

Electronic PCR: bridging the gap between genome mapping and genome sequencing

Gregory D. Schuler

A crucial event in the history of the Human Genome Project was the decision to use sequence-tagged sites (STSs) as common landmarks for genomic mapping. Following several years of constructing STS-based maps of ever-increasing detail, the emphasis has recently shifted towards large-scale genomic sequencing. A computational procedure called 'electronic PCR' allows STS landmarks to be revealed as data emerge from the sequencing pipeline, thereby bridging the gap between mapping and sequencing activities.

In the late 1980s, as the challenge of mapping and sequencing the human genome was being contemplated, Olson and colleagues put forth the visionary proposal of using sequence-tagged sites (STSs) as landmarks of the genome, thereby providing a 'common language' for unifying the mapping efforts taking place in numerous laboratories around the world¹. An STS consists of a pair of short oligonucleotide primer sequences, together with the precise polymerase chain reaction (PCR) conditions for using them to detect a site that is unique within the genome. By using the DNA sequence itself to define genomic landmarks, we can circumvent the 'language barrier' caused by multiple synonymous identifiers being attached to the same marker. More importantly, an STS is free of context within any particular cloned DNA segment, which obviates the need to maintain and distribute any biological materials. All of the information needed to recover a genomic landmark can be stored in a computer database and transmitted electronically to any laboratory in the world.

Because STSs are defined by their PCR-primer sequences, these landmarks of the genome can be recovered in DNA sequences through a computational strategy called 'electronic PCR'² (e-PCR). Briefly, the procedure involves searching for subsequences that match the two primers and then demanding that they be in the correct order, orientation and spacing to give rise to a PCR product of the correct molecular weight (Fig. 1). In order to tolerate both natural variation and sequencing error, a 'margin' parameter can be used to specify the maximum amount by which the computed PCR-product size may deviate from the known true size. A program that implements this basic process has been developed (available at <ftp://ncbi.nlm.nih.gov/pub/schuler/e-PCR/>) and a World Wide Web resource is available for the on-line e-PCR analysis of user-supplied DNA sequences (<http://www.ncbi.nlm.nih.gov/cgi-bin/STS/nph-sts>). These implementations make use of a computer trick known as 'hashing' to speed up the search for the primers sequences significantly². As a result, it is possible to compare an

entire database of STSs to a database of DNA sequences in a reasonable period of time. By contrast, a general-purpose program such as BLAST³ would perform the sequence comparisons more slowly, require post-processing of the results (to enforce the spacing of the primers) and be prone to reporting false results (especially when searching for STSs based on repeated DNA, such as microsatellites). Various freely available pattern-matching programs (e.g. using 'regular expression' strategies) could also be used, but they are slower still, although less susceptible to reporting false positives.

Genome mapping

For the first five years of its official existence, the Human Genome Project has focused on the construction of increasingly detailed physical maps of the genome. Two STS-based physical maps of the entire human genome have been developed, by the Whitehead Institute⁴ and by the Stanford University Genome Center⁵. These maps are based partly (Whitehead) or entirely (Stanford) on radiation-hybrid (RH) methodology, in which the frequency of X-ray-induced DNA breakage between two STSs is used to estimate the distance between them. In addition, several chromosome-specific maps have been developed, mostly using overlapping yeast artificial chromosome (YAC) inserts to order the STSs⁶⁻¹¹. Although these maps have many applications, one hope was that they would help to define 'sequence-ready' reagents: sets of ordered clone inserts with maximal coverage and minimal redundancy. Unfortunately, most existing maps do not have sufficient resolution to be used for this purpose. Because the clones typically used as sequencing substrates accommodate inserts of 100–200 kb, an average map resolution of about 50 kb would be needed to determine the relationships between these clones easily. Nevertheless, sequencing is already under way for those targets with the most detailed regional or chromosome-specific maps.

One of the most compelling reasons for mapping and sequencing the human genome is to facilitate the identification of genes responsible for inherited human disorders, and a high-resolution genetic map is an indispensable tool for this task. A significant milestone is the completion of the Génethon genetic map, which contains 5264 STSs developed from microsatellite

G. D. Schuler (schuler@ncbi.nlm.nih.gov) is at the National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA.

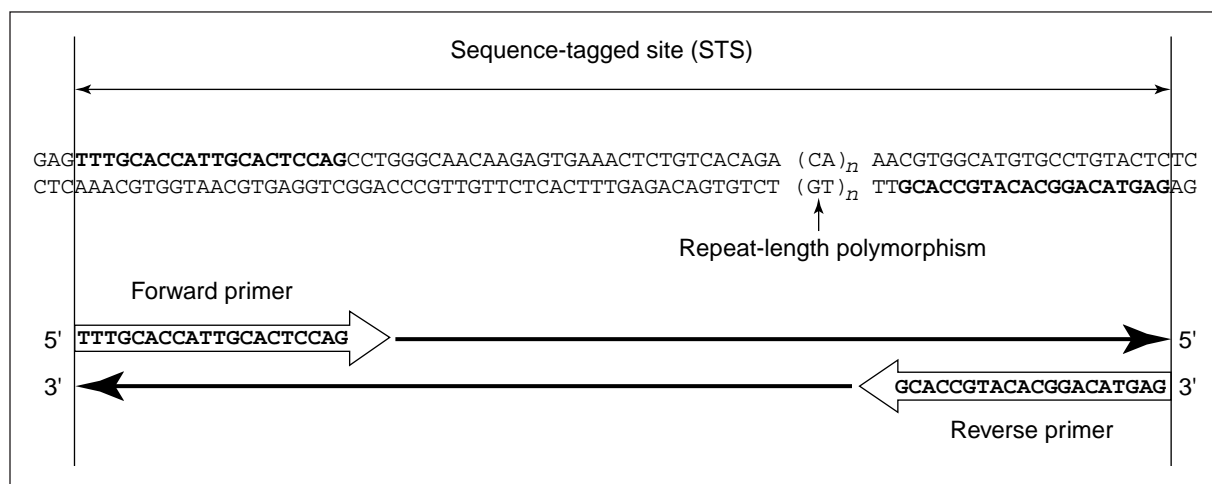


Figure 1

A sequence-tagged site (STS) is defined by a pair of short oligonucleotide primers pointing in opposing directions, so that successive cycles of primer-annealing and -extension will give rise to a detectable product of a specific size, determined by the spacing of the primer-binding sites. In order to be successfully mapped, an STS should occur only once in the genome. When using e-PCR to find the site in a DNA sequence, it should be allowed to be in either orientation. A successful match would thus include either the forward primer followed by the inverse of the reverse primer or the reverse primer followed by the inverse of the forward primer. In either case, the distance between the outer termini of the primers must be within some specified margin of the known PCR product size.

sequences¹². This abundant class of microsatellite sequence contains $(CA)_n$ dinucleotide repeats, which have the property that the number of repeating units varies between different alleles. By choosing PCR primers from unique sequences flanking the repeats, the polymorphic character of the locus is revealed by allele-specific differences in the molecular weight of the PCR product (Fig. 1). Polymorphic STS are essential for genetic mapping, which relies on the ability to track the inheritance of genetic differences. The 'margin' feature of the e-PCR program, even at its default setting, should allow nearly all polymorphic STSs to be detected, despite these natural variations in PCR-product size.

Using a genetic map, the gene responsible for a certain phenotype of interest may be localized to an interval flanked by two genetic landmarks, but the experimental techniques traditionally used to identify genes in such an interval are expensive, labor-intensive and time-consuming. High-resolution gene maps are of particular benefit for this enterprise because, by providing regional inventories of genes, they may in some cases obviate the need for experimental gene-finding methods. An international consortium of RH-mapping laboratories has recently produced a gene map containing 20 128 cDNA-based markers representing roughly 16 000 distinct genes¹³. An update to this work, in which the number of mapped genes has been roughly doubled, has recently been completed¹⁴. Two RH panels (Genebridge 4 and Stanford G3) were used for localizing genes, together with approximately 1000 Génethon microsatellite markers that were used to provide a common framework for data integration. Moreover, the inclusion of microsatellite markers is of particular importance because it immediately allows the entire gene map to be integrated with the genetic map, which is in keeping with its expected application in supporting positional-cloning activities.

Genome sequencing

Although improvements to the maps are continuing, the past three years have seen a major shift in emphasis from mapping to sequencing. In the USA, the NIH has funded several genome centers to develop methodology for very-large-scale sequencing. In the UK, similar work is well under way at the Sanger Centre, with support from the Wellcome Trust. Additional genome-sequencing projects are evolving in Germany, France and Japan. Even though this renewed focus on sequencing is only a few years old, a substantial amount of data has already been generated. As of 31 December 1997, there were 1147 finished human-genomic-sequence entries in GenBank, with lengths totaling 78.5 Mb. Despite these advances, however, it is clear that further increases in sequencing capacity will be needed to complete the entire genome of roughly 3000 Mb by the target date of 2005.

The use of e-PCR in the rapid, automatic analysis of genomic sequences will become increasingly important as the rate of data accumulation continues to increase. Finding mapped STSs in these sequences allows entries to be classified by their position in the genome. As a demonstration of this strategy, the 78.5 Mb of available human genomic DNA was subjected to e-PCR analysis to test for the presence of STSs from four whole-genome maps (Table 1). For example, when the 1147 GenBank entries were tested against the Génethon genetic map, STSs were found in 124 of them (11%); these entries contained 17% (13.2 Mb) of the sequenced human DNA. In general, greater amounts of DNA could be anchored to those maps containing the most STSs (e.g. the Whitehead physical map). When all the maps were used together, 44% of the GenBank entries contained an STS from at least one map, which allowed 45.6 Mb (58% of the total) to be positioned on the genome. These values are less than the sums of the individual maps because of the fact that many STSs have been included on

Table 1. Proportion of genomic-sequence data that can be anchored to a genomic map

Map	Entries	Percentage anchored	Megabases	Percentage anchored
Généthon genetic map	124	11	13.2	17
Whitehead physical map	417	36	39.6	50
Stanford physical map	191	17	19.7	25
RH Consortium gene map	294	26	25.9	33
Any of the above	510	44	45.6	58

Table 2. An estimate of progress in human genome sequencing

Chromosome	Size (Mb)	DNA sequenced		Généthon coverage	
		Mb	Percentage	Sites	Percentage
1	263	1.6	0.6	3	0.7
2	255	0.6	0.2	0	0.0
3	214	0.8	0.4	2	0.6
4	203	2.2	1.1	3	1.1
5	194	1.7	0.9	5	1.6
6	183	4.7	2.6	11	3.5
7	171	14.4	8.4	27	9.9
8	155	0.3	0.2	1	0.4
9	145	1.8	1.2	2	1.1
10	144	0.5	0.3	2	0.7
11	144	4.0	2.8	6	2.2
12	143	1.2	0.9	5	2.0
13	98	1.7	1.7	4	2.4
14	93	1.2	1.3	3	1.9
15	89	0.4	0.4	1	0.7
16	98	5.6	5.7	9	5.0
17	92	1.9	2.0	7	3.8
18	85	0.0	0.0	0	0.0
19	67	2.5	3.7	6	5.0
20	72	0.1	0.2	1	0.7
21	39	0.7	1.8	0	0.0
22	43	9.9	23.1	16	23.9
X	164	20.1	12.3	24	11.1
Y	28	0.5	1.9	n/a	n/a
Total	3182	78.5	2.5	138	2.6

several maps to serve as reference points for map integration.

At first glance, this analysis does not seem to be much of a benefit, considering that a sequence-ready map is generally used to select clones for sequencing in the first place. However, a small number of submitted sequences have, in fact, been received without any form of map information. Furthermore, in the current set of entries, e-PCR analysis shows that five sequence entries (0.4%) have either incorrect chromosome assignments or are chimeric, as evidenced by the presence of multiple STSs that localize to a different chromosome on at least two genomic maps (these errors are routinely communicated to the relevant genome centers).

As human-genomic-sequence data continue to accumulate, it will become more important to have a means

of monitoring the progress that has been made on each of the human chromosomes. Until more-sophisticated procedures are available, two simple estimates of current progress are offered here (Table 2). The first is to sum the lengths of the sequences assigned to each chromosome and divide by estimates of chromosome sizes drawn from the literature¹⁵. This approach will tend to overestimate the true progress because overlaps at the ends of the sequences (typically 50–200 bases, but occasionally very large) are not taken into account. In addition, this measure may be affected by systematic error in the chromosome-size determinations.

A second approach is to make use of the e-PCR results generated above by calculating the fraction of STSs from a particular chromosome that have been found in a genomic sequence. Using the Généthon marker set, chromosome 22 shows the greatest apparent degree of completion, with 23.9% of the STSs (16 of 67) covered by the available sequence data (Table 2). This estimate compares favorably with the 23.1% completion figure based on summing the lengths of the GenBank sequence entries. Other chromosomes showing at least 5% completion are those with whole-chromosome-sequencing projects well under way: chromosomes 7, 16, 19 and X. Overall, it appears that about 2.6% of the human genome has been sequenced. This estimate based on STS coverage is not affected by sequence redundancy (if a site appears in multiple sequences, it is counted only once) and is independent of chromosome-size estimates (being based only on the number of STSs for each chromosome). On the other hand, it is a coarser estimate owing to the small number of STSs involved and their uneven distribution. In addition, a possibility for bias exists, should the choice of sequencing targets be influenced by the availability of mapped markers. However, in most cases, targets are chosen for other properties (e.g. gene density) and STSs are developed as needed. The Généthon marker set was chosen for the experiment described here because the STSs are not derived from genes, thereby avoiding the potential bias of gene-rich regions being selected for early sequencing. However, by making this choice, the Y chromosome cannot be assessed because there is no genetic map for this chromosome (owing to a lack of meiotic recombination). Despite the fact that each of the approaches shown in Table 2 has its strengths and weaknesses, the results obtained are remarkably similar.

Bridging the gap

The common language of STSs provides the ability to correlate maps of ever-increasing resolution, and the use of e-PCR to detect STSs in DNA sequences carries this process to its ultimate conclusion. The map with highest possible resolution is the genomic sequence itself but, well in advance of assembling the final genomic reference sequence, localized regions can be brought into focus as they are completed. Within a sequenced region, exact base-pair distances between STSs can be determined. One application of this is to provide estimates of the accuracy of existing physical maps; calculations of errors in the sequenced segments can be extrapolated to the map as a whole. Furthermore, it becomes easier to compare maps with each other. The usual barrier to this process is the paucity of common markers (having a common language is only of benefit if the parties use

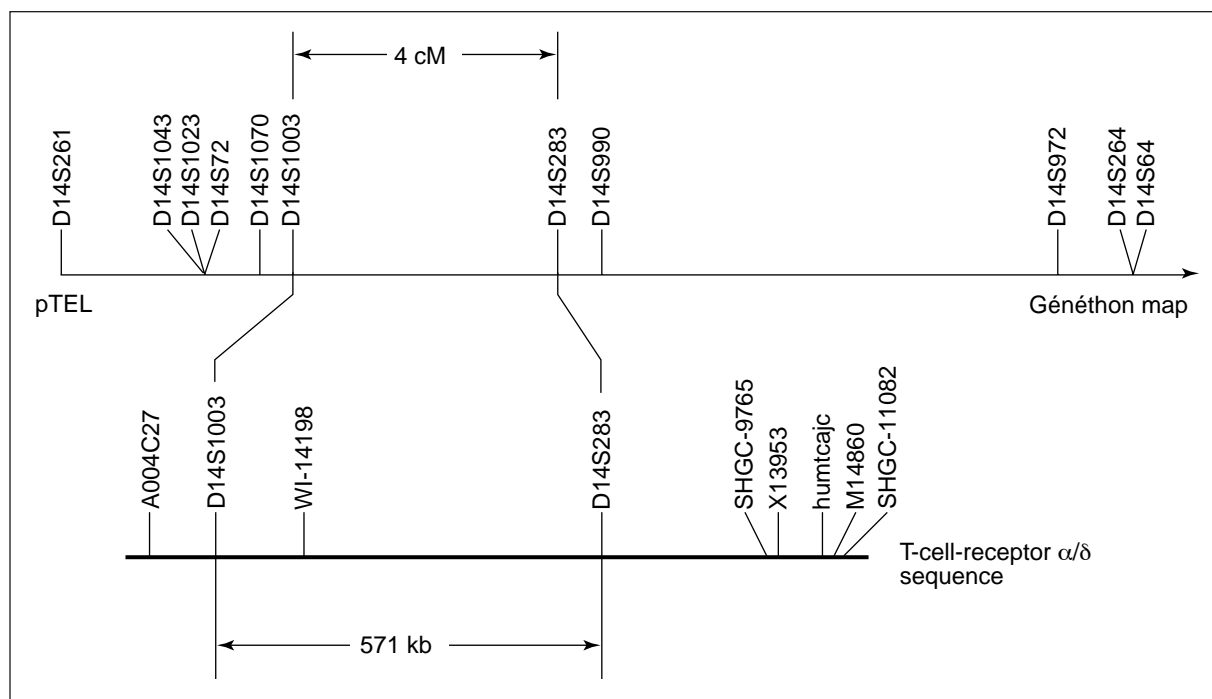


Figure 2

Placing the T-cell-receptor alpha-delta-gene sequence on the genetic map. The line drawn for the TRC alpha-delta genomic sequence represents a 1.07 Mb composite sequence constructed from five GenBank entries (AE000658–AE000662). Two genetic markers (D14S1003 and D14S283) found in this sequence by e-PCR analysis allow the sequence to be positioned and oriented with respect to the Génethon map of chromosome 14. A 16.3 centi-Morgan portion of this map is shown at the top of the figure, originating near the short-arm telomere (pTEL) and continuing rightward towards the centromere. The remaining markers shown on the sequence are cDNA-based markers, several of which correspond to the known positions of the alpha and delta constant-region genes.

at least some of the same words!) but, given the sequence, STSs from all maps can be localized by e-PCR and related to one another. These map comparisons are of particular interest when relating a genetic map to a physical map because it becomes possible to discern 'hot' and 'cold' spots of genetic recombination.

A useful illustration of the utility of correlating map and sequence data is provided by the sequence of the T-cell-receptor (TCR) alpha-delta-gene cluster¹⁶. The five GenBank entries spanning this region (AE000658–AE000662) were assembled to form a 1 071 650-base composite sequence, which was then subjected to e-PCR analysis using the STSs from the Génethon genetic map and the RH-consortium gene map. Two genetic markers (D14S1003 and D14S283) were found, allowing the sequence to be positioned and oriented with respect to the Génethon map of chromosome 14 (Fig. 2). Despite being separated by a genetic distance of 4 centi-Morgans, these markers are only 571 kb apart in the sequence, indicating that the region is a 'hot spot' for meiotic recombination. In addition, six cDNA-based markers were found, several of which correspond to the known positions of the TCR alpha and delta constant-region genes.

The end product of the Human Genome Project is a complete reference sequence of the human genome. Although only 2.5% of the bases have yet been sequenced, it is not too early to begin to assemble this reference sequence. The human sequence map presented in the Entrez Genomes division is an attempt to do exactly this (<http://www.ncbi.nlm.nih.gov/Entrez/>). Using the STSs they contain, islands of contiguous sequence have been positioned using two physical maps

(from Whitehead and Stanford) and two genetic maps [from Génethon and the Cooperative Human Linkage Center (CHLC)]. It can be anticipated that e-PCR will continue to play a role in both the assembly and the validation of the human reference sequence as it approaches completion.

Acknowledgments

Thanks to M. Boguski for his critical review of the manuscript. The software for constructing a composite sequence for a clone contig was provided by J. Zhang, who is also involved in maintaining the integrated human maps and sequences found in the Entrez Genomes division. The WWW resource for performing e-PCR was developed by S. Shavirin.

References

- 1 Olson, M., Hood, L., Cantor, C. and Botstein, D. (1989) *Science* 245, 1434–1435
- 2 Schuler, G. (1997) *Genome Res.* 7, 541–550
- 3 Altschul, S. F. *et al.* (1997) *Nucleic Acids Res.* 25, 3389–3402
- 4 Hudson, T. J. *et al.* (1995) *Science* 270, 1945–1954
- 5 Stewart, E. A. *et al.* (1997) *Genome Res.* 7, 422–433
- 6 Gemmill, R. M. *et al.* (1995) *Nature* 377, 299–319
- 7 Bouffard, G. G. *et al.* (1997) *Genome Res.* 7, 673–692
- 8 Qin, S. *et al.* (1996) *Proc. Natl. Acad. Sci. U. S. A.* 93, 3149–3154
- 9 Krauter, K. *et al.* (1995) *Nature* 377, 321–333
- 10 Doggett, N. A. *et al.* (1995) *Nature* 377, 335–365
- 11 Collins, J. E. *et al.* (1995) *Nature* 377 (suppl.), 367–371
- 12 Dib, C. *et al.* (1996) *Nature* 380, 152–154
- 13 Schuler, G. D. *et al.* (1996) *Science* 274, 540–546
- 14 Deloukas, P. *et al.* *Science* (in press)
- 15 Morton, N. E. (1991) *Proc. Natl. Acad. Sci. U. S. A.* 88, 7474–7476
- 16 Boysen, C., Simon, M. I. and Hood, L. (1997) *Genome Res.* 7, 330–338