Identifying expressed genes

Katherine J. Martin* and Arthur B. Pardee

Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA 02115

The study of expressed genes has had a great impact on biological research (1). Expressed genes are the basic functional units of genomic DNA. Because these regions cannot be identified from genomic sequence information *per se*, the gene's products, messenger RNAs or proteins, must be isolated from cells, directly sequenced, and identified. As we steadily build up sizable expression databases that currently include more than 92,000 of the roughly 100,000 total human genes, the process of identifying the remaining undiscovered genes is becoming progressively more difficult. Existing databases of expressed genes now include virtually all of the abundantly expressed human genes—the easier to reach "low hanging fruit on the tree," as well as many middle and rarely expressed genes. The genes that are still undiscovered are expressed at low levels or are specifically expressed only in certain cell types, developmental stages, or growth conditions. Such genes hold the promise of including key regulatory factors responsible for differentiated phenotypes, developmental progression, or cell growth regulation. As we move forward to identify these genes, highly efficient methods of removing, i.e., subtracting, the bulk of identified, abundant genes from cDNA libraries are required.

In this issue of PNAS, Wang and colleagues (2) discover a flaw in current subtraction methods, which are now widely used to identify novel expressed genes. They show that the long poly(A) regions present in most expressed mRNAs generate a serious problem in subtraction reactions. Long poly(dT) regions of tester cDNA, which is generated from the RNA of interest, randomly hybridize with long poly(dA) regions of driver cDNA generated from the comparison cell type, resulting in template loss. This loss particularly affects low abundance mRNAs. Wang *et al*. predict that this flaw will limit the usefulness of current subtraction methods and result in a fall-off in gene identification rates before the identification of all genes is completed. They report a conceptually and technically straightforward solution that significantly enhances the efficiency with which novel expressed genes are identified. They generate subtraction libraries by using short, anchored oligo(dT) primers that anneal at the 5′ ends of poly(A) regions of mRNAs. Hence, the cDNAs produced are devoid of long poly(dA) regions.

Inception of Expression Analysis

Information on expressed genes has classically originated from studies of individual cDNAs, identified and cloned by virtue of their particular importance to a specific topic of research. For the past 20 years, DNA sequence information for these functionally characterized genes has been entered into databases including GenBank, the European Molecular Biology Laboratory, and the DNA Data Base of Japan, which share their respective contents.

A random approach recently has been used as a part of a large-scale effort to collect DNA sequence information for all expressed human genes. This approach entails sequencing partial cDNA clones generated from mRNA and is termed expressed sequence tag (EST) analysis (3). EST sequences generally represent 200–800 bp of first-pass sequence information extending in from mRNA 3′ ends. Two large public EST sequencing projects, the EST project and the Cancer Genome and Anatomy Project (CGAP, http://www.ncbi.nlm.nih.gov/ncigap), have been initiated to rapidly identify, i.e. obtain at least partial sequence information for all expressed genes.

The first EST project (3) was begun in 1991 and to date has accumulated sequences for a total of approximately 48,000 different genes. Rates of novel gene discovery by the EST project were initially high, but have declined sharply in recent years. Ninety percent of the 48,000 ESTs were accumulated in the 4 years before 1997 and only modest numbers were added in the past 3 years. The second large-scale expressed sequence tag project, the Cancer Genome and Anatomy Project (CGAP) (http://www.ncbi.nlm.nih.gov/ncicgap), was begun in late 1996. CGAP currently is maintaining high rates of gene discovery by applying the latest techniques in tissue procurement, cDNA library preparation, and bioinformatics to sequence expressed genes in cancerous, precancerous, and normal cell lines and tissues. CGAP now has accumulated more than 44,000 novel genes that were not already discovered by the EST project. Sequences from both of the EST sequencing projects are collected in the database of ESTs (dbEST, http://www.ncbi.nlm.nih.gov/dbEST), which is maintained by the National Center for Biotechnology Information. GenBank and dbEST sequences then are organized into a nonredundant list of unique genes by the UniGene project (http://www.ncbi.nlm.nih.gov/UniGene), which is considered the most regularly updated source for high-quality, nonredundant information on expressed genes. CGAP also is creating a public database, SAGEmap, to provide quantitative gene expression data (4).

Public human expression databases now are believed to include a large percentage of all genes. UniGene currently lists sequence information for 92,571 different expressed genes (UniGene build #108, Feb. 19, 2000). It is noted that the algorithms used by UniGene to cluster redundant sequences are experimental and hence this number may increase or decrease with improvements and the addition of new sequences. Further, some sequences currently considered to be different genes may in fact represent nonoverlapping regions of the same gene. Hence, more complete sequence information, e.g., from genomic data, also may reduce the UniGene tally. The ultimate target is also uncertain. Estimates of the total number of expressed human genes range between 60,000 and 150,000 (5–7). Serial analysis of gene expression (SAGE) results indicate that 46% of genes currently have no matches in existing databases (2, 7), hence predicting a total of 130,000 genes. Although precise values for the total number of expressed human genes or the number that have already been identified are uncertain, it is likely that current databases are close to complete.

The dbEST obtains its information from many different types of libraries and tissues. Libraries include PCR-amplified and unamplified libraries, normalized, subtracted, and unaltered libraries, as well

See companion article on page 4162.

*To whom reprint requests should be addressed.

as libraries generated with a new reverse transcriptase that produces very long cDNA clones. Libraries are prepared from more than 1,000 different human tissues and cell lines, including normal and cancer cells, different developmental stages and growth states, as well as microdissected, bulk, and pooled tissues.

A parallel ongoing effort is to sequence the complete human genome. Draft sequence currently is reported for 47% (Human Genome Project, http://www.ncbi.nlm.nih.gov/genome/seq) and more than 90% (Celera Genomics, http://www.celera.com) of the genome. Upon completion, these projects will provide the basic fundamental DNA sequence information for the entire 3×10^9 bp of the human genome. Genomics has made key contributions to biology and medicine. However, its greater value may be as a component of integrated resources of genomic data plus data on the 3% of its sequence that is translated. Integrated databases, linked to functional information on molecular processes and disease states, hold the potential for revolutionizing methods of basic research and disease intervention (8–10).

In addition to library subtraction and normalization, other methods of novel gene discovery currently are used. These include methods that incorporate a subtraction step and methods that do not. Subtraction-based methods include, for example, representational difference analysis (11). Methods that do not involve subtraction include differential display (DD) (12) and related fingerprinting techniques, SAGE, differential library screening, and negative selection (13, 14).

DD is a PCR-based method whereby cDNAs made from two mRNA samples are compared on side-by-side tracks of a sequencing gel. Each mRNA is represented as a single band and differential bands can isolated and cloned. In parallel comparisons, DD was found to have less bias toward abundant genes than EST sequencing or subtractive hybridization (15). In recent years, false positive rates for DD have been reduced from 50% to close to 5% (16). The method currently is widely applied to solve specific research questions (nearly 1,400 citations in Medline), although it contributes only a minor fraction of EST and CGAP entries.

SAGE is a streamlined method of sequencing genes from any type of expression library, subtracted, normalized, or unaltered (13). In SAGE, 10-bp stretches of cDNA obtained from a precise location relative to the mRNA 3′ end are concatemerized so that a single lane of DNA sequencing identifies many genes. SAGE is used as a tool for both quantitative analysis of gene expression levels and novel gene discovery.

Subtraction and Normalization Methods

As sequencing continues, the remaining unidentified genes are becoming progressively harder to find because they are of progressively lower abundance and are more cell-type restricted. It is important that databases include even these most scarce and tissue-specific genes, as these have the potential to include many of the most biologically interesting regulatory factors. Subtraction and normalization are key methods in the process of identifying such genes.

Normalization methods selectively reduce the level of representation of abundant genes so that the resulting mRNAs preparations contain both abundant and rare genes at similar levels. Before normalization, a typical cell expresses 1,000–2,000 different abundant and middle abundant messages at levels of >500 and 15–500 copies/cell, respectively. These represent 50–65% of the cell's total mRNA mass (17, 18). In contrast, approximately 15,000 different rare messages are expressed at a level of <5 copies per cell and constitute the remaining percentage of a cell's mRNA mass.

Subtraction methods allow one to compare two mRNA preparations and remove from one (the tester) genes that are present in the other (the driver). This technique is useful in identifying tissue-specific genes. Current EST sequencing combines subtraction and normalization methods with advanced tissue preparation methods, such as the isolation of RNA from bulk, microdissected, and pooled tissues, to increase the diversity and reduce the redundancy of messages in a given sample.

To perform a subtraction, first-strand cDNA is generated from tester RNA by using poly(dT) primers. Double-stranded cDNA then is generated from driver RNA again by using the poly(dT) primers. Tester cDNA is mixed with an excess of driver cDNA, and the cDNAs are denatured and allowed to reanneal. Double-stranded cDNA is eliminated by adsorption to a hydroxyapatite column. Normalization methods use the same basic principles, except that a single RNA preparation is used to generate both tester and driver cDNA and hybridization time is limited.

Improving Expressed Gene Identification

Wang and colleagues (2) describe the fall-off in rates of new gene discovery by the EST project. "The rate of novel gene identification through the EST project declined dramatically from 10.6% of EST sequences in 1996 (36,000 novel sequences) to only 2.7% of EST sequences collected in 1998 (638 novel sequences)." They predict a similar decline in CGAP gene identification rates once current methodologies become limiting. They identify a problematic area of current subtraction and normalization methods that may limit the usefulness of these methods.

Conventionally, reverse transcription is performed with a poly(dT) primer that anchors randomly along the approximately 200-bp poly(A) tails of mRNAs and creates long poly(dA/dT) sequences in the cDNAs. During subtraction, rare cDNAs are lost at a high rate because of random hybridization of their long poly(dA) regions with driver poly(dT). Wang *et al.* (2) demonstrate a method that overcomes this problem: the construction of subtraction libraries by using short, anchored oligo(dT) primers. These primers are composed of 11 dTs plus one or two other 3′ bases that anchor their hybridization at the 5′ ends of poly(dA) sequences. The short poly(dA/dT) tails produced are less likely to cause the removal of rare cDNAs. Such short anchored primers previously were used in the DD technique (12) to target reverse transcriptase and PCR to the 3′ ends of expressed genes.

Wang *et al.* (2) demonstrate that these anchored primers do indeed produce short 3′ dA/dT sequences. These sequences reduced loss of a synthetic tester template during subtraction reactions. They also reduced loss of rare cellular genes. A selected group of rare colon-specific mRNAs was assayed after subtraction. They were increased several-fold in the poly(dA/dT)(-) libraries made by using the anchored primers, as compared with conventional dA/dT(+) libraries; four of the five rare colon specific genes were retained at 1.4- to 7.8-fold higher levels. Furthermore, Moloney murine leukemia virus reverse transcriptase was more effective than avian myeloblastosis virus reverse transcriptase in reducing rare template loss.

A point of interest is that the dC-anchored primer behaved like poly(dT), by randomly hybridizing along poly(A) tails of mRNAs. Therefore, two base-anchored primers were used in its place in which the penultimate base was dC plus an ultimate dC, dA, or dG. Interestingly, we find with DD that dC is a perfectly satisfactory 3′ base perhaps because the repeated 30 rounds of PCR progressively shorten the 3′ tails to a minimum.

New Directions

Expression approaches already have had a great impact on biological research (1, 19). Expression databases are receiving thousands of queries per day by researchers. The approach is providing researchers with new tools to address complex molecular events. In a general sense, expression approaches are contributing to the current shift in research directions from the classical emphasis on individual genes and molecules to the investigation of patterns of gene expression and the elucidation of comprehensive functional networks. New technologies based on the expression approach include cDNA microarrays and computer-based expression analyses.

Microarrays in particular have had an overwhelming recent popularity. cDNA tags for thousands of expressed genes are arrayed on a glass or membrane support. When incubated with labeled first-strand cDNA produced from cellular RNA, each tag hybridizes to its cognate mRNA and allows relative quantitation of the expression levels. This technology allows widespread changes in expression patterns to be probed in a single experiment.

Important clinical applications also are developing from expression technology. Expression information linked to functional information on molecular processes and disease states can accurately classify patients according to distinct subcategories of disease. Such classification has the potential to assist clinicians in making prognosis and predicting the most effective therapies.

The power of hybridization arrays can be seen by the questions now being answered by this method. For example, two groups have studied large-scale gene expression changes in leukemias and lymphomas and found that groups of genes can be readily identified that sort blood samples *a priori* into different disease groups (8, 9). Studies of breast tumor tissue biopsies also showed that clinically relevant gene expression patterns could be identified (10, 20). These expression patterns have the potential to very accurately identify prognostic groups and predict the most effective therapies.

Currently, the most comprehensive hybridization arrays include 20–30% of the 130,000 genes. Hybridization array methods will become more powerful when tags for all expressed genes are included. For the most part, these genes are already in databases. Arrays with complete collections of tags, as well as the equipment to hybridize and interpret the results, will likely follow. A significant challenge currently is to develop approaches to determine the meaning of changes in individual gene expression levels, especially in a biological and functional context.

In summary, Wang and colleagues (2) have identified a flaw in current subtraction methods that are widely used to identify novel expressed genes. They predict that this will limit the usefulness of current subtraction methods and result in a fall-off in gene identification rates. As a solution, they construct subtraction libraries by using short, anchored oligo(dT) primers that do not copy the long poly(A) regions of mRNAs. They present data showing that this technological improvement has the potential to improve rates of novel gene identification, which will enhance the assembly of a database that includes all expressed human genes.

We thank Drs. Kornelia Polyak, Scott McGinnis, and Todd Golub for helpful input and review of the manuscript.

1. Sager, R. (1997) *Proc. Natl. Acad. Sci. USA* **94,** 952–955.
2. Wang, S. M., Fears, S. C., Zhang, L., Chen, J.-J. & Rowley, J. D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 4162–4167.
3. Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., *et al*. (1991) *Science* **252,** 1651–1656.
4. Lal, A., Lash, A. E., Altschul, S. F., Velculescu, V., Zhang, L., McLendon, R. E., Marra, M. A., Prange, C., Morin, P. J., Polyak, K., *et al*. (1999) *Cancer Res*. **59,** 5403–5407.
5. Cohen, J. (1997) *Science* **275,** 767–772.
6. Bishop, J. O., Morton, J. G., Rosbach, M. & Richardson, M. (1974) *Nature (London)* **250,** 199–204.
7. Velculescu, V. E., Madden, S. L., Zhang, L., Lash, A. E., Yu, J., Rago, C., Lal, A., Wang, C. J., Beaudry, G. A., Ciriello, K. M., *et al*. (1999) *Nat. Genet.* **23,** 387–388.
8. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al*. (1999) *Science* **286,** 531–537.
9. Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., *et al*. (2000) *Nature (London)* **403,** 503–511.
10. Martin, K. J., Kritzman, B. M., Price, L. M., Koh, B., Kwan, C. P., Zhang, Z., Mackay, A., O'Hare, M. J., Kaelin, C. M., Mutter, G. M., *et al*. (2000) *Cancer Res.,* in press.
11. Lisitsyn, N., Lisitsyn, N. & Wigler, M. (1993) *Science* **259,** 946–951.
12. Liang, P. & Pardee, A. B. (1992) *Science* **257,** 967–971.
13. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) *Science* **270,** 484–487.
14. Nelson, P. S., Hawkins, V., Schummer, M., Bumgarner, R., Ng, W. L., Ideker, T., Ferguson C. & Hood, L. (1999) *Genet. Anal.* **15,** 209–215.
15. Wan, J. S., Sharp, S. J., Poirier, G. M., Wagaman, P. C., Chambers, J., Pyati, J., Hom, Y. L., Galindo, J. E., Huvar, A., Peterson, P. A., *et al*. (1996) *Nat. Biotechnol.* **14,** 1685–1691.
16. Martin, K. J., Kwan, C. P., O'Hare, M. J., Pardee, A. B. & Sager, R. (1998) *BioTechniques* **24,** 1018–1026.
17. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. & Watson, J. D. (1994) *Molecular Biology of the Cell* (Garland, New York).
18. Bonaldo, M. F., Lennon, G. & Soares, M. B. (1996) *Genome Res.* **6,** 791–806.
19. Boguski, M. S. (1995) *Trends Biochem Sci.* **20,** 295–296.
20. Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C., *et al*. (1999) *Proc. Natl. Acad. Sci. USA* **96,** 9212–9217.