Commentary

Navigating the human transcriptome

Robert L. Strausberg*† and Gregory J. Riggins‡

Cancer Genomics Office, National Cancer Institute, Bethesda, MD 20892; and ‡Departments of Pathology and Genetics, Duke University Medical Center, Durham, NC 27710

The potential coding capacity of the human genome is currently a topic of great interest. The number of genes predicted from the recent human-genome analysis was at the lower end of previous estimates, which had ranged between about 30,000 and 120,000 (1, 2). Whereas estimates of gene number are likely to increase based on additional experimental evidence and improved gene-finding algorithms, it is clear that gene number is only one mechanism for creating the genetic diversity required to encode the full complement of human proteins. The scientific literature richly describes the presence and functional significance of alternatively processed forms of human transcripts that are derived from different transcription initiation sites, alternative exon splicing, and multiple polyadenylation sites (3–5). Determining the various transcript forms and investigating the purpose of these complex mixtures of instructions will be the next great endeavor toward understanding human biology.

Large-scale analysis of the transcriptome originates from the expressed sequence tag (EST) concept popularized by Venter and coworkers (6). In the EST approach, clones from cDNA libraries are subjected to single-pass sequencing, such that a unique identifier is assigned to each cDNA. Initially, these tags were about 300 nucleotides in length, but sequence tags of more than 700 nucleotides are now common. Many scientists quickly realized the value of using EST sequences to identify new genes and characterize genes expressed in normal and diseased tissues. Public and private efforts soon emerged to capitalize on this opportunity. Key to the success of the public efforts was the formation of the IMAGE consortium (7), an academic-industrial partnership to distribute clones produced by the public efforts. The Merck Gene Index (MGI; ref. 8) and the Cancer Genome Anatomy Project (CGAP; ref. 9) have produced many of the human clones distributed by IMAGE. A guiding principle for these efforts was the immediate distribution of project resources: clones were distributed through the IMAGE consortium (http://image.llnl.gov/) and sequences through a GenBank division, dbEST (10).

CGAP and MGI have sought to develop a comprehensive catalog of human and mouse ESTs. Each project has produced cDNA libraries based on the use primers for first-strand synthesis that are anchored at the 3′ transcript end [the poly(A) sequence]. Sometimes, the resulting cDNA molecules represent the entire transcript. More often, they are incomplete, either because of mRNA degradation or incomplete enzymatic processing during conversion of mRNA to cDNA. Thus, for a given transcript, there might be several different forms of cDNA in the library. To facilitate gene cataloging, the MGI and CGAP sequenced the 3′ cDNA end, starting from the poly(A) sequence (see Fig. 1). With this approach, it is more likely that sequences from transcripts derived from the same gene will be recognized as such (although alternative polyadenylation sites add complication.) In the MGI project, sequencing from the 5′ clone end also was performed to provide more protein-encoding sequences (because many of the clones are incomplete, 5′ clone-end sequences often define not the true 5′ end of a transcript, but, instead, are within coding regions.).

Complementing the traditional EST approaches have been imaginative new strategies. One of these approaches, Serial Analysis of Gene Expression (SAGE; ref. 11), produces short sequence tags (usually 14 nucleotides in length) located adjacent to defined restriction sites near the 3′ end of the cDNA. Therefore, one advantage of SAGE is that, unlike the EST approach, each transcript has a unique tag, thereby facilitating transcript quantification. In addition, these tags are concatemerized, such that 30 or more gene tags can be read from a single sequencing lane, substantially increasing the efficiency of gene cataloging. However, because the tags are so short, the informatics challenges become greater. The CGAP project, working together with the National Center for Biotechnology Information (NCBI), has generated a SAGE database, SAGEmap (12). This database now includes over 4,000,000 gene tags, complementing the more than 3,700,000 human ESTs in dbEST.

In this issue, Camargo *et al*. describe a previously untried strategy, the ORF ESTs (ORESTES) approach developed by the Fundaçao de Amparo à Pesquisa do Estado de Sao Paulo/Ludwig Institute for Cancer Research (FAPESP/LICR)-Human Cancer Genome Project (13). The power of ORESTES is that a high proportion of the sequence tags are in the coding regions of transcripts (see Fig. 1). In the ORESTES approach, low-stringency PCR conditions are used to produce cDNA libraries from which a relatively small number of individual clones are sequenced. Several thousand ORESTES libraries have been produced, each with different primers, such that each library is expected to contain unique cDNA sequences. As described by Camargo *et al*., the experimental results confirm the theoretical expectations of ORESTES in that these sequences are spaced throughout the transcript, thereby providing a scaffold to complete full-length transcript sequences. Moreover, the approach has a normalization effect for a broader sampling of the many different transcripts, with less dependence on expression levels. This approach facilitates discovery of genes with low expression levels. The volume of tags is significant as well; 700,000 ORESTES represent nearly 20% of human dbEST.

As shown in Fig. 1, complete gene sequence assembly from ESTs can be quite challenging, because each tag is relatively short, and often the tags don't share homologous sequences. To organize the disparate EST sequences, the NCBI developed UniGene (14). From the outset, UniGene was designed to bin all transcript sequences from a gene into one cluster, thereby serving as a platform for transcript profiling efforts, including those based on microarrays (15, 16). However, because UniGene, in principle, groups all transcript forms of a gene into one cluster, it is too imprecise to serve as the foundation for defining the scope of the human transcriptome.

To proceed effectively with transcriptome efforts, the cDNA sequencing and clone needs will be more demanding than in the past. Currently, there is a significant shift in emphasis of the large-scale cDNA efforts toward the sequencing of complete human transcripts. In 1999, the National

See companion article on page 12103.

†To whom reprint requests should be addressed at: National Cancer Institute, 31 Center Drive, Room 11A03, Bethesda, MD 20892. E-mail: rls@nih.gov.

COMMENTARY

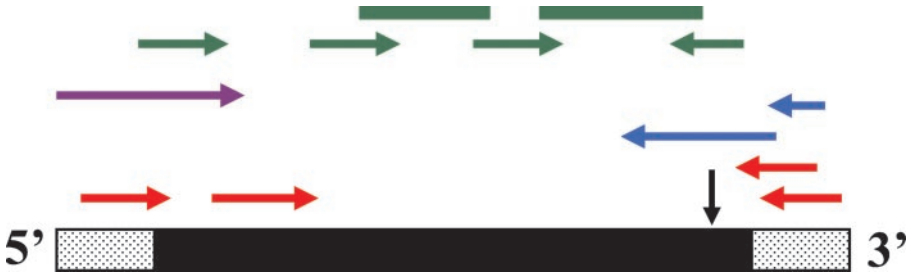

Fig. 1. Idealized schematic view of ESTs generated by various public projects. A full-length cDNA is represented by the black bar, with nontranslated regions indicated by the stippled sections. The red arrows depict 5′ and 3′ ESTs based on the MGI approach; blue arrows depict the CGAP 3′ approach. Note that at both the 5′ and 3′ ends, alternate EST positions are possible based on transcript variations or incomplete cDNA synthesis. The purple arrow indicates the MGC EST strategy in which 5′ ESTs are generated to search for full-ORF clones. ORESTES tags are shown by the green arrows, which are spaced more evenly throughout the cDNA sequence. The green bars denote regions where sequence gaps exist that might be subject to the transcript-finishing approach of Camargo *et al.* The black arrow indicates where a SAGE tag might be located. This arrow is vertical to indicate that SAGE tags are located at a precise site within a transcript.

Institutes of Health announced the Mammalian Gene Collection Project (ref. 17; http://mgc.nci.nih.gov), which is focused toward the identification and complete sequencing of human and mouse full-ORF cDNAs. To date, that project has produced over 5,000 human sequences (deposited in GenBank). In addition, the German Genome Project recently completed full-ORF human cDNA sequences derived from ≈1,500 human genes (18). Moreover, substantial progress in the sequencing of mouse full-ORF cDNAs already has been reported (19).

The transcript-finishing approach of Camargo *et al.* presented in this issue represents a valuable addition to these efforts. This strategy utilizes the ORESTES scaffold EST sequences to build primers for reverse transcription (RT)-PCR reactions to bridge gaps, thereby confirming the membership of ESTs in a common transcript and providing intervening sequence information. When combined with our increasing knowledge of the human-genome sequence (and improved gene models), the transcript-finishing approach will likely provide a convenient means for delineating the boundaries of each gene and providing complete transcript sequences.

Realizing the full potential of cDNA sequencing and annotation efforts will require the development of new schemes and databases for capturing information about each of the human transcripts. Newer databases, such as the NCBI's RefSeq (20), annotate alternatively spliced forms of transcripts. For example, RefSeq lists at least 14 distinct species of BRCA1 transcripts based on inclusion or exclusion of particular exons. Complete delineation of the human transcriptome will require a public database that is firmly rooted by the use of precise sequence-based annotation to describe all exons (including sequence variation within a particular exon) and alternate forms of processing at the 5′ and 3′ ends for each transcript species. Such a database will serve as a platform for the functional annotation of each transcript species through links to primary data sets and the scientific literature.

One of the great strengths of the human transcriptome effort is that it incorporates the talents of an international consortium and utilizes many different experimental strategies. In addition, although data and clone-release policies differ somewhat among the groups, there is an overall spirit of open sharing. Transcriptome analysis differs from the recently reported human-genome sequencing efforts in that the "biological space" of the transcriptome still remains to be defined. Many of the transcript tags annotated in dbEST and SAGEmap are from projects focused on the molecular characterization of cancer. The reasons for emphasis on cancer are clear, and it also is evident that comprehensive study of the transcriptome will require more substantial study of normal human tissues and cells, as well as those from various disease states.

Imperative to an elucidation of the transcriptome will be the development of new technologies and scientific strategies. We will need to identify and analyze not only different transcripts from a single gene, but we will also need to examine the entirety of the transcript population of cells and tissues such that we can begin to understand the networks of interactions encoded by various transcript forms. For example, the development of DNA chips that facilitate identification and quantification of specific transcripts might be used to address these questions. Undoubtedly, innovation will be a hallmark of transcriptome research for the next several years. That this creative spirit is strong is richly documented in the work of Camargo *et al.*

The authors thank Dr. Lynette Grouse for helpful discussions during the preparation of this manuscript.

1. Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001) *Nature (London)* **409,** 860–921.
2. Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001) *Science* **291,** 1304–1351.
3. Gong, Q. H., Cho, J. W., Huang, T., Potter, C., Gholami, N., Basu, N. K., Kubota, S., Carvalho, S., Pennington, M. W., Owens, I. S., *et al.* (2001) *Pharmacogenetics* **11,** 357–368.
4. Yi, X., White, D. M., Aisner, D. L., Baur, J. A., Wright, W. E. & Shay, J. W. (2000) *Neoplasia* **2,** 433–440.
5. Edwalds-Gilbert, G., Veraldi, K. L. & Milcarek, C. (1997) *Nucleic Acids Res.* **25,** 2547–2561.
6. Adams, M. D., Dubnick, M., Kerlavage, A. R., Moreno, R., Kelley, J. M., Utterback, T. R., Nagle, J. W., Fields, C. & Venter, J. C. (1992) *Nature (London)* **355,** 632–634.
7. Lennon, G., Auffray, C., Polymeropoulos, M. & Soares, M. B. (1996) *Genomics* **33,** 151–152.
8. Williamson, A. R. (1999) *Drug Discov. Today* **4,** 115–122.
9. Strausberg, R. L., Buetow, K. H., Emmert-Buck, M. R. & Klausner, R. D. (2000) *Trends Genet.* **16,** 103–106.
10. Boguski, M. S., Lowe, T. M. J. & Tolstoshev, C. M. (1993) *Nat. Genet.* **4,** 332–333.
11. Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995) *Science* **270,** 484–487.
12. 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.
13. Camargo, A. A., Samaia, H. P. B., Dias-Neto, E., Simao, D. F., Migotto, I. A., Briones, M. R. S., Costa, F. F., Nagai, M. A., Verjovski-Almeida, S., Zago, M. A., *et al.* (2001) *Proc. Natl. Acad. Sci. USA* **98,** 12103–12108.
14. Wheeler, D. L., Church, D. M., Lash, A. E., Leipe, D. D., Madden, T. L., Pontius, J. U., Schuler, G. D., Schriml, L. M., Tatusova, T. A., Wagner, L., *et al.* (2001) *Nucleic Acids Research* **29,** 11–16.
15. Luo, J., Duggan, D. J., Chen, Y. D., Sauvageot, J., Ewing, C. M., Bittner, M. L., Trent, J. M. & Isaacs, W. B. (2001) *Cancer Res.* **61,** 4683–4688.
16. 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.
17. Strausberg, R. L., Feingold, E. A., Klausner, R. D. & Collins, F. S. (1999) *Science* **286,** 455–457.
18. Wiemann, S., Weil, B., Wellenreuther, R., Gassenhuber, J., Glassl, S., Ansorge, W., Bocher, M., Blocker, H., Bauersachs, S., Blum, H., *et al.* (2001) *Genome Res.* **11,** 422–435.
19. Kawai, J., Shinagawa, A., Shibata, K., Yoshino, M., Itoh, M., Ishii, Y., Arakawa, T., Hara, A., Fukunishi, Y., Konno, H., *et al.* (2001) *Nature (London)* **409,** 685–690.
20. Pruitt, K. D. & Maglott, D. R. (2001) *Nucleic Acids Res.* **29,** 137–140.