mRNA levels. However, to look at protein levels of many samples simultaneously a tissue microarray system has been developed (19,20). This system permits up to one thousand small tissue samples obtained with a narrow gauge biopsy needle to be arrayed in a single block of tissue. This block of tissue can then be used to produce hundreds of slides that can be probed by immunohistochemistry. In this way a standard set of the same samples can be probed for expression levels of many different genes. A digital imaging system is used to record and read the data. The results must also be scored in some fashion by signal intensity, done manually at this point in technology development. Finally, a good antibody is needed for each gene of interest that will

Figure 2. Serial analysis of gene expression (SAGE) anatomic viewer results (<http://cgap.nci.nih.gov/SAGE>). The gene expression profile of matrix metalloproteinase 1 (GenBank # NM_002421) is shown in various cells by their anatomic origin. MMP1 is highly expressed in ovary cancer cells when compared to the normal cells.



work in the normally available formalin-fixed tissue. This approach has the potential to be able to calculate gene expression correlations with a vast archive of preserved tumor material.

Bioinformatics and statistics

The huge amount of data produced by the large-scale approaches described above poses a significant problem for those trying to extract useful information. Today, it is almost impossible to approach these datasets without proper use of computational tools, either locally or in a remote site. There are two groups of researchers that use bioinformatics tools on a daily basis: a larger group composed of molecular biologists and biochemists, among others, who use internet sites of interest where services are provided, and a second smaller group that is responsible for the development of their own tools, used for their own research but sometimes made available to the community. Frequently, this second group is responsible for the development of sites that offer bioinformatics tools to the first group. These two groups differ dramatically regarding the methodology they use and the expertise they have. The former group is composed only of biologists and has very limited computational skills. Biologists, computer scientists, and mathematicians who have a very strong expertise in computational science and programming compose the latter group.

Programming is a crucial part of the routine of a bioinformatics laboratory. The most used programming language is PERL (<http://www.perl.org>). The wide use of PERL in bioinformatics is mainly due to some features of its structure. Since PERL was developed to help system administrators in their daily tasks, the language deals very well with “strings”, which makes it highly appropriate for analyses involving DNA and protein sequences.

Relational databases are also very important and used in almost every task in bioinformatics.

There are several options available but the most used is MySQL (<http://www.mysql.com>), which is relatively simple and is available as an open source database. An intriguing example of a relational database is AceDB (<http://www.acedb.org/>). This relational database was developed exclusively for the *Caenorhabditis elegans* genome project and since then has been widely used by the bioinformatics community.

SAGE Genie (11) is a nice example of how bioinformatics can have an impact on a specific field. The raw data used in SAGE Genie are publicly available and have been used for other initiatives. SAGE Genie, through its computational resources, brings a new perspective to the problem of tag-to-gene and gene-to-tag assignments by scoring different databases according to the representation of the 3' most SAGE tag. In addition, it incorporates data from EST databases to identify transcript variants that would generate a different 3' most SAGE tag.

Statistical tools have been extensively used in analysis involving expression profiling especially of microarrays. The identification of patterns of gene expression and the grouping of genes based on gene expression classes requires a sophisticated statistical analysis. Several methods have been used for the clustering of gene expression data including hierarchical clustering, mutual information and self-organizing maps. In many ways, the computational analysis of gene expression resembles the computational approaches adopted for phylogenetic studies. In both cases, it is almost impossible to find the “best” approach and the use of multiple techniques is ideal to explore different aspects of the data.

A crucial problem nowadays is how to integrate different types of data in a searchable database. This is especially critical in expression profiling studies since heterogeneous kinds of information are available, like EST, SAGE, microarrays and all the proteomic arsenal of techniques. We are still

in a very early stage of defining formats and nomenclature and ahead of us lies the most important challenge that is to integrate all the data with biology and medicine for the development of testable hypotheses.

Advances in cancer research using serial analysis of gene expression

The ability to evaluate the expression pattern of thousands of genes in a quantitative fashion, without prior sequence information, is one of the most attractive features of SAGE. In the past few years, SAGE analysis has been performed in patients with brain, breast, colon, pancreatic, lung, bladder and ovarian cancers and has successfully located new oncogenes, candidate tumor suppressor genes, invasion-related genes, growth-controlling genes, hypoxia-induced genes, and tumor markers. Some of these applications are described below.

Colon cancer

The first application of SAGE to human tissues was to a colon cancer (21). Comparing colon tumors to normal colon epithelium showed that less than 1.5% of the transcripts were differentially expressed. Many genes elevated in colon cancer represented products known to be involved in growth and proliferation, while genes found in normal colon were often related to differentiation. SAGE was used more recently to locate candidate biomarkers for metastasis in colon cancer (22).

Ovarian cancer

Ovarian cancer treatment would benefit from early detection markers, since most ovarian cancers have metastasized prior to detection. SAGE analyzed a total of 385,000 transcripts from ten different ovarian libraries with the purpose of discovering ovarian cancer markers (23). From these data, tran-

scripts were identified that were high in all three primary ovarian cancers and low in all three nonmalignant specimens. A total of 27 genes were identified that met these criteria and that were overexpressed more than 10-fold in ovarian tumors. Interestingly, a majority of those genes were predicted to encode membrane or secreted proteins, making them candidates for biomarkers for tumor targeting. Many of these secreted genes encoded protease inhibitors.

Brain cancers

SAGE has been used to study the most common adult malignant brain tumor, glioblastoma multiforme (GBM). The first SAGE analysis of GBM compared over 200,000 transcript tags from primary GBM and normal brain cortex (14). Approximately 1% of the genes detected were differentially expressed and included angiogenesis factors such as vascular endothelial growth factor (EGF), cell cycle regulators and transcription factors. These data were also used by the Cancer Genome Anatomy Project to help start the public SAGEmap database and are available online at this site. Cancer-induced genes mined from these data were further tested using real-time PCR and Western and Northern blotting to see if candidate tumor markers could be identified (18). Most of the tumor overexpressed genes predicted by SAGE could be confirmed in a subset of glioblastomas. In general, a particular antigen was only highly expressed at most in about one third of the GBM tested, probably due to the molecular heterogeneity of this cancer. However, in combination, 75% of the tumors had at least one antigen that was strongly expressed, and not present in a panel of normal neural tissues. Two antigens were located that coded for cell surface proteins, and may be useful for targeting gliomas with antibody-based therapy.

Brain tumors other than GBM have been studied by expression profiling. SAGE has

also been used to analyze medulloblastomas, the major malignant pediatric brain tumor (24). Detailed SAGE expression profiles are also available for medulloblastomas and a variety of gliomas at the Cancer Genome Anatomy Project SAGE Genie web site (11).

Hypoxic malignant cells

Brain tumors have one of the highest rates of new blood vessel growth of any cancer and one of the stimuli of new vessel growth is hypoxia. To obtain a blood supply and protect against cellular damage and death, oxygen-deprived tumor cells alter gene expression, resulting in resistance to therapy. SAGE was used to locate genes specific for hypoxic cells that could contribute to angiogenesis or other pathological effects of tumor hypoxia. A human GBM cell line was used as a model to compare the expression profile of hypoxic glioblastoma cells to the same cells grown under normal oxygen conditions (25). Ten new genes, not previously known to be hypoxia-responsive were located, including an angiopoietin-related gene, *Angiopoietin-Like 4*. Transcription from these genes, in general, was elevated by hypoxia in other types of malignancy, and the transcript or protein was found specifically for some of the genes in hypoxic regions of solid tumors.

Tumor vascular endothelium

Endothelial cells provide the blood supply and support the critical growth of solid tumors. Targeting tumor antigens located in tumor endothelial cells may provide a strategy for antitumor therapy (26-28). SAGE was used to identify genes differentially expressed between the endothelial cells from either normal colon or colon adenocarcinoma (29). The study detected 79 different genes differentially expressed between these tissues, including 46 that were 10-fold or more elevated in tumor-associated endothelial

cells. Of the top 25 tags more differentially expressed, six were previously recognized as markers of angiogenic vessels and at least seven encoded proteins involved in extracellular matrix formation or remodeling. These matrix-related processes are likely to be crucial to the growth of new vessels. In addition, 14 SAGE tags elevated in the tumor corresponded to novel, noncategorized genes. To validate the expression pattern of these genes, we focused on nine genes that were named tumor endothelial marker (TEM), and designated TEM-1 to TEM-9. On the basis of these results, it was suggested that endothelium growing in a tumor is more like developing endothelium, and that these differences may be clinically relevant. Further experiments confirmed the tumor endothelium-specific expression of these genes, not only for colorectal tumors but also for other major tumor types. These TEM or other genes identified in this study may become targets of antiangiogenic therapies.

Tumor invasion and the extracellular matrix

EGF receptor (EGFR) and EGFRvIII have been implicated in invasion and the higher virulence of tumors. In brain, the expression of EGFRvIII mutant protein enhances the malignant phenotype of gliomas *in vivo* by increasing cell proliferation and decreasing cell death (30). To better understand the role of EGFRvIII in tumor progression, we looked for downstream transcriptional targets by SAGE and DNA array analysis (31) using a glioblastoma cell line as a model. Thirty-eight genes for which EGFRvIII elevated transcript levels were identified. The highly expressed genes included extracellular matrix components, metalloproteases, collagen, and a serine protease. The blockade of EGFR showed that the transcript targets were inhibited in a concentration- and time-dependent manner. The targets of EGFRvIII identified in our study provide insight into the molecular mechanism of EGFRvIII-enhanced

invasion and are potential tumor markers for the screening of drugs for EGFR inhibition.

Cancer-related pathways

SAGE was used to identify many genes whose expression is believed to mediate p53-induced apoptosis (12). Many of these genes were novel and were predicted to encode proteins involved in oxidative stress, thus providing a new paradigm for the mechanism of p53-mediated apoptosis. Similarly, SAGE was used to identify downstream targets of the APC/ β -catenin pathway, a pathway activated in the vast majority of colon cancers (13).

Using a different approach, estrogen-responsive breast cancer cells were treated with estrogen and analyzed by SAGE for expression changes leading to the identification of many possibly useful estrogen-regulated genes (32).

There is no doubt that gene profiling techniques will play a major role in the dissection of the myriad of molecular pathways important in human cancer. The examples given above represent a small fraction of the efforts that have already been dedicated to this goal.

Future directions

SAGE is a useful method for profiling as many of the expressed transcripts in a cell population as is currently possible within a reasonable amount of time. It provides one of the best means to obtain a quantitative profile of expressed transcripts present in a particular tissue, but the technique is time consuming and laborious. There is a large archive of public SAGE data that can be readily accessed (11) and used to help build a database of information necessary to address a particular question. Data obtained by SAGE not only improve our understanding of tumor development and progression but also might be helpful to better understand

growth regulatory pathways and to identify new diagnostic and prognostic markers that, either alone or in combination, can improve the accuracy of cancer diagnosis and can be used as a potential target for drug therapy.

Improved bioinformatics and computational methods allow the data to be queried more easily, but much progress is still necessary to be able to integrate SAGE and other sources of molecular information in a meaningful standard format.

Validation of candidate biomarkers at the RNA level is now much quicker with the use of real-time PCR techniques. The application of *in situ* hybridization or immunohistochemistry can be used to determine if all cells within a tumor are expressing the marker - or if there is some small population of normal cells that highly expresses the gene of interest. When it is necessary to screen large sample sets for protein levels, immunohistochemistry using tissue microarrays can provide a rapid approach (19). Various improvements in proteomic technology may also eventually provide a means to assay proteins at a level as comprehensive as currently available for mRNA (33).

A general conclusion that can be drawn from gene expression profiling of cancer is that tumors, even with identical histopathology, are highly heterogeneous at the expression level. This makes it challenging, but still possible, to classify tumors at the molecular level. It is also difficult to locate tumor-specific markers, and a combination of markers or therapeutic targets will probably be necessary for what we now call a single tumor type based on histology.

The rate-limiting step for tumor marker application or discovery is still the work required to show that the marker will be clinically useful. It is therefore important that the best candidate markers or antigens can be predicted with some degree of accuracy from gene expression data. It still remains to be seen if the candidate markers or antigens discovered initially by SAGE will

produce useful clinical tests or therapies. Although this process will take several years, it seems appropriate to use the most comprehensive data sets possible and careful validation of the limited number of candidates prior to embarking on the laborious task of further developing a tumor-specific gene for clinical use.

References

- Vogelstein B & Kinzler KW (1993). The multistep nature of cancer. *Trends in Genetics*, 9: 138-141.
- Swendeman SL & La Quaglia MP (1996). cDNA subtraction hybridization: a review and an application to neuroblastoma. *Seminars in Pediatric Surgery*, 5: 149-154.
- Sagerstrom CG, Sun BI & Sive HL (1997). Subtractive cloning: past, present, and future. *Annual Review of Biochemistry*, 66: 751-783.
- Gress TM, Hoheisel JD, Lennon GG, Zehetner G & Lehrach H (1992). Hybridization fingerprinting of high-density cDNA-library arrays with cDNA pools derived from whole tissues. *Mammalian Genome*, 3: 609-619.
- Augenlicht LH, Wahrman MZ, Halsey H, Anderson L, Taylor J & Lipkin M (1987). Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate *in vitro*. *Cancer Research*, 47: 6017-6021.
- Adams MD, Soares MB, Kerlavage AR, Fields C & Venter JC (1993). Rapid cDNA sequencing (expressed sequence tags) from a directionally cloned human infant brain cDNA library. *Nature Genetics*, 4: 373-380.
- Williamson AR (1999). The Merck Gene Index Project. *Drug Discovery Today*, 4: 115-122.
- Strausberg RL, Dahl CA & Klausner RD (1997). New opportunities for uncovering the molecular basis of cancer. *Nature Genetics*, 15: 415-416.
- Camargo AA, Samaia HP, Dias-Neto E et al. (2001). The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proceedings of the National Academy of Sciences, USA*, 98: 12103-12108.
- Velculescu VE, Zhang L, Vogelstein B & Kinzler KW (1995). Serial analysis of gene expression. *Science*, 270: 484-487.
- Boon K, Shoemaker J, Osório EC et al. (2002). An anatomy of normal and malignant gene expression. *Proceedings of the National Academy of Sciences, USA*, 99: 11287-11292.
- Polyak K, Xia Y, Zweier JL, Kinzler KW & Vogelstein B (1997). A model for p53-induced apoptosis. *Nature*, 389: 300-305.
- He TC, Sparks AB, Rago C, Hermeking H, Zawel L, da Costa LT, Morin PJ, Vogelstein B & Kinzler KW (1998). Identification of c-MYC as a target of the APC pathway. *Science*, 281: 1509-1512.
- Lal A, Lash AE, Altschul SF et al. (1999). A public database for gene expression in human cancers. *Cancer Research*, 59: 5403-5407.
- Strausberg RL, Buetow KH, Emmert-Buck MR & Klausner RD (2000). The cancer genome anatomy project: building an annotated gene index. *Trends in Genetics*, 16: 103-106.
- Riggins GJ & Strausberg RL (2001). Genome and genetic resources from the Cancer Genome Anatomy Project. *Human Molecular Genetics*, 10: 663-667.
- Morrison TB, Weis JJ & Wittwer CT (1998). Quantification of low-copy transcripts by continuous SYBR Green I monitoring during amplification. *Biotechniques*, 24: 954-958.
- Loding WT, Lal A, Siu IM, Loney TL, Wikstrand CJ, Marra MA, Prange C, Bigner DD, Strausberg RL & Riggins GJ (2000). Identifying potential tumor markers and antigens by database mining and rapid expression screening. *Genome Research*, 10: 1393-1402.
- Kononen J, Bubendorf L, Kallioniemi A, Barlund M, Schraml P, Leighton S, Torhorst J, Mihatsch MJ, Sauter G & Kallioniemi OP (1998). Tissue microarrays for high-throughput molecular profiling of tumor specimens. *Nature Medicine*, 4: 844-847.
- Moch H, Schraml P, Bubendorf L, Mirlacher M, Kononen J, Gasser T, Mihatsch MJ, Kallioniemi OP & Sauter G (1999). High-throughput tissue microarray analysis to evaluate genes uncovered by cDNA microarray screening in renal cell carcinoma. *American Journal of Pathology*, 154: 981-986.
- Zhang L, Zhou W, Velculescu VE, Kern SE, Hruban RH, Hamilton SR, Vogelstein B & Kinzler KW (1997). Gene expression profiles in normal and cancer cells. *Science*, 276: 1268-1272.
- Parle-McDermott A, McWilliam P, Tighe O, Dunican D & Croke DT (2000). Serial analysis of gene expression identifies putative metastasis-associated transcripts in colon tumour cell lines. *British Journal of Cancer*, 83: 725-728.
- Hough CD, Sherman-Baust CA, Pizer ES, Montz FJ, Im DD, Rosenshein NB, Cho KR, Riggins GJ & Morin PJ (2000). Large-scale serial analysis of gene expression reveals genes differentially expressed in ovarian cancer. *Cancer Research*, 60: 6281-6287.
- Michiels EMC, Oussoren E, Van Groenigen M, Pauws E, Bossuyt MM, Voûte PA & Baas F (1999). Genes differentially expressed in medulloblastoma and fetal brain. *Physiological Genomics*, 1: 83-91.
- Lal A, Peters H, St Croix B, Haroon ZA, Dewhirst MW, Strausberg RL, Kaanders JH, van der Kogel AJ & Riggins GJ (2001). Transcriptional response to hypoxia in human tumors. *Journal of the National Cancer Institute*, 93: 1337-1343.
- Folkman J (1971). Tumor angiogenesis: therapeutic implications. *New England Journal of Medicine*, 285: 1182-1186.
- Folkman J (1985). Angiogenesis and its inhibitors. *Important Advances in Oncology*, 3: 42-62.
- Folkman J (1996). Fighting cancer by attacking its blood supply. *Scientific American*, 275: 150-154.
- St Croix B, Rago C, Velculescu V et al. (2000). Genes expressed in human tumor endothelium. *Science*, 289: 1197-1202.
- Nishikawa R, Ji XD, Harmon RC, Lazar CS, Gill GN, Cavenee WK & Huang HJ (1994). A mutant epidermal growth factor receptor common in human glioma confers enhanced tumorigenicity. *Proceedings of the National Academy of Sciences, USA*, 91: 7727-7731.
- Lal A, Glazer CA, Martinson HM, Friedman HS, Archer GE, Sampson JH & Riggins GJ (2002). Mutant epidermal growth factor receptor up-regulates molecular effectors of tumor invasion. *Cancer Research*, 62: 3335-3339.
- Charpentier AH, Bednarek AK, Daniel RL, Hawkins KA, Lafilin KJ, Gaddis S, MacLeod MC & Aldaz CM (2000). Effects of estrogen on global gene expression: identification of novel targets of estrogen action. *Cancer Research*, 60: 5977-5983.
- Figeys D & Pinto D (2001). Proteomics on a chip: promising developments. *Electrophoresis*, 22: 208-216.