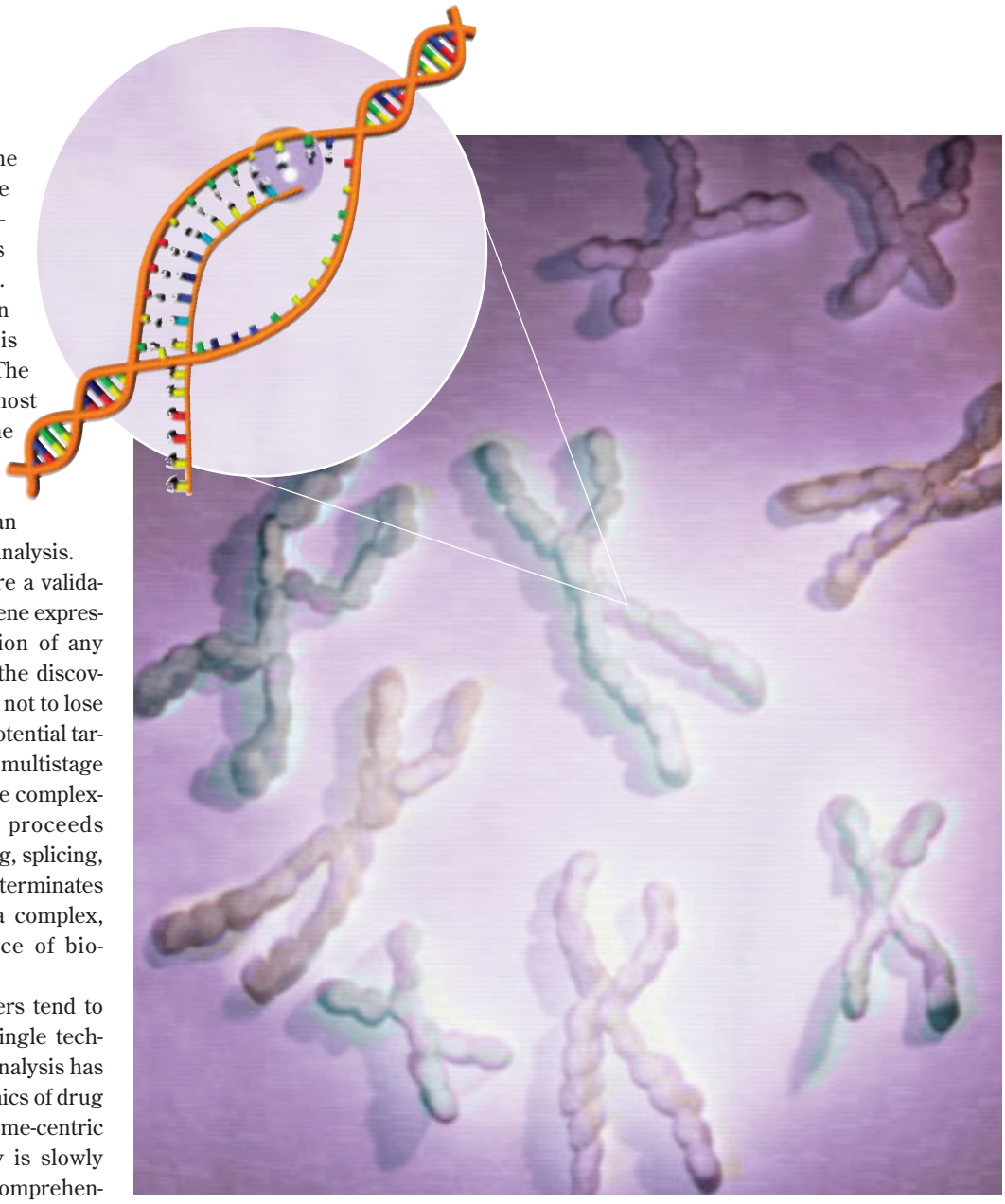
TARGETING THE TRANSCRIPTOME

Many levels of complexity, many biochemical techniques, but no high-throughput solution

BY JEFFREY AUGEN



It is well known that the drug discovery pipeline is less than 5% efficient—most projects entering the pipeline fail. Moreover, the most common reason for ending a project is failure to validate a target. The missing subtlety is that most targets entering the pipeline are discovered using contemporary high-throughput techniques rather than more complete metabolic analysis. The 95% failure rate is more a validation of the complexities of gene expression than it is an indication of any particular shortcoming in the discovery process. It is important not to lose sight of the fact that each potential target is the end point of a multistage process that begins with the complexities of gene structure, proceeds through a myriad of copying, splicing, and regulatory steps, and terminates with transcription—itself a complex, highly regulated sequence of biochemical reactions.

The fact that researchers tend to depend too heavily on a single technique such as microarray analysis has a real impact on the economics of drug discovery. Thus, the genome-centric view of molecular biology is slowly being replaced by a more comprehen-

PHOTO: PHOTODISC

sive systems view. One of the most important elements of this approach is a comprehensive understanding of the transcriptional state of all genes involved in a specific metabolic profile. This collection of transcribed elements has recently come to be known as the transcriptome.

Any technique used to study the transcripts within a cell must be capable of spanning a broad dynamic range—from single-digit copy counts to large numbers, often in the thousands. Accuracy is important because at the single-digit level, small changes in the number of copies of certain messages can significantly affect metabolism and disease. This picture is further complicated by the fact that many species of RNA that play an important role in metabolism are never translated into protein. Some are degraded by the RNA silencing machinery in the cell, whereas others are prevented from engaging in protein translation. These control mechanisms can result in a message that is highly abundant within the cell being translated into a small number of protein molecules. Similarly, although regulatory messages (miRNA, siRNA) are relatively straightforward to spot because they are reproducibly short, they are spliced from longer transcripts that have the potential to cause confusion.

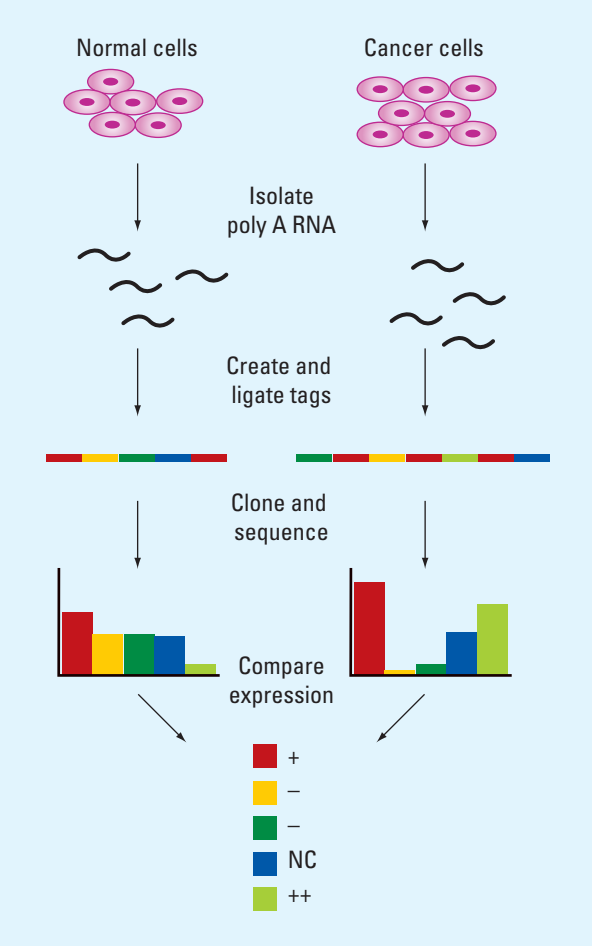


Figure 1. Time for SAGE. Using serial analysis of gene expression, or SAGE, researchers compare the expression of short tag sequences in normal and perturbed tissue samples to look for disease- or treatment-specific changes.

Microarrays

Microarrays have emerged as the tool of choice for studying expression profiles. One drawback of microarray analysis, however, is related to its inability to distinguish very-low-abundance transcripts—those present in single-digit copy counts in which the copy range is large. A more common problem involves the quantification of mRNA species from a large population of cells where the molecule of interest is only present in a small subpopulation. In such situations, the species of interest is likely to be diluted beyond the detection limit by more abundant transcripts appearing across the broader population. Increasing the absolute amount of the hybridized target is not usually helpful because it is the relative abundance of each transcript in the RNA pool, coupled with probe characteristics, that determines the sensitivity of the array for each sequence (*1*). Unfortunately, it is also difficult to identify transcripts that are upregulated by less than 50% using microarrays, a significant problem for researchers in areas such as oncology and neuroscience, where subtle changes in gene expression are critical to understanding the differences between disease and health.

SAGE advice

A variety of techniques have emerged to address shortcomings in microarray technology. Most of them focus on precise measurements and accuracy across a broad dynamic range. The first, known as serial analysis of gene expression (SAGE), can be used to discover every expressed transcript in a cell regardless of whether that message was previously known (*2*).

SAGE is designed to measure the expression of a small "tag" rather than the entire transcription product of a gene (*3*). The tag, defined as a 10-base-pair region directly adjacent to the 3′-end of the first occurrence of the sequence CATG, is generated through a process of well-defined enzymatic digestions and oligonucleotide purification steps. The process of generating SAGE tags is outlined in Figure 1.

The simplicity of SAGE analysis makes it an excellent tool for measuring the appearance and disappearance of a small number of highly regulated transcripts under a well-characterized set of conditions. The technique has been used to study expression of the *p53* gene, known to be inactive in many human cancers. Although expression of the gene is known to induce either stable growth arrest or programmed cell death (apoptosis), the mechanism underlying development of *p53*-dependent apoptosis remains incompletely understood (*4*). SAGE is an appropriate technique for such analysis because the transcripts each come from well-characterized genes that have been completely sequenced.

Recent studies of human *p53* expression have revealed the upregulation of 14 transcripts, 8 of which are known to code for proteins involved in cellular responses to oxidative stress. These observations stimulated additional biochemical experimentation that ultimately suggested a three-step

process for *p53*-induced apoptosis: transcriptional induction of redox-related genes, formation of reactive oxygen species, and oxidative degradation of mitochondrial components. Leakage of calcium and proteinaceous components from the damaged mitochondria stimulates caspases that are ubiquitously activated during the apoptotic process.

However, like microarrays, SAGE has deficiencies. Most notable is its dependence on small "tag" sequences. The mRNA splicing process has the capability of confounding techniques like SAGE that rely on partial sequences to identify expressed transcripts. Splice variants that differ outside the region being studied cannot be distinguished by such techniques. Even when splice variants are distinguishable, subtle differences outside the region of study will not be revealed.

The solution to this problem is to use a combination of techniques. For example, SAGE can reveal the presence of a specific gene, and microarray technology can be used to enumerate the complete set of splice variants. However, it is important to point out that microarray results can also be difficult to interpret in this context because each spot in the array corresponds to a short portion of a transcript (~25 bases). Although overlaps between the short cDNA templates contained in the array can be used to assemble complete sequences, the presence of unspliced messages and message fragments removed during splicing can confuse the interpretation. In general, it is almost impossible to use microarrays, or any other high-throughput technique, to unambiguously identify the specific combinations of exon variants that occur together in individual mRNA molecules.

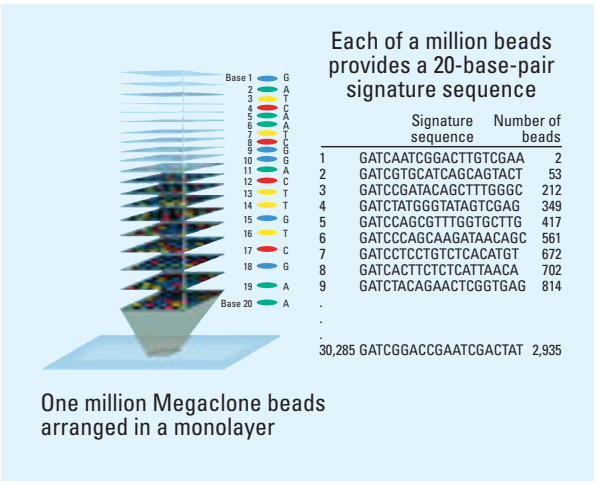


Figure 2. Taking a bead. The resin-based Megaclone method allows researchers to amplify, clone, and sequence millions of transcripts in parallel. (Image courtesy of Lynx Therapeutics.)

Parallel efforts

Researchers at Lynx Therapeutics (*www.lynxgen.com*) recently developed a new technology that generates accurate and precise copy counts for every transcript in a sample regardless of the number of transcripts or the range of copy counts. This patented technology was invented by Sydney Brenner, whose goal was to be able to obtain an accurate accounting of the complete transcriptional profile of a cell—major and minor messages, transient messages, short-lived transcripts, and control sequences. The strategy involves several innovative approaches to amplifying, cloning, and sequencing millions of transcripts in parallel (*5–7*) using two new technologies.

The first method, Megaclone, allows a cDNA copy of each message to be cloned onto the surface of a 5-µm bead. Through message amplification and selection processes, each bead is ultimately decorated with 100,000 copies of the same sequence. Thus, a message present in 20 copies will appear on 20 different beads with 100,000 copies per bead. (To see Megaclone in action, visit *www.lynxgen.com/wt/animation.php3?page_name=megaclone_ani*.)

The second technique, massively parallel signature sequencing (MPSS), uses sequence-dependent fluorescent signatures to successively identify four-base-pair segments and ultimately construct sequences from the bead-bound cDNA. With more than 1 million beads immobilized in a single-layer array inside a flow cell, solvents and reagents can be washed over the samples in each cycle of the process. Sequence data are ultimately matched against more complete, well-characterized entries in the public database infrastructure. (To see MPSS in action, visit *www.lynxgen.com/wt/animation.php3?page_name=mpss_ani*.)

The combination of Megaclone and MPSS (Figure 2) represents a complete departure from analog detection techniques because each transcript is separately cloned, amplified, and sequenced as many times as it appears in the cell without any dependence on comparisons or extrapolations. Because the technique involves precise counting, accuracy is maintained across a broad range of copy counts. Finally, the digital nature of the Megaclone–MPSS strategy facilitates precise comparisons between experiments and eliminates the need to make final adjustments to expression values.

Researchers at the National Cancer Institute (*www.cancer.gov*) recently used MPSS to characterize the transcriptomes of two human cell lines. The experiments involved analyzing transcripts from cell lines derived from normal breast epithelium (HB4a), and a colon adenocarcinoma (HCT-116) that represented nearly 17,000 genes, of which about 55% were expressed at fewer than 10 transcripts per million. The vast majority of the transcripts found in these cells can be mapped to known genes and their polyadenylation variants. Among the genes that could be identified from their signature sequences, approximately 8500 were expressed by both cell lines, whereas 6000 showed cellular specificity (*8*).

However, the Megaclone–MPSS technique is not without drawbacks. Like SAGE, MPSS suffers from an inability to provide complete information about SNPs and other polymorphisms. Because each signature sequence represents a final enzyme digestion product (the length of cDNA between the bead-bound tag and the first CATG), any additional sequence information existing beyond the message cleavage site is automatically lost. An upregulated tran-

script containing a SNP outside the signature sequence will simply appear to be upregulated, and the fact that a polymorphism is present will be lost in the analysis.

One could envision a more complex situation in which a disease group associated with the upregulation of a specific gene contains two different subpopulations: one with a SNP in the gene, and one without. The upregulation would be quantified by MPSS, but the SNP might fall outside the signature sequence region. These data would not be helpful in explaining the presence of two subgroups, but they might have some diagnostic value for the overall disease. Moreover, if the subgroups required two distinctly different treatment regimens, MPSS alone would fail as a complete diagnostic. By comparison, microarray analysis would likely identify both the SNP and the upregulation event but might have other problems related to dynamic range if key transcripts were up- or down-regulated across a broad range of copy counts.

Improving the efficiency of discovery is a difficult problem because there is no simple set of rules that can be used to select analysis techniques.

In general, if the goal is to achieve very accurate copy counts for all transcripts across a broad range, then MPSS is likely to be a valuable tool. Conversely, if the goal is to interrogate every gene at the single-nucleotide level, possibly to identify SNPs or other small aberrations, then microarray technology is likely to represent a better fit. Many situations will require both sets of capabilities. The fact that MPSS data is often used to search public databases populated with more complete sequence information obtained through microarray analysis supports this view. The reverse is also true in the sense that microarray analysis can be used to identify interesting transcripts that can then be quantified at the single-copy count level with MPSS.

The splice variant problem

Messenger RNA splice variants can confound almost any high-throughput expression profiling technique. SAGE and Megaclone–MPSS are especially vulnerable because they rely on partial sequences to identify expressed transcripts—splice variants whose differences exist outside the tag sequence region cannot be distinguished by these techniques. Furthermore, using SAGE or MPSS-derived sequences as probes to search large gene sequence databases is unlikely to help because basic sequence data is not useful as a predictor of transcriptional splicing. Conversely, once SAGE or MPSS has revealed the presence of a specific gene, microarray technology can be used to enumerate the complete set of splice variants.

Unambiguous identification of individual mRNA isoforms can only be obtained using low-throughput approaches based on amplification and end-to-end sequencing of individual transcripts. Nuclease protection assays (NPAs) are often used in this context to test for the presence of individual sequence elements in an isolated transcript or population of transcripts. The NPA is based on hybridization of single-stranded, well-characterized antisense probes to an RNA sample. After hybridization, remaining unhybridized probes and sample RNA are removed by digestion with a mixture of nucleases. After nuclease inactivation, the remaining double-stranded probe–target hybrids are precipitated and purified for further sequence analysis. NPAs can reveal the presence of specific exons in a mixture of mRNA species. The combination of high-throughput profiling approaches with a carefully planned sequence of NPAs can be a powerful solution to the problem of comprehensive mRNA profiling. These approaches are complemented by a growing base of transcription profiles and complete transcriptome maps that are rapidly becoming part of the public database infrastructure for bioinformatics.

Nuclease protection techniques have been helpful in identifying a variety of specific genetic elements that exhibit regulatory functions in various disease states. For example, the technique was recently used to study the promoter region affecting transcription of the AP-2γ protein, which is known to be overexpressed in a high percentage of breast cancer tumors. DNase footprinting led to the identification of three binding sites in this region, two of which are required for both promoter function and cell-type-specific activity. One of the regions, SP-3, was identified as having a critical control function with regard to expression levels of the AP-2γ protein (*9*).

Improving the efficiency of discovery is a difficult problem because there is no simple set of rules that can be used to select analysis techniques. The answer is likely to lie in a broad-based approach that involves building libraries of information about the genome, transcriptome, and proteome. In the absence of such information, individual discovery projects will continue to grow into broad biochemical investigations that require analysis at many different levels. The alternative is a <5% success rate for projects entering the pipeline.

References

(1) Mimics, K.; et al. *Trends Neurosci.* **2001,** *24,* 479–486.
(2) Madden, S.; Wang, C.; Landes, G. *Drug Discov. Today* **2000,** *5,* 415–425.
(3) Roberts, G.; Smith, C. W. J. *Curr. Opin. Chem. Biol.* **2002,** *6,* 375–383.
(4) Polyak, K.; et al. *Nature* **1997,** *389,* 300–305.
(5) Brenner, S.; et al. *Nat. Biotechnol.* **2000,** *18,* 630–634.
(6) Brenner, S.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2000,** *97,* 1665–1670.
(7) Reinartz, J.; et al. *Brief. Func. Genom. Proteom.* **2002,** *1,* 95–104.
(8) Jongeneel, C. V.; et al. *Proc. Natl. Acad. Sci. U.S.A.* **2003,** *100,* 4702–4705.
(9) Hasleton, M. D.; Ibbitt, J. C.; Hurst, H. C. *Biochem. J.* **2003,** *373,* 925–932.

Jeffrey Augen is president and CEO of TurboWorx (*www.turboworx.com*). ■