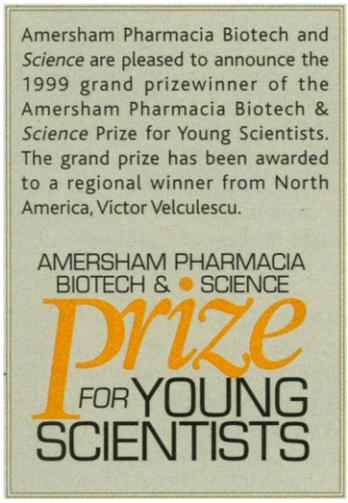
ESSAY: AMERSHAM PHARMACIA BIOTECH & *SCIENCE* PRIZE

Tantalizing Transcriptomes—SAGE and Its Use in Global Gene Expression Analysis

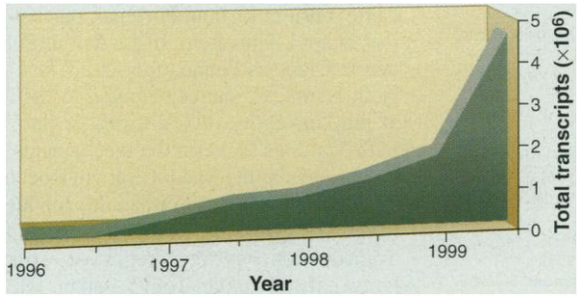
Victor E. Velculescu

Elucidating the genomic sequence of higher organisms, including that of humans, is now an achievable goal. This type of analysis, however, only represents one level of genetic complexity. A second and equally important level of complexity is the ordered and timely expression of these genes within an organism. Gene expression is encoded in eukaryotic cells by the precise quantity of messenger RNA transcript copies for each of the active genes inside a particular cell. Of the approximately 100,000 genes in the human genome, only a fraction are thought to be active in each cell type, with the expression of each gene ranging from a few to several thousand transcript molecules per cell (*1*). The several thousand different cell types that constitute the human body are each thought to have unique patterns of gene expression specifically designed for particular physiologic functions. A variety of internal or external factors can modulate these gene expression patterns, leading to altered physiologic or disease states. It is this complexity of gene expression patterns that allows an organism's genome to provide, at least in part, the diversity of information required for biologic life.

Investigators have long sought to obtain global snapshots of gene expression patterns, not only to better understand basic cellular biology but also to provide insight into human disease. Unfortunately, previous methods of evaluating gene expression have not been useful in providing overall gene expression patterns, as they have only allowed for analysis of a limited number of genes or were nonquantitative (*2–4*). The more recent methods of hybridization-based analyses using immobilized complementary DNAs (*5*) or oligonucleotides (*6*) are limited by their ability to analyze only previously isolated genes.



For my thesis, I developed a method called serial analysis of gene expression (SAGE) that allows the rapid and detailed characterization of gene expression patterns. The SAGE method is based on the isolation of unique sequence tags from individual transcripts and concatenation of tags serially into long DNA molecules. Rapid sequencing of concatemer clones reveals individual tags and allows quantitation and identification of cellular transcripts. Our initial efforts focused on developing and validating the method in a well-characterized system (*7*). These experiments, by analyzing transcripts from human pancreas, demonstrated the feasibility of SAGE and revealed a quantitative expression pattern characteristic of pancreatic function.



Cumulative transcripts analyzed by SAGE worldwide. For additional information on SAGE publications please see www.sage.net or www.ncbi.nlm.nih.gov/SAGE.

These successes led us to reason that SAGE could, in principle, be used to characterize the entire set of genes expressed from a eukaryotic genome. Such global patterns of gene expression can be represented by a "transcriptome," conveying the identity of each expressed gene and its level of expression for a defined population of cells. Unlike the genome, which is essentially a static entity, the transcriptome can be modulated by both external and internal factors and thereby serves as a dynamic link between an organism's genome and its physical characteristics.

A transcriptome as defined above has not been characterized in any eukaryotic or prokaryotic organism, largely because of technological limitations. General features of gene expression patterns were elucidated more than two decades ago through RNA-DNA hybridization measurements (*1*), but these studies did not provide much information about the identities of the expressed genes within each expression class. Data on the expression levels of individual genes have accumulated as new genes were discovered, but in only a few instances have the absolute levels of expression of particular genes been measured and compared to other genes in the same cell type. Description of any cell's transcriptome would therefore provide new information useful for understanding numerous aspects of cell biology and biochemistry.

Using SAGE, I was able to provide the first description of a transcriptome, in the yeast *Saccharomyces cerevisiae* (*8*). This organism was chosen because it is widely used to clarify the biochemical and physiologic parameters underlying eukaryotic cellular functions and because, at the time, it was the only eukaryote in which the entire genome had been defined at the nucleotide level (*9*). These studies, comprising more than 60,000 transcripts, identified the expression of 4665 genes from the yeast genome, with expression levels ranging from 0.3 to more than 200 transcript copies per cell. Surprisingly, several hundred new genes were identified that had not been previously predicted by analysis of the complete yeast genome sequence (*8, 10*).

With the demonstration of the utility of SAGE and its ability to provide global expression analyses, I applied SAGE to analyze the complexity of human cancer. Although hundreds of studies over the past 20 years have pointed out differences in the expression of one or a few genes, no comprehensive study of gene expression in cancer cells has ever been reported. It was, therefore, not known how many genes were expressed differentially in tumor versus normal cells, and whether these changes played a role in driving cancer progression. As a step in unraveling the complexities of gene expression differences in cancer cells, I have used SAGE to characterize gene expression differences regulated by the *ras*

The author is at the Johns Hopkins Oncology Center, 424 North Bond Street, Baltimore, MD 21231, USA. E-mail: velculescu@jhmi.edu

oncogene, providing potential molecular mechanisms for *ras*-mediated oncogenesis. Others have used SAGE to comprehensively analyze gene expression profiles in gastrointestinal (*11*) and lung cancers (*12*), to identify transcriptional targets of p53 that regulate apoptosis (*13*) and G_2 cell cycle arrest (*14*), and to identify myc as a downregulated target of the *APC* tumor suppressor gene (*15*).

Our studies in human cancer cells with SAGE have allowed us to characterize human transcriptomes (*16*). Analysis of 3.5 million transcripts from 19 normal and diseased tissues has identified the expression of about 84,000 genes (providing an estimate of the minimum number of genes contained in the human genome) and revealed that more than 43,000 genes can be expressed in a single cell type. Additionally, these analyses identified specific genes that were exclusively expressed in individual cell types, the set of genes that were expressed in all cell types, and a small number of genes that were uniformly elevated in many cancers as compared to their normal counterparts.

The analyses performed by SAGE have allowed a heretofore unavailable picture of global gene expression patterns, drawing us deeper into the genetic complexity of eukaryotic life. The number of systems that SAGE can be applied to and the abundance of data generated by such analyses will be limited only by experimental ingenuity. Like the genome sequence itself, these studies provide a wealth of information integral to future experimentation in normal and disease states.

References and Notes

1. B. Lewin, *Gene Expression* (Wiley, New York, 1980).
2. M. Hedrick, D. I. Cohen, E. A. Nielsen, M. M. Davis, *Nature* **308**, 149 (1984).
3. P. Liang and A. B. Pardee, *Science* **257**, 967 (1992).
4. M. D. Adams *et al.*, *Nature* **377**, 3 (1995).
5. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
6. D. J. Lockhart *et al.*, *Nature Biotechnol.* **14**, 1675 (1996).
7. V. E. Velculescu, L. Zhang, B. Vogelstein, K. W. Kinzler, *Science* **270**, 484 (1995).
8. V. E. Velculescu *et al.*, *Cell* **88**, 243 (1997).
9. A. Goffeau *et al.*, *Science* **274**, 546 (1996).
10. M. A. Basrai, V. E. Velculescu, K. W. Kinzler, P. Hieter, *Mol. Cell. Biol.* **19**, 7041 (1999).
11. L. Zhang *et al.*, *Science* **276**, 1268 (1997).
12. K. Hibi *et al.*, *Cancer Res.* **58**, 5690 (1998).
13. K. Polyak, Y. Xia, J. L. Zweier, K. W. Kinzler, B. Vogelstein, *Nature* **389**, 300 (1997).
14. H. Hermeking *et al.*, *Mol. Cell* **1** (1997).
15. T. C. He *et al.*, *Science* **281**, 1509 (1998).
16. V. E. Velculescu *et al.*, *Nature Genet.*, in press.

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

1999 Grand Prize Winner

Victor E. Velculescu

Victor E. Velculescu was born on 16 August 1970 in Bucharest, Romania, and grew up in Thousand Oaks, California. In 1992, he earned his bachelor's degree in the Department of Biological Sciences at Stanford University. He later attended the Johns Hopkins University School of Medicine where he was awarded his Ph.D. in the Program of Human Genetics and Molecular Biology in 1998 and his M.D. in 1999. His doctoral work on the development and application of SAGE to analyze gene expression patterns was performed in the laboratory of Ken Kinzler at the Johns Hopkins Oncology Center. Since January 1999, Dr. Velculescu has been working as a postdoctoral fellow in the laboratory of Bert Vogelstein at the Johns Hopkins Oncology Center.

Regional Winners

Europe: Giles E. Hardingham, for his essay, "The Flexibility of the Calcium Signal in Activating Gene Expression," reporting work performed at the Medical Research Council's (MRC) Laboratory of Molecular Biology, in Cambridge, UK. Dr. Hardingham was born in Wimbledon, London, in 1973 and was raised in a family of scientists. After studying Natural Sciences at Cambridge University, he moved to Hilmar Bading's group at the MRC, where he performed his doctoral research. He is now an MRC Fellow at the MRC laboratory of Molecular Biology and a Research Fellow of Clare College, Cambridge, where he continues to study the activation of gene expression in hippocampal neurons.

North America: Lisa Goodrich, for her essay, "Patching Together Development and Disease," based on work performed in the Department of Developmental Biology at Stanford University. Dr. Goodrich was born in Washington, DC, and grew up in Boston, MA. She worked as an undergraduate at Harvard in the laboratory of John Dowling, in which she studied the development of the retina in zebrafish. After a year working on the development of ultraviolet photoreceptors in salmon in the laboratory of Yvette Kunz at University College Dublin in Ireland, she joined Matthew Scott's lab at Stanford for her doctoral work on the role of vertebrate *patched* genes in neural development and disease. Dr. Goodrich is now a postdoctoral fellow with Marc Tessier-Lavigne at the University of California, San Francisco where she is screening mouse genes for novel axon guidance receptors.

The second North American regional winner was Marilia Cascalho, for her essay, "Mismatch Repair and Somatic Hypermutation—A Tale of a Double-Edged Sword," based on her thesis research at the University of California at San Francisco. Dr. Cascalho, born in Lisbon, Portugal, received her M.D. degree in 1986 at the Lisbon Medical School, University of Lisbon in Portugal. During a fellowship funded by the Juvenile Diabetes Foundation with Åke Lernmark at the Hagedorn Research Laboratory in Denmark, she contributed to the identification of a diabetes autoantigen. Then at the University of California at San Francisco, she performed her doctoral work with Matthias Wabl on the mechanisms of somatic mutation. Since 1999, Dr. Cascalho has been at the Mayo Clinic in Rochester, MN.

Japan: Toshimasa Yamauchi, for his essay, "Discovery of Novel Cross-Talk Between the Cytokine Receptor Superfamily and the Growth Factor Receptor Signal Transduction Pathway," reporting work done at the Graduate School of Medicine, University of Tokyo, Tokyo, Japan. Dr. Yamauchi was born in Kyoto, Japan, and received his M.D. from the Faculty of Medicine and his Ph.D. at the University of Tokyo, Graduate School of Medicine. He described the cross-talk between cytokine and growth factor receptors while working under the direction of Takashi Kadowaki and Yoshio Yazaki, in the Third Department of Internal Medicine, Faculty of Medicine, at the University of Tokyo. Dr. Yamauchi holds a Research Fellowship from the Japan Society for the Promotion of Young Scientists at the Graduate School of Medicine, University of Tokyo, where he is studying the signal transduction pathways of hormones.

The full text of the essays written by the regional winners and information about applying for next year's award can be seen on *Science* Online at www.sciencemag.org/feature/data/pharmacia/1999.shl

References and Notes

[3] **Differential Display of Eukaryotic Messenger RNA by Means of the Polymerase Chain Reaction**
Peng Liang; Arthur B. Pardee
Science, New Series, Vol. 257, No. 5072. (Aug. 14, 1992), pp. 967-971.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819920814%293%3A257%3A5072%3C967%3ADDOEMR%3E2.0.CO%3B2-O

[5] **Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray**
Mark Schena; Dari Shalon; Ronald W. Davis; Patrick O. Brown
Science, New Series, Vol. 270, No. 5235. (Oct. 20, 1995), pp. 467-470.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819951020%293%3A270%3A5235%3C467%3AQMOGEP%3E2.0.CO%3B2-B

[7] **Serial Analysis of Gene Expression**
Victor E. Velculescu; Lin Zhang; Bert Vogelstein; Kenneth W. Kinzler
Science, New Series, Vol. 270, No. 5235. (Oct. 20, 1995), pp. 484-487.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819951020%293%3A270%3A5235%3C484%3ASAOGE%3E2.0.CO%3B2-S

[11] **Gene Expression Profiles in Normal and Cancer Cells**
Lin Zhang; Wei Zhou; Victor E. Velculescu; Scott E. Kern; Ralph H. Hruban; Stanley R. Hamilton; Bert Vogelstein; Kenneth W. Kinzler
Science, New Series, Vol. 276, No. 5316. (May 23, 1997), pp. 1268-1272.
Stable URL:
http://links.jstor.org/sici?sici=0036-8075%2819970523%293%3A276%3A5316%3C1268%3AGEPINA%3E2.0.CO%3B2-B

LINKED CITATIONS

[15] **Identification of c-MYC as a Target of the APC Pathway**

Tong-Chuan He; Andrew B. Sparks; Carlo Rago; Heiko Hermeking; Leigh Zawel; Luis T. da Costa; Patrice J. Morin; Bert Vogelstein; Kenneth W. Kinzler

Science, New Series, Vol. 281, No. 5382. (Sep. 4, 1998), pp. 1509-1512.

Stable URL:

http://links.jstor.org/sici?sici=0036-8075%2819980904%293%3A281%3A5382%3C1509%3AIOCAAT%3E2.0.CO%3B2-U