Finding and Mapping New Genes Faster than Ever: Revisited

James M. Sikela¹

Department of Biochemistry and Molecular Genetics, University of Colorado School of Medicine, Aurora, Colorado 80045-0508

ABSTRACT This article recounts some of the early days of the Human Genome Project, covering the important and sometimes controversial role that complementary DNA-based approaches played in the discovery and mapping of the majority of human genes. It also describes my involvement in this effort and my lab's development of methods for rapid sequence identification and mapping of human genes.

Beginnings of the Human Genome Project

The Human Genome Project (HGP) was the most ambitious biology project ever undertaken and, partly for this reason, there were many uncertainties in the early days about how best to execute it. Among the concerns were the high cost and the inadequacies of the DNA sequencing and bioinformatics capabilities available. Actually, such uncertainty provided a receptive environment for innovation. While the National Institutes of Health (NIH) and Department of Energy (DOE) worked out plans in the United States, the Human Genome Organization (HUGO) was formed to organize international efforts. One of the ideas being discussed at this time was that sequencing complementary DNAs (cDNAs) was a potentially useful strategy. This view held that cDNAs provided a smaller target than the whole genome, would cost less, and could serve as a "pilot" for full genome sequencing. In addition, cDNAs provided quick access to functionally important parts of the genome, *i.e.*, genes, and as a result were likely to provide valuable insight into human biology and disease. However, cDNA sequencing was not included as part of the initial HGP program (Pearson and Soll 1991) for reasons that were not entirely clear. For example, David Nelson of Baylor College of Medicine complained at an early HGP meeting that his grant on cDNAs was not funded because, according to NIH, "genes are not part of the purview of the Human Genome Project"!

A significant part of the enthusiasm behind the HGP proposal was that its main objective was to focus more on the technical rather than on the biological problem. Unlike other types of studies that dealt with complex cellular and molecular components with different spatial and temporal distributions, the HGP's main focus would be on obtaining only one type of information (the order of the four bases) from a linear, one-dimensional target (DNA). It would just obtain lots of it. That, together with assembling the results, was the primary goal of the project. And while a large part of the appeal was that it would be just a matter of scaling up an already proven (and Nobel Prize-winning) technology, there was also the lure of the enormity of the accomplishment: the first look at the sequence of our entire genome, the instructions for making a human being.

A Personal Journey During pre-EST Days

In 1989, as part of a Veterans Administration Schizophrenia Center research grant awarded to Bob Freedman of the Department of Psychiatry at the University of Colorado School of Medicine, I proposed to sequence human cDNAs using automated DNA sequencers. Our application was funded that year and, after I visited Applied Biosystems (ABI) in August, we used some of the funds to purchase an automated sequencer (the first in Colorado, an ABI370). Because it was a Veterans Administration (VA)-funded project, it was placed in a lab at the Denver VA hospital, across the street from my lab at the University of Colorado School of Medicine. The sequencing team that I assembled included Steve Cohen, an undergraduate who initially got things going on the ABI machine; my first graduate student,

Andrea Wilcox; lab tech Jan Hopkins; and postdoc Fred Khan, who came from Bruce Roe's lab at the University of Oklahoma. One of the best parts of being a scientist is the ability to test a big idea that has not been tried. With a brand new sequencer, grant funds to run it, and our big idea in place, we were enjoying the best science had to offer.

Virtually all DNA sequencing approaches require an oligonucleotide primer (a short piece of single-stranded DNA) to initiate DNA synthesis. The first ABI sequencers were designed to sequence into inserts of clones using M13 vector primers that were adjacent to the insert, and they were quite good at this. Going beyond such "single-pass" reads and obtaining sequences farther into the insert required considerably more effort. At that time, about the only way to obtain additional sequence information from a DNA insert was to design custom primers for each newly generated DNA sequence and "walk" along the insert, a slow and expensive approach. So it occurred to me, why not use ABI sequencers to sequence into the insert of randomly selected clones from cDNA libraries? In that way, one could quickly obtain a partial DNA sequence of many human genes, information that could serve as a type of "bar code" that would uniquely identify each gene. And with such information and the individual cDNA clones in hand, one could map human genes to the genome, although it was not obvious how mapping could be done in a high-throughput manner.

In October 1989, I attended the First Annual Human Genome Conference in San Diego and was struck by two ideas that were independently presented. No one mentioned the possibility that they could be combined. First, Sydney Brenner pushed the idea that the HGP should focus on cDNA sequencing. An audience member at the conference asked: "But Sydney, what about the regulatory regions?" Brenner responded: "We have to leave something for the next generation to work on."

Second, Maynard Olson's idea of sequence-tagged sites (STSs) was broached. An STS is a pair of PCR primers that amplify a unique segment of the human genome (Olsen *et al.* 1989). Their appeal was that they provide a PCR-based, and therefore rapid and readily transferable, genome mapping tool. The STS idea was proposed as a physical mapping strategy using genomic DNA, but nothing was mentioned about using STSs to map genes or cDNAs.

After hearing Brenner's talk, and later as I sat among the packed audience at the STS talk, I had a "Eureka!" moment that literally made me shiver: Why not make STSs from single-pass cDNA sequences? With such gene-based STSs, one could rapidly map human genes.

Gene-Based STSs for Rapid Mapping of Human Genes

After returning to Colorado I realized that using the 3'-untranslated regions (3' UTRs) of cDNAs for STSs offered major advantages, the most important of which was that 3' UTRs lack introns. This meant that the 3' UTR sequence

obtained from the cDNA would be the same as that found in genomic DNA, and therefore PCR primers designed from such cDNA sequences should work well on genomic DNA. Such primer pairs could be used to quickly map each gene to its genomic location using, for example, human-rodent somatic cell hybrids.¹

Because 3' UTRs lie adjacent to the poly(A) tail of messenger RNAs (mRNAs), they are readily accessible simply by sequencing into the insert of oligo(dT)-primed cDNA libraries. And because 3' UTRs are more species-specific than coding sequence, they are less likely to also amplify rodent DNA in the somatic cell (or radiation) hybrid. The 3' UTRs are also more gene-specific than coding regions, making it more likely that genes in a family can be distinguished from one another. Finally, using sequences immediately upstream of the poly(A) site places essentially all gene-based STSs at the same position within each cDNA, increasing the chance that only one STS is used for each gene (using sequence at the 5' ends of cDNAs would unintentionally, and unknowingly, lead to the generation of multiple STSs for the same gene—an inefficient outcome—because cDNA inserts are often truncated at the 5' end and therefore can "stop" at various positions in the same gene).

In 1989 and early 1990, excited by the potential of the idea, my lab began pursuing this approach. We started using M13 primers and our ABI370 instrument for single-pass sequencing of randomly selected cDNA clones (using brain cDNA libraries because of our interest in schizophrenia and other brain disorders). We then used the sequences to develop gene-based STSs from 3' UTR regions, followed by mapping the genes by PCR (Figure 1). To our knowledge, no one had done this before, so many things needed to be worked out. For example, in the past cDNA libraries were mainly screened, usually by probing plaques of phage clones, to identify one or a few relevant clones. In contrast, this new approach picked cDNA clones randomly. When we first tried the method and got nothing but a long run of "A's" from many clones, we thought that we were not running the ABI machine in the right way. Later, we realized that the clones were the problem as many contained nothing but poly(A), revealing the low quality of some commercially available libraries. While many clones did include mRNA sequences, they were often upstream of long stretches of poly(A), which interfered with obtaining meaningful sequence data. So we developed an anchored oligo(dT) primer method that allowed one to obtain cDNA sequence

¹Human-rodent somatic cell hybrids are a panel of rodent (typically hamster) cell lines, each of which contains a known subset of human chromosomes. For gene mapping, DNA from the panel is interrogated with gene-specific primers for the presence or absence of the human PCR product. Comparison of these results with a previously generated "key," indicating which human chromosomes are in which hybrid cell line, results in the identification of the specific chromosome on which the gene is located. Radiation hybrids that contain fragmented pieces of human chromosomes that are generated by radiation exposure provide higher-resolution mapping. YAC libraries, containing large pieces of human genomic DNA gridded onto microtiter plates, could also be used for PCR-based mapping. And pooling strategies enable mapping a gene to a specific YAC clone with as few as 32 PCR reactions.

Figure 1 Diagram of the strategy used for initial identification and mapping of the great majority of all human genes. Single-pass sequences from randomly chosen cDNA clones are generated. These ESTs can be from either the 5' end or the 3' end of the cDNA insert (or both). To make the human gene map, 3' sequences were used because they allowed one to easily obtain sequence from the 3' UTR of the cDNA. 3' UTR sequences were ideal for development of gene-based STSs, which could be used for high-throughput, PCR-based mapping of each gene.

directly upstream of the poly(A) tail regardless of tail length (Khan *et al.* 1991). The approach solved the poly(A) problem, but it needed to be adapted to ABI's automated sequencers, which had been optimized for M13 vector primers.

Competition

At a September 1990 meeting in Steamboat Springs of the University of Colorado Alcohol Research Center (ARC), I gave a talk that laid out the approach and what we had accomplished so far. Afterward, David Goldman, an ARC advisor who was an intramural investigator at the NIH, commented that he thought that his NIH colleague Craig Venter had just begun to sequence cDNAs. This was the first time I heard that someone else was carrying out a similar project. Since no articles had yet been published on the approach, we wrote up some of what we had accomplished and published the proof-of-principle in *Nucleic Acids Research* in April 1991 (Wilcox *et al.* 1991). Our article laid out the major steps of the cDNA sequencing–mapping approach, including the 3' UTR gene-based STS idea as well as an efficient cDNA library normalization step. While we had generated numerous sequence reads using the ABI sequencer with M13 vector primers, the process was not yet very efficient due to the poly(A) length problem. As a result, while we included some single-pass cDNA sequence data in the article, the article focused primarily on the gene-mapping strategy. Two months after our article was published, Venter published his findings, dubbing these single-pass cDNA sequences “expressed sequence tags” or ESTs (Adams *et al.* 1991).

The Field Heats Up

This work was now drawing much more attention, and with it came controversy. There were concerns that it would divert resources (funds) from the effort to sequence the whole human genome: if cDNAs were “done” by the HGP,

there would be less motivation to provide the funds necessary to carry out the more difficult and more expensive work of obtaining the sequence of the genome.

Despite these concerns, the DOE decided to fund a major cDNA initiative to sequence and map cDNAs as part of their HGP effort. Because we had already been doing this, we were awarded a DOE grant (started in October 1991) to sequence and map more human brain cDNAs. Several other groups received grants, and some of the divisions of labor were the following: my lab and Venter's were the primary sequencing groups; Bento Soares at Columbia University would focus on cDNA library development; my lab and that of another early pioneer of mapping ESTs, Mihaelis Polymeropoulos at NIH, would map genes; and a cDNA clone repository would be set up by the American Type Culture Collection.

The sequencing–mapping effort was significantly aided by several key improvements in cDNA library construction that came from the Soares lab (Soares 1994; Soares *et al.* 1994). These improvements included (1) making “normalized” cDNA libraries that allowed genes to be identified more efficiently (*i.e.*, normalization reduced the representation of cDNA clones of abundantly expressed genes); (2) making cDNA libraries from multiple human tissues to allow access to more genes; and (3) making directionally cloned cDNA inserts so one could choose to sequence only the 3' end (ideal for mapping) or the 5' end or both. Finally, the poly(A) problem was minimized by the use of saturating amounts of oligo(dT) primers to reverse-transcribe cDNA from poly(A)⁺ mRNAs to ensure that only the primer that sat nearest the 3' UTR would be extended into the 3' UTR. In this way, only a few A's next to the 3' UTR were included in the cDNA clone, allowing sequencing into the 3' UTR using the same M13 vector primer for each clone.

In 1992 we published an article in *Nature Genetics* (Khan *et al.* 1992) that presented >1000 single-pass cDNA sequences using the new Soares libraries and the physical mapping of many of the genes. We also showed that cDNA sequences containing “CA repeats” (polymorphic genetic markers) could be genetically mapped, providing anchor points that allowed integration of all three types of genome maps: gene map, physical map, and genetic map.

While these events were unfolding, Venter chose to file patents on the partial cDNA sequences (ESTs) that his lab was generating. Not surprisingly, commercial interest in the field became substantial, with multiple companies (The Institute for Genomic Research/Human Genome Sciences, Incyte) being formed primarily to focus on cDNAs. Other academic groups, notably the labs of Kenichi Matsubara from Japan and Charles Auffray from France, had also begun to sequence human cDNAs, and HUGO organized a number of international meetings to allow researchers to compare strategies and results of this fast-developing field (Sikela and Auffray 1993).

Since my lab, like Venter's, was funded by the DOE cDNA effort, the question came up as to whether we

would also file patents on our cDNA sequences. DOE's advice was that no stigma would be associated with either choice we made. We decided to release all our cDNA sequences to public databases with no strings attached. Subsequent events related to Alzheimer's disease vindicated our decision.

Discovery of Presenilin 2 Validated the cDNA Approach

In 1995 Nikki Walter, who along with T. J. Stevens were my main genome-focused lab techs, informed me that people from three different labs had called her on the same day to ask for the same cDNA clone. She asked the last caller why they were interested, and they said they worked on Alzheimer's disease (ALZ) and wanted the clone because they thought it was relevant to ALZ. But that was all they would say.

A few weeks later Rudy Tanzi and Gerald Schellenberg published their discovery of a new ALZ disease gene, Presenilin 2 (PSEN2) (Levy-Lahad *et al.* 1995). They mentioned in the article that they found the gene by using the recently discovered PSEN1 gene sequence (Sherrington *et al.* 1995) to search a public DNA database and hit a brain EST my lab had submitted. The search indicated that the EST was related, but not identical, to PSEN1 and therefore was likely to be a "sister gene." There were some ALZ families from the Volga River region in Russia in which the PSEN1 gene was normal, and the Tanzi lab immediately suspected that our EST corresponded to the gene that was mutated in the Volga River families. They were right, and after some quick follow-up analysis, they confirmed that it was a new ALZ gene.² All three labs that called Nikki that day must have carried out the same BLAST search using the PSEN1 sequence, the first day that the PSEN1 sequence became public. Tanzi's lab, which already had access to the PSEN1 gene sequence, had done that weeks previously, hit our EST, and "won" the race to find Presenilin 2. This demonstrated how ESTs could be used to easily discover important new disease genes. A computer search of a public gene sequence database, which took only a few seconds, was all it took to find the likely culprit.

"A Milestone in the Field"

While Venter had decided to keep all his EST sequences private and explore their commercialization, we continued to make ours publicly available. We had been fine-tuning the steps in the cDNA sequencing–gene-mapping pipeline for several years, and by the mid-1990s it was accurate and streamlined. The optimized cDNA libraries allowed ABI sequencers to easily provide 3' UTR sequences from many cDNA clones. Resulting gene-based STSs could be generated and efficiently mapped to both radiation hybrids and YAC

clones, resources that would allow genes to be mapped to specific chromosomal regions. In 1995 my lab published an article that showed how this optimized gene-based STS approach could be scaled up to generate a human gene map (Berry *et al.* 1995). A companion commentary (Boguski and Schuler 1995) called our article a "milestone in the field," and also described a coordinated plan to scale up the method in a big way.

Scaling Up the Approach

cDNAs had become such a major area of interest that an international effort was organized to scale up cDNA-based gene finding and mapping, with everything made public and freely available. My lab and others from the United States, Europe, and Japan signed on to help. There were two main objectives:

1. To scale up single-pass cDNA sequencing (EST sequencing) to "identify" most human genes. The consortium would carry out automated, single-pass sequencing of randomly selected human cDNAs on a large scale, bankrolled primarily by the drug company Merck. All sequence data would immediately be put into public databases, no strings attached, with no patent applications.
2. To scale up the gene-mapping approach that I had developed to generate the first comprehensive human gene maps. Gene-mapping groups from around the world would use gene-based STSs, derived from 3' UTRs of cDNAs, to map each gene to a chromosomal location, primarily using panels of human–rodent radiation hybrid cell lines and human YAC libraries. I had previously pitched the idea of making only one STS for each gene as "one gene = one STS," and it was made part of the large-scale mapping strategy. To avoid duplication of effort, these gene-specific STS primer sequences would be centrally distributed, *e.g.*, by the National Center for Biotechnology Information, to groups doing the mapping. Gene-mapping data and STS sequences would quickly be deposited into public databases.

This international effort was essentially a culmination of the time and effort that had gone into developing and optimizing the approach, and it worked extremely well. After only a year of operation, the consortium's efforts resulted in the identification (*i.e.*, sequence-tagging) and mapping of >15,000 human genes. The results were published in the annual genome issue of *Science* in 1996 (Schuler *et al.* 1996) (<http://www.ncbi.nlm.nih.gov/science96/>). Two years later the consortium reported a total of 30,000 genes mapped using the cDNA-based strategy (Deloukas *et al.* 1998) (<http://www.ncbi.nlm.nih.gov/genemap99/>). The cDNA ride was over, but it was satisfying to know that my lab had played a significant role in the effort that had identified and mapped most human genes.

My parents never had the opportunity to attend college, but without their support and encouragement of my

²This story is also described in more detail in two books: *Decoding Darkness* by Rudy Tanzi and *Cracking the Genome* by Kevin Davies.

interest in science and DNA I would never have had the chance to contribute to the HGP. My mother, who helped me manually read thousands of nucleotides of DNA sequence from autoradiograms as I finished my Ph.D. thesis, passed away in 1992, so unfortunately I was not able to share much of this with her. When our 1995 article (Berry *et al.* 1995) came out, I sent my father a copy with an attached sticky note that said: "It looks like a method we developed is going to be used to make the world's first detailed human gene map."

While a sequence of the human genome followed in only 5 years, the millions of partial cDNA sequences that had been generated by the large-scale cDNA effort continued to be important, providing a resource for identifying genes and exons in the human genome.

Literature Cited

- Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos *et al.*, 1991 Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252(5013): 1651–1656.
- Berry, R., T. J. Stevens, N. A. Walter, A. S. Wilcox, T. Rubano *et al.*, 1995 Gene based sequence-tagged sites (STSs) as the basis for a human gene map. *Nat. Genet.* 10: 415–423.
- Boguski, M. S., and G. D. Schuler, 1995 ESTablishing a human transcript map. *Nat. Genet.* 10: 369–371.
- Deloukas, P., G. D. Schuler, G. Gyapay, E. M. Beasley, C. Soderlund *et al.* 1998 A physical map of 30,000 human genes. *Science* 282: 744–746.
- Khan, A. S., A. S. Wilcox, J. A. Hopkins, and J. M. Sikela, 1991 Efficient double stranded sequencing of cDNA clones containing long poly(A) tails using anchored poly(dT) primers. *Nucleic Acids Res.* 19: 1715.
- Khan, A. S., A. S. Wilcox, M. H. Polymeropoulos, J. A. Hopkins, T. J. Stevens *et al.*, 1992 Single pass sequencing and physical and genetic mapping of human brain cDNAs. *Nat. Genet.* 2: 180–185.
- Levy-Lahad, E., W. Wasco, and P. Poorkaj, D. M. Romano, J. Oshima *et al.* (1995) Candidate gene for the chromosome 1 familial Alzheimer's disease locus. *Science* 269 (5226): 973–977.
- Olson, M., L. Hood, C. Cantor, and D. Botstein, 1989 A common language for physical mapping of the human genome. *Science* 245(4925): 1434–1435.
- Pearson, M. L., and D. Söll, 1991 The Human Genome Project: a paradigm for information management in the life sciences. *FASEB* 5: 35–39.
- Schuler, G. D., M. S. Boguski, E. A. Stewart, L.D. Stein, G. Gyapay *et al.* 1996 A gene map of the human genome. *Science* 274: 540–546.
- Sherrington, R., E. I. Rogaev, Y. Liang, E. A. Rogaeva, G. Levesque *et al.* 1995 Cloning of a gene bearing mis-sense mutations in early-onset familial Alzheimer's disease. *Nature* 375 (6534): 754–760.
- Sikela, J. M., and C. Auffray, 1993 Finding new genes faster than ever. *Nat. Genet.* 3: 189–191.
- Soares, M. B., 1994 Construction of directionally cloned cDNA libraries in phagemid vectors, pp. 110–114 in *Automated DNA Sequencing and Analysis*, edited by M. D. Adams, C. Fields, and J. C. Venter. Academic Press, New York.
- Soares, M. B., M. F. Bonaldo, P. Jelenc, L. Su, L. Lawton *et al.*, 1994 Construction and characterization of a normalized cDNA library. *Proc. Natl. Acad. Sci. USA* 91: 9228–9232.
- Wilcox, A. S., A. S. Khan, J. A. Hopkins, and J. M. Sikela, 1991 Use of 3' untranslated sequences of human cDNAs for rapid chromosome assignment and conversion to STSs: implications for an expression map of the genome. *Nucleic Acids Res.* 12: 1837–1843.

Communicating editor: A. S. Wilkins