ESTablishing a human transcript map

Mark S. Boguski & Gregory D. Schuler

National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, Maryland 20894, USA

By all indications, the Human Genome Project is about to enter its climactic phase of genomic sequencing on a massive scale[1]. Then perhaps we will finally discover the raison d'être of that 97% of the genome which does not encode proteins[2]. But long before that, perhaps in only a year or two, most human genes will be sequence-tagged and placed on various physical maps. Such a 'transcript map' (or 'expression map') of the genome will be an important part of the sequencing infrastructure, as well as a critical resource for the positional candidate approach to gene cloning[3].

One of the specific goals of the U.S. Human Genome Project is the construction of a high resolution, STS map of the genome[4]. Until recently, the great majority of STSs have been derived from anonymous genomic sequences. This has been hugely successful, efficient and elegant in its simplicity, for both genome-wide and chromosome-specific mapping. However, as James Sikela and coworkers proposed in 1991, the development of STSs from the 3' untranslated regions (3'UTR) of mRNAs has the advantage of supplying not only a marker, but also a gene for the map[5]. The two main advantages of using 3'UTRs rather than other portions of an mRNA sequence are that: (i) they almost never contain introns and thus the PCR product size is the same for cDNA and genomic DNA templates; and (ii) the sequences of 3'UTRs are not as well-conserved as coding sequences and thus it is much easier to distinguish between individual genes and paralogous gene family members that may be quite closely-related in their coding sequences.

One of the early problems with use of gene-based STSs for mapping was that there simply were not enough unique human gene sequences. All of that changed with the advent of EST sequencing[6], at which time several groups began mapping ESTs, albeit on a limited scale and usually only to the resolution of a chromosome assignment[7–9]. The paper from Sikela's group in this issue[10] reports the mapping of 318 cDNAs on the CEPH mega-YACs and is a milestone in the field, serving as a bridge between the limited EST mapping of the past and the thousands upon thousands of gene-based markers that will be appearing on maps in the near future.

The large mapping centers took up gene-based STS mapping using YACs and radiation hybrids (RH) in earnest about six months ago, spurred by the availability of 1500 human cDNA sequence fragments being produced per day and submitted to GenBank® by the Washington University Genome Sequencing Center, under contract with Merck & Co. (see Box). The goal of this project is to provide up to 400,000 ESTs for the public domain by March 31, 1996. These ESTs represent 3' and 5' sequences from approximately 200,000 oligo(dT)-primed, directionally-cloned human cDNAs. This project supplements and dramatically extends the 65,000 human EST sequences that have been placed in the public domain by other groups. Perhaps half the cost of developing an STS is in the initial sequencing, and thus the ready availability of so many cDNA 3'UTRs has made gene-based, STS mapping more attractive, economical and efficient than ever before (although this potential cost savings is obviated, to some extent, by a somewhat higher mapping failure rate with EST-based STSs compared with random genomic sequences). To facilitate collaborations and coordinate activities among different sequencing and

Box Electronic information resources

Below are World Wide Web uniform resource locators (URLs), e-mail addresses and anonymous file transfer protocol (ftp) sites for resources described in the text. Individual mapping center addresses may be found in Table 3 of ref. 16.

EST

Sequences, contact information, homology and mapping data and so on:
http://www.ncbi.nlm.nih.gov/dbEST/index.html
Data is "mirrored" at:
http://www.ebi.ac.uk/dbest/dbest_index.html
See also:
ftp://ncbi.nlm.nih.gov/repository/dbEST

ABI trace data corresponding to EST sequences, laboratory and computer protocols:
http://genome.wustl.edu/est/esthmpg.html

I.M.A.G.E. consortium library arrays:
http://www-bio.llnl.gov/bbrp/image/image.html

Other forms of access and additional information:

For homology search results and sequencing and mapping data submission and retrieval:
info@ncbi.nlm.nih.gov

For trace data and trace editing programs useful in optimizing PCR primer design:
est@watson.wustl.edu

For ordering information for physical DNA clones corresponding to ESTs and contributing libraries to the project: info@image.llnl.gov

STS

"Home page" for STS Division of GenBank:
http://www.ncbi.nlm.nih.gov/dbSTS/index.html
Data is "mirrored" at the European Bioinformatics Institute:
http://www.ebi.ac.uk/dbest/dbsts_index.html

Genome Data Base:
http://gdbwww.gdb.org/gdbdoc/topq.html

RH mapping consortium

E-mail general inquiries to datalib@ebi.ac.uk and include the word 'rhdb' in the subject line.

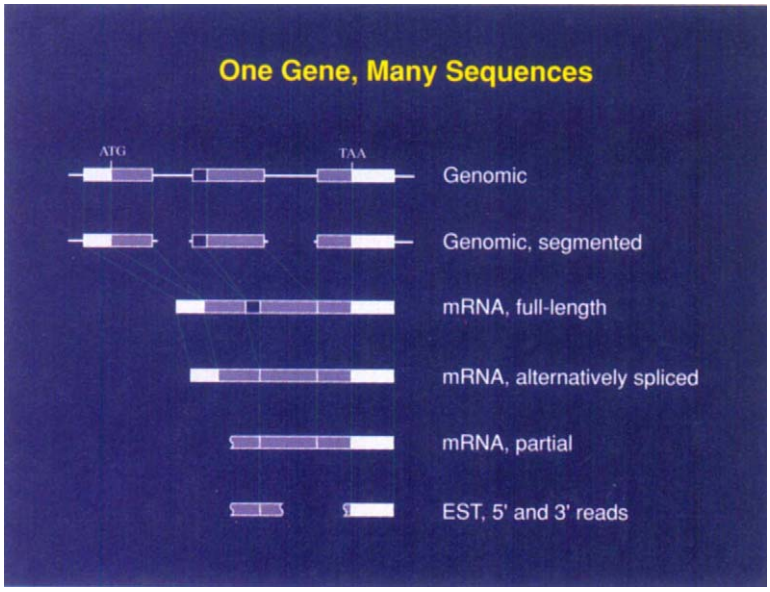

Fig. 1 One gene, many sequences. GenBank[17] is a comprehensive source of sequence data, but selecting candidates for physical mapping can be difficult. This is in large part due to the presence of multiple sequence records that, while not identical to one another, are still derived from the same gene. Shown here are some of the many forms these sequences can assume, including both genomic clones and mRNAs. Gene sequence entries possess differing amounts of flanking and intronic sequence. Sequences of mRNAs can be incomplete or contain variation because of alternative splicing. Finally, ESTs are both fragmentary and have a higher error rate. In the UniGene set, all these sequences are placed in a cluster if they share statistically significant DNA sequence similarity in the 3'UTR. For the Washington University-Merck & Co. ESTs, these sequences are derived from oligo(dT)-primed mRNA, with directional cloning and sequencing from both the 5' and 3' ends. Clustering uses only 3' ends, but use of common clone identifiers places corresponding 5' ends into the clusters.

mapping groups, the Human Genome Organization (HUGO) has been sponsoring a series of cDNA mapping workshops. Reports of the first two meetings are available[11,12].

When using ESTs for mapping, another complicating factor (apart from a higher failure rate) is that, despite the use of normalized cDNA libraries to produce the EST sequences[13], there is still a considerable degree of redundancy in the data, not to mention the overlap with more completely characterized, traditional GenBank entries representing 'functionally cloned'[3] mRNAs and genes. Thus a considerable amount of processing of the sequence data is necessary to provide novel, non-redundant mapping candidates. To address this problem, the NCBI devised a system to screen all ESTs against existing human genes and mRNAs in GenBank; the remainder were collapsed into clusters of sequences likely to be derived from the same gene. This was not a trivial task. First, GenBank is an historical archive and thus has a large degree of internal redundancy[14] and multiple representations of essentially the same data (Fig. 1). From 5,734 well-annotated human sequences in GenBank (release 88.0), a non-redundant set of 3,125 unique human 3'UTRs (referred to as the *UniGene* set) was created to serve as a source of mapping candidates itself, as well as a standard for comparison with and classification of new ESTs. Those 3' ESTs that do not match with sequences in the UniGene set are considered new human genes and redundancy among these is reduced by sequence comparison and clustering to produce unique mapping candidates. This set, referred to as *UniEST*, currently contains 14,457 clusters (derived from 33,675 3'EST sequences containing polyadenylation sites and screened for vector sequences and known repetitive elements). While many sequences in the UniGene set (as opposed to the UniEST set) have already been mapped, this was done over a number of years without consistency in methodology, resolution or annotation. It is therefore desirable that all these functionally-cloned genes also be placed on high-resolution physical maps with UniEST sequences.

Even with the UniGene and UniEST sets, and the Genexpress Index (R. Houlgatte, C. Auffray, personal communication), there is still the danger of duplicated effort if a mapping center selects an EST from these mapping sources without considering which ESTs are being mapped elsewhere. (A small amount of duplication is necessary and desirable for cross-referencing different maps.) A consortium of RH mapping groups (consisting of Whitehead Institute/MIT, Stanford, Oxford and Cambridge Universities, the Sanger Center and Généthon) has addressed this problem by depositing into a central database sequences (from any source) that they intend to map. This database, RHalloc, detects cases where two or more groups have selected the same EST for mapping and notifies those groups accordingly (Box).

If half the cost of gene-based mapping is in obtaining the sequences, another significant expense is in synthesizing the PCR primers. Most groups now working on transcript maps are bearing the cost out of their own funds, but The Institute for Genomic Research (TIGR) is in the process of donating approximately 15,000 EST-based primer pairs (of their own design, based upon publicly-available sequences) to the mapping effort[12].

So what progress toward transcript maps has been made? Members of the RH mapping consortium are each currently mapping hundreds of ESTs per month and scaling up with the goal of placing 30,000–40,000 gene-based STSs on radiation hybrid maps by the summer of 1996 (these maps will also contain genetic markers, anonymous STSs and CpG island markers). The MIT genome center is also mapping ESTs to YACs and has already released and deposited about 2,000 gene-based STSs in the public databases (Box and below). The Cooperative Human Linkage Center (CHLC) has found ESTs to be a useful source of polymorphic markers for genetic mapping.

Acknowledgements

We thank Eric Green for helpful discussions and a critical reading of the manuscript. We are indebted to L. Stein for providing the Généthon and MIT maps, to K. Buetow for providing the CHLC data, to M. Cavanaugh for extracting data from the Genome Data Base, and to J. Zhang importing the data into Chromoscope.

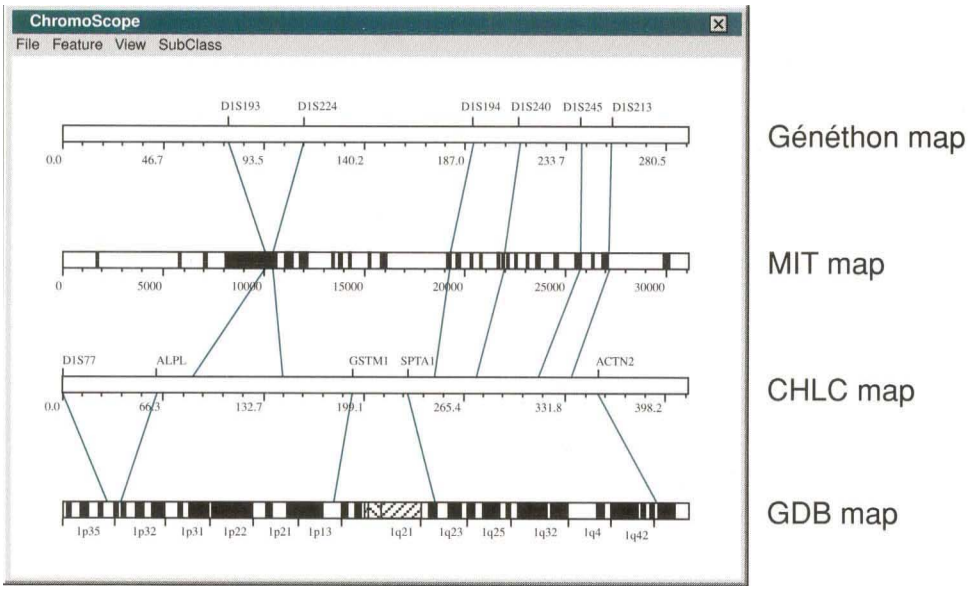

Fig. 2 Prototype graphical user interface for visualization and retrieval of mapping data derived from different mapping approaches. Different human chromosome 1 maps are shown in the Chromoscope program, a network client-server application with links to GenBank, MEDLINE, and other databases[18]. Lines drawn between maps represent common markers (it is possible to 'zoom in' and see all markers). Black rectangles in the MIT map represent YAC contigs. For the GDB map, a cytogenetic representation is shown.

Of course this data is useless without electronic information systems that permit search and retrieval over the Internet. Several of the large centers provide data releases at their own ftp and World Wide Web sites in addition to submitting it to the Genome Data Base and GenBank. GenBank has two special divisions for these data, dbEST and dbSTS, each of which contains homology information and can be queried by map location via e-mail (Box). Work is in progress on network-based, graphical user interfaces for map visualization and sequence retrieval (Fig. 2). As more and more gene-based markers accrue, homology information and links to the scientific literature will greatly amplify the value of these markers. Disease gene hunters will find such data and retrieval systems invaluable for their work.

Finally, turning once again to the sequencing phase of the genome project, the transcript map will be an important resource for genomic sequencing. To sequence the genome of the nematode, *C. elegans*[15], the strategy was to select the gene-dense regions (defined genetically) as initial targets for sequencing; the availability of a human transcript map will provide data on gene density and make it possible to use a similar strategy for sequencing human chromosomes. Another lesson drawn from the nematode project is that, even with a high gene density (one gene per 5 kb) and the presence of many fewer dispersed repeats (compared with mammals), ESTs still proved invaluable for estimating the total number of genes and for finding the exons by alignment with genomic sequences. The combination of EST sequences and map locations should be very helpful in defining genes in the vast expanses of human genomic DNA.

Even before the human 'sequence map' is densely-populated, the transcript map will shed new light on global aspects of gene organization, evolution and expression. □

1. Boguski, M.S. *Trends Biochem. Sci.* **20**, 295–296 (1995).
2. Nowak, R. *Science* **263**, 608–610 (1994).
3. Collins, F.S. *Nature Genet.* **9**, 347–350 (1995).
4. Collins, F. & Galas, D. *Science* **262**, 43–46 (1993).
5. Wilcox, A.S., Khan, A.S., Hopkins, J.A. & Sikela, J.M. *Nucl. Acids Res.* **19**, 1837–1843 (1991).
6. Adams, M.D. *et al. Science* **252**, 1651–1656 (1991).
7. Durkin, A.S., Maglott, D.R. & Nierman, W.C. *Genomics* **14**, 808–810 (1992).
8. Khan, A.S. *et al. Nature Genet.* **2**, 180–185 (1992).
9. Polymeropoulos, M.H. *et al. Genomics* **12**, 492–496 (1992).
10. Berry, R. *et al. Nature Genet.* **10**, 415–423 (1995).
11. Stewart, A. *HUGO Genome Digest* **2**, 1–4 (1995).
12. Stewart, A. *HUGO Genome Digest* **2**, 6–9 (1995).
13. Soares, M.B. *et al. Proc. natn. Acad. Sci. U.S.A.* **91**, 9228–9232 (1994).
14. Aitschul, S.F., Boguski, M.S., Gish, W. & Wootton, J.C. *Nature Genet.* **6**, 119–129 (1994).
15. Wilson, R. *et al. Nature* **368**, 32–38 (1994).
16. Murray, J.C. *et al. Science* **265**, 2049–2054 (1994).
17. Benson, D.A., Boguski, M., Lipman, D.J. & Ostell, J. *Nucl. Acids Res.* **22**, 3441–3444 (1994).
18. Zhang, J., Ostell, J. & Rudd, K.E. in *27th Hawaii International Conference on System Sciences* (ed. Hunter, L.) 58–67 (IEEE Computer Society Press, Maui, 1994).

Not all converted yet

Jonathan Howard

Institute for Genetics, University of Cologne, Zuelpicher Strasse 47, 50674 Cologne, Germany

It can be a bit misleading to use the term 'gene conversion' for events occurring in the mammalian germline. As distinct from double reciprocal crossing over, gene conversion results in a net loss of genetic information from one of the two participating duplexes, since a nucleotide sequence characteristic of one duplex (the recipient) is replaced by a homologous sequence characteristic of another (the donor). To validate formally any particular example of gene conversion it is therefore necessary to recover all the products of the crossing-over in which the event occurred to establish whether the transfer of genetic information was directional or reciprocal. While this can be done with relative ease in fungi, for example, it is a tough task if the cells in question are mammalian germ cells, and by this pedantic criterion it is clear that the experiments reported on page 407 of this issue by Zangenberg and colleagues[1] from