Jpn J Human Genet **38**, 381–390, 1993

A MATHEMATICALLY DESIGNED STS PRIMER WITHOUT ANY MISMATCHES FOR DIRECT SEQUENCING OF COSMID DNA CLONES

Xiaoren TANG,[1] Yifei WANG,[2] Yasuhiko NAKATA,[1]
Hai-Ou LI,[1] Akiko FUJITA,[1] Hui GAO,[1]
Akinori SARAI,[2] and Kazushige YOKOYAMA[1,*]

[1]*DNA Bank and* [2]*Genetic Information Bank, Tsukuba Life Science Center, RIKEN (The Institute of Physical and Chemical Research), 3-1-1 Koyadai, Tsukuba, Ibaraki 305, Japan*

Summary We report here a new method for the direct sequencing of large DNA inserts of cosmid clones from human chromosomes using STS primers of 14 nucleotides without any mismatches, which are designed from results of a mathematical analysis. It is clear that STS primer of 14 nucleotides is optimum for direct sequencing of cosmid recombinant DNA clones. We also provide examples of direct sequencing of cosmid clones of human chromosome 21 using these STS primers.

Key Words STS primer, cosmid, mathematical design, direct sequencing

INTRODUCTION

Human chromosomes have been well characterized genetically as part of an effort to identify the many genes that are potentially involved in important inherited disorders (Scoggin and Patterson, 1982; Siddique *et al.*, 1991; St. George-Hyslop *et al.*, 1987). The mapping of human chromosomes requires detailed information of nucleotide sequences. Several standard steps in general use for molecular cloning, such as screening, subcloning, and nucleotide sequencing, are required to generate such an information. Sequence-tagged sites (STSs), which are short tracts of operationally unique DNA sequences, have recently been used for generation of megabase contigs of particular chromosome regions (Olson *et al.*, 1990). Here, we describe the design of STS primers of optimum length by computerized mathematical analysis, as well as a new method of primer-directed sequencing of large DNAs from cosmid recombinant clones by use of optimized STS primers without any intermediate molecular-cloning steps.

Received July 15, 1993; Revised version received and accepted October 1, 1993.
*To whom correspondence should be addressed.

MATERIALS AND METHODS

The random primers were designed using the *JUMBLE* algorithm of *EuGene*™ software (Baylor College of Medicine, Houston, TX) on a Sun-4/2 computer. In the case of section 16, which contained ten random primers, a sequence of 16 nucleotides (AAAATTTTCCCCGGGG) was chosen as the original primer. This original primer was then randomized to ten random primers, namely, AATGCTTGGATAGCCC, ATAGAGTTACCCGCTG, GAGGACGATCTACTTC, GAAGTAGATCGTCCTC, GTTAGCTCGCAGAATC, TCGTTCGATGCAAGAC, CAAGGCTCATCTGTAG, CATAGTAGTGACCCGT, TAGCCGATGAAGTCCT, and GATCGAGCGATCATCT. The sequences of ten random primers were matched with

Table 1. The frequencies of matching of random primers to the total population of nucleotide sequences that have been entered in *GenBank*.

No. of section	No. of nucleotides (mer)	Number of random primers contained in a section	Nucleotide sequence [a] of one example of the random primers (5'→3')	Average number of matches / Total number of sequences in *GenBank*	Frequency of matches (%)
4	4	10	gatc	5.5×10^5 6.5×10^4	859
5	5	10	tagcc	1.4×10^5 "	215
6	6	10	catagt	3.2×10^4 "	50
7	7	10	caaggct	0.8×10^4 "	12
8	8	10	tcgttcga	0.2×10^4 "	3
9	9	10	gttagctcg	0.4×10^3 "	0.7
10	10	10	gaagtagatc	96 "	0.1
11	11	10	gaggacgatct	29.5 "	0.4×10^{-1}
12	12	10	atagagttaccc	6.1 "	0.9×10^{-2}
13	13	10	aatgcttggatag	1.2 "	0.1×10^{-2}
14	14	10	gatcgagcgatcat	0.3 "	0.4×10^{-3}
15	15	10	tagccgatgaagtcc	0.0 "	0.0
16	16	10	catagtagtgacccgt	0.0 "	0.0

[a] Only one sequence is given as an example from ten sequences of random primers.

the total sequence (1.3×10^8 bases, see *GenBank* Release 77.0) using GCG software on VAX-4000 computer. Only one of random primers, CATAGTAGTGACCC-GT, is shown as an example in Table 1. By means of this method, sequences of 130 random primers, divided into thirteen experimental sections from no. 4 to no. 16 (see Table 1) were generated and matched with the total sequence in *GenBank*.

The cosmid clones were constructed from the human-Chinese hamster hybrid 153E9a3 cell line (Patterson *et al.*, 1985) using the pWE15 vector (Toyobo Inc.,

Table 2. The efficiencies of matching of STS primers to cosmid DNAs for direct nucleotide sequencing.

No.	STS primer (5'→3')	Length of nucleotide (mer)	Original DNA marker (name)	Locus (name)	Accession no.	Cosmid as template DNA (name)	Sequencing using STS primers	
							distinct*	indistinct*
1	cctggtcaggctccc	15	pPW511-1H	D21S52	M94593	pC511	○	
2	cctggtcaggctcc	14	pPW511-1H	D21S52	M94593	pC511	○	
3	cctggtcaggctc	13	pPW511-1H	D21S52	M94593	pC511		○
4	tcctcatctgtaaaa	15	pPW512-6B	D21S53	M94595	pC512	○	
5	tcctcatctgtaaa	14	pPW512-6B	D21S53	M94595	pC512	○	
6	tcctcatctgtaa	13	pPW512-6B	D21S53	M94595	pC512		○
7	tcctcatctgta	12	pPW512-6B	D21S53	M94595	pC512		○
8	ccacagtgcctggcg	15	pPW512-6B	D21S53	M94595	pC512	○	
9	ccacagtgcctggc	14	pPW512-6B	D21S53	M94595	pC512	○	
10	ccacagtgcctgg	13	pPW512-6B	D21S53	M94595	pC512		○
11	ccacagtgcctg	12	pPW512-6B	D21S53	M94595	pC512		○
12	cttactattgctga	14	pPW513-5H	D21S54	M94598	pC513	○	
13	cttactattgctg	13	pPW513-5H	D21S54	M94598	pC513		○
14	actggtgtcttatt	14	pPW267C	D21S12	M94591	pC267	○	
15	actggtgtcttat	13	pPW267C	D21S12	M94591	pC267		○
16	ggcatattgctacc	14	pPW552-3H	D21S59	M94601	pC552-2	○	
17	ggcatattgctac	13	pPW552-3H	D21S59	M94601	pC552-2		○

*The symbol ○ under "distinct" indicates that the DNA of the cosmid could be clearly sequenced using an STS primer. The same symbol under "indistinct" indicater that sequencing of cosmid DNA could not be performed easily because of indistinct bands on nucleotide-sequencing gels.

Tokyo) by a standard method (Sambrook *et al.*, 1989).

PCR was performed using the DNA thermal cycler (Cetus, Norwalk, CT) in a total volume of 10 μl that contained 50 ng of genomic DNA, 10 pmol of each primer, 1.4 mM $MgCl_2$, 200 mM dNTPs, 50 mM KCl, 10 mM Tris-HCl (pH 8.3), 0.6 unit of DNA polymerase from *Thermus aquaticus* (Taq; Toyobo Inc., Tokyo), and 0.01% gelatin. Amplification was performed for 25 cycles with denaturation at 94°C for 1 min, annealing at 55°C for 2 min, and extension at 72°C for 3 min. Amplified DNA products were separated on a 2-mm-thick sequencing gel containing 10% polyacrylamide/15% glycerol. PCR-specific fragments were recovered and used as probes for screening cosmid clones (pC511, pC512, pC513, pC267, and pC522-2, see Table 2) by single colony hybridization.

The cosmid clones DNAs that contained a DNA fragment of 40 to 45 kbp from a human chromosome were purified and then sequenced by a modified version of the method for primer-directed sequencing of cosmid DNA. Each annealing reaction was performed in a total volume of 10 μl that contained 10–15 μg of denatured cosmid DNA, 2 μl of 5× reaction buffer and 1 μl of 0.5 pmol of STS primer (15-, 14-, 13-, or 12-mer) or T3 promoter primer (20-mer; Promege Co., Madison, WI) or pBR322-primer P1 (16-mer; Takara Shuzo Co., Kyoto) as controls. The reaction mixture was warmed to 65°C for 30 min and then cooled slowly to room temperature over a period of about 30 min. After annealing, the reaction mixture was supplemented with 1 μl of 0.1 M DTT, 2 μl of diluted labeling mix, 0.5 μl of [α-^{32}P]dCTP, 2 μl of diluted enzyme from a kit (Sequenase® Version 2.0; Toyobo Inc., Tokyo) and incubated for 10 min at 37°C. Aliquots of 3.5 μl were removed and mixed with each of four ddNTP termination mixes (ddATP or ddGTP or ddCTP or ddTTP), with subsequent incubation at 37°C for 10 min. The reactions were stopped by addition of 4 μl of stop solution. Then 1.2 μl of each sample was used for sequencing in the usual way (Slightom and Sieu, 1992; Tang *et al.*, 1992).

RESULTS AND DISCUSSION

Several methods for the direct sequencing of large DNA fragments such as those involving genomic cosmid and P1 phage recombinant clones, have been reported (Slightom and Sieu, 1992). However, when these methods are used, it is sometimes difficult to sequence the DNA because of poor annealing of the primer to large pieces of DNA when the length of the STS primer is more than 20 nucleotides (Sambrook *et al.*, 1989; Lathe, 1985; Ikuta *et al.*, 1987). Thus, it is now important to know the optimum length of STS primers that is long enough for specific annealing and short enough to assure perfect matches between the primers and large fragment of DNA. In order to determine the shortest practical lengths of the STS primers, it is necessary to estimate the frequency of matching of the primer to the genome of interest. For example, in the case of hybridization blotting of

DNA, the number of independent perfect matches (K) for a synthetic oligonucleotide probe of length L in a genome of complexity C can be calculated from the following equation: $K=(1/4)^L \times 2C$ (Lathe, 1985). When $4^L=2C$, the sequence of the oligonucleotide would be expected to occur only once in a genome. In the case of the human genome, $2C=6\times10^9$ and $L=16$. Thus, an oligonucleotide probe of 16 nucleotides would be expected to be matched with and to hybridize to a human genome only once. However, it is well known that many repeated sequences, such as moderately repetitive DNAs, highly repetitive DNAs, RNA genes (rRNA, tRNA), satellite DNA, histone gene and a large amount of poly A sequences, constitute over 50% of human genome (Britten and Davidson, 1971; Darnell, 1976, 1983; Pardue and Gall, 1970; Singer and Skowronski, 1985; Sinclair and Brown, 1971; Jelinek and Schmid, 1982). For example, repeated sequences with four identical nucleotides (AAAA or GGGG or CCCC or TTTT) can be matched 9.1×10^5 times in the total DNA sequence (8.3×10^7 bases) entered in *GenBank* (see Release 71.0 of *GenBank*, Bilofsky *et al.*, 1986). There is no doubt that the more identical nucleotides are repeated within a sequence, the more matches there will be between this sequence and human genomic DNA. Indeed, the actual value for the complexity (C) of the human genome is lower than the theoretical value given above when we take the repeated sequences into account (Williams, 1989). However at the moment we do not know the exact value of the complexity (C) of the human genome without such repetitive DNA sequences. Therefore we take into account the actual value of the complexity (C) as the total sum of the nucleotide sequences entered into *GenBank*. Thus, STS primers of shorter length (L) should be useful for the sequencing of cosmid DNAs provided that the primers do not contain repeated sequences with sequences of four or more of identical nucleotides.

A total of 130 primers with different sequences and lengths were randomly designed by the *JUMBLE* algorithm of *EuGene*™ software (Baylor College of Medicine, Houston, TX) on a Sun-4/2 computer. The randomly selected primers were divided into thirteen experimental sections that were individually assigned different numbers from four to sixteen (Table 1). The randomly selected primers in each section were of the same length but had different sequences from one another. All of the sequences in the experimental sections were surveyed and we examined whether or not there were any matches with the total sequence (1.3×10^8 bases) in the *GenBank* Release 77.0 using the Pattern-Search program of GCG software (Genetics Computer Group Sequence Analysis Software Package, University of Wisconsin, WI) on a VAX-4000 computer. The sequence of a random primer and an individual sequence entered into *GenBank* were assigned the letters W and U. The signal function, Sign (W, U), can be defined as follows:

$$\text{Sign}(W,U)=\begin{cases}1 \text{ if W is identical to U}\\0 \text{ if W is different from U}\end{cases}$$

Thus, it is postulated that the frequencies of matching of random primers to the

total sequence in *GenBank* can be calculated from the following equation:

$$f_r(k) = \frac{1}{mn} \sum_{i=1}^{m} \sum_{S_j \in \mathbf{B}} \sum_{L^*_{k,j} \subset S_j} \mathrm{Sign}(L_{k,i}, L^*_{k,j}) \qquad [1]$$

where $f_r(k)$ is the frequency of matching of random primers with a length of k nucleotides, **B** is a set that contains all of the sequences in *GenBank*, n is the number of total nucleotide sequences, S_j is *j*th sequence in **B**, $L_{k,i}$ is the *i*th random primer from the experimental section with a length of k nucleotides, m is the number of random primers in an experimental section and is set at ten, and $L^*_{k,j}$ is an arbitrary fragment of length k that occurs in the sequence S_j.

After analyzing the data for random primers, it was clear that the average number of matches was 1.2 for section 13, 0.3 for section 14, and 0 for sections 15 and 16 (see Table 1). The frequency of matches is shown in Fig. 1. In other words, the randomly selected different sequences in the section 14, which contained ten random primers of 14 nucleotides each, could be matched less than once with 1.3×10^8 bases of total sequence (*GenBank* Release 77.0). It is evident that the complexity (C) of the human genome (3×10^9 bases) is much higher than that of human genes entered into *GenBank* (2.5×10^7 bases). Moreover we do not know the estimated value of the complexity (C) of human genome without any repetitive DNA sequences. However, at least we can reduce the actual value of complexity (C) by the following reasons. 1) Human genome contains 50% or more repetitive DNA sequences. 2) The total sequence in *GenBank* is almost specific and non-repetitive. 3) The design of STS primer without repetitive DNA sequences should be taken into account. 4) The DNA sequences entered into *GenBank* are under a bias to-

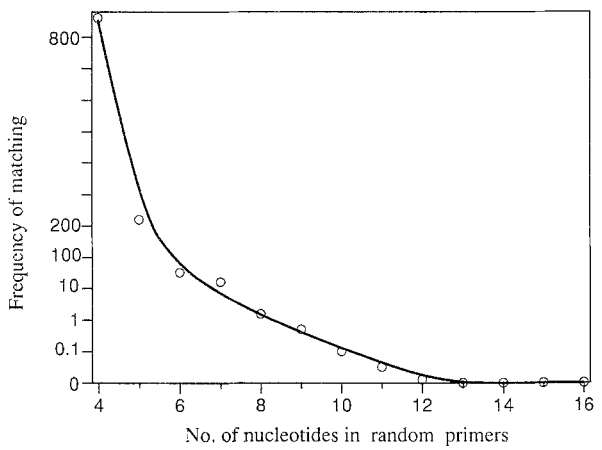

Fig. 1. The frequency of matching of random primers (see Table 1) was calculated using the Pattern-Search program of GCG software on a VAX-4000 computer. The circles indicate the frequency of matches as described in Table 1.

ward the gene sequences included the coding sequences. It is well known that only one or two percent of human genome is expressed (MacLean *et al.*, 1983). Thus the actual value of complexity (C) of human genome should be lower as we expected. Therefore we elected to calculate the optimum size of primers on the basis of the total sequences of genes entered into *GenBank*, instead of the complexity of human genome (3×10^9 bases). Thus, we propose that an STS primer of 14 nucleotides should allow specific annealing and more efficient sequencing of cosmid DNA, although we cannot rule out the possibility that the STS primer of more than 14 nucleotides is optimum for direct sequencing of human genomic DNA clones of larger insert size.

Both 5′- and 3′-flanking sequences of STS markers (Tang *et al.*, 1992) were used as primers for amplification of DNA with the DNA thermal cycler in the presence of human placental DNA (Fig. 2). Amplified fragments of DNA were recovered and used as probes for selection of cosmid clones from a 153E9a3 hybrid cell genomic library (Patterson *et al.*, 1985). One cosmid clone (pC511, Fig. 3a) that contained sequences homologous to the pPW511-1H DNA marker (Locus D21S52, accession number M94593) was chosen. An STS primer of 13, 14, or 15 nucleotides was designed on the basis of pPW511-1H and used for direct sequencing of pC511 cosmid DNA (Table 2 and Fig. 3). It was obvious that DNA

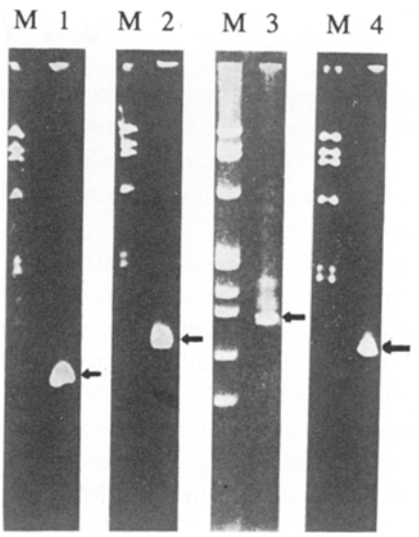

Fig. 2. Results of polyacrylamide gel electrophoresis of amplified DNA products generated by PCR. The following substrate DNAs were individually amplified with pairs of STS primers (forward/reverse, Tang *et al.*, 1992) by the PCR reaction. Lane 1, pPW511-1H; lane 2, pPW512-6B; lane 3, pPW513-5H; lane 4, pPW552-3H. Amplified human-specific DNA is indicated by arrowheads (106 bp, lane 1; 145 bp, lane 2; 173 bp, lane 3; 159 bp, lane 4). Lanes marked M contain fragments of ϕX 174 DNA that had been digested with *Hae*III.

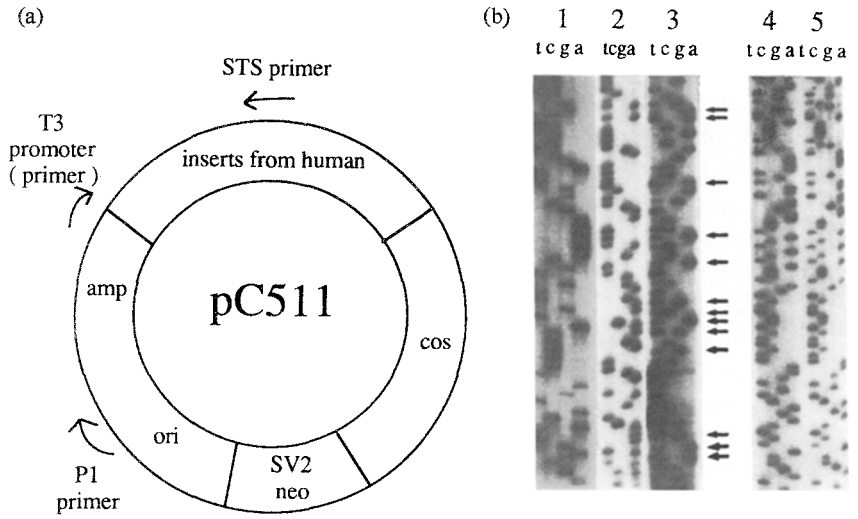

Fig. 3. Direct sequencing of cosmid DNA (pC511) that contained a DNA fragment (insert size of 45 kbp) from a human chromosome in the pPW15 vector (a), performed with Sequenase® version 2.0 on a 6% polyacrylamide gel (b), pBR322 primer P1 (lane 1), T3 promoter primer (lane 2), and STS primers of 13 (5′-cctggtcaggctc, lane 3), 14 (5′-cctggtcaggctcc, lane 4), or 15 (5′-cctggtcaggctccc, lane 5) nucleotides were used for sequencing. Indistinct sequencing bands (arrow) can be observed in lane 3.

of pC511 could easily be sequenced using the STS primers of 14 or 15 nucleotides (Fig. 3b; lanes 4 and 5). However, the STS primer of 13 nucleotides was inadequate for generation of distinct bands during nucleotide sequencing (Fig. 3b; lane 3) presumably because of mismatches between primer and template DNA.

By using this method, STS primers of 12, 13, 14, or 15 nucleotides were designed by reference to nucleotide sequences of other DNA markers (pPW512-6B, pPW513-5H, pPW267C, pPW552-3H) on the basis of rule as described by Williams (1989) and were then used for direct sequencing of corresponding cosmid clones (pC512, pC513, pC267, pC552-2). Details of the sequencing of DNA with different STS primers are shown in Table 2. Clearly, there were no problems associated with direct sequencing of cosmid cloned DNA with all 14-mer STS primers, but $\geq 50\%$ of DNA samples failed to give distinct bands of nucleotides on gels when 13-mer or 12-mer STS primers were used. Moreover, the 14-mer STS primers for direct sequencing of cosmid cloned DNA always gave distinct bands of nucleotides on gels as far as we have sequenced more than one hundred cosmid cloned DNAs (data not shown). It is also possible that the nucleotide at the 3′-end of the oligonucleotide primer may influence the priming ability of different primers to cosmid DNA clones because the primers ending in purines provide more efficient priming. This effect could be more exaggerated in the case of shorter primers. Thus in attempt to exclude this possibility, we analyzed the priming capability of

the different composition of the 3′-end of the oligonucleotide primers, especially to compare the sequence preference of the purine and the pyrimidine residue at the 3′-end of the primer. As shown in lanes 4 to 12 in Table 2, the optimum shorter STS primers for the direct sequencing of cosmid cloned DNA was 14 nucleotides in both cases with either the purine or the pyrimidine residue at the 3′-end. Thus, these data clearly demonstrated that the hypothetical frequency of matching of random primers in Table 1 adequately represents the frequency of matches generated by the STS primers used here (Table 2).

It is generally accepted that STS primers of 15–20 nucleotides can be used as primers for sequencing of DNA. However, an STS primer of 14 nucleotides is shorter and less expensive and easier to synthesize than a longer one. Moreover, an STS primer of 14 nucleotides can be used more efficiently for direct sequencing of cosmid recombinant clones without problems because it is specific and anneals more easily to large pieces of DNA. Taken together, our results suggest that the optimal length of an STS primer is 14 nucleotides for "direct sequencing," especially under the condition of the ambient temperature of annealing with large DNA inserts, such as P1 phage cloned DNA, yeast artificial chromosome (YAC) DNA and cosmid cloned DNA (see Table 2 and unpublished data).

The information about STSs in a database can be obtained easily and can help us to isolate the respective YAC clones or cosmid recombinant clones. It has also been known that large pieces of DNA that contain important genes can be sequenced directly without any intermediate molecular-cloning steps. Our results indicate that direct sequencing of cosmid cloned DNA, with an optimized STS primer of 14 nucleotides without any mismatches, will be useful for construction of a long-range physical map of human chromosomes.

Acknowledgments We thank Mayumi Nifuku and Sachiyo Shimizu for excellent secretarial assistance. This work was supported by Special Coordination Funds of the Science and Technology Agency of the Japanese Government and by grants from the Life Science Research Project of RIKEN.

REFERENCES

Bilofsky HS, Burks C, Fickeet JW, Goad WB, Lewitter FI, Rindone WP, Swindell CD, Tung CS (1986): The GenBank genetic sequence databank. Nucleic Acids Res **14**: 1–4

Britten RJ, Davidson EH (1971): Repetitive and non-repetitive DNA sequences and a speculation on the origins of evolutionary novelty. Quant Rev Biol **46**: 111–138

Darnell JE (1976): mRNA structure and function. Prog Nucl Acid Res Mol Biol **19**: 493–511

Darnell JE (1983): The processing of RNA. Sci Am **249**: 89–99

Ikuta S, Takagi K, Wallace RB, Itakura K (1987): Dissociation kinetics of 19 base paired oligonucleotide-DNA duplexes containing different single mismatched base pairs. Nucleic Acids Res **15**: 797–801

Jelinek WR, Schmid CW (1982): Repetitive sequences in eukaryotic DNA and their expression. Annu Rev Biochem **51**: 813–844

Lathe R (1985): Synthetic oligonucleotide probes deduced from amino acid sequence data theoretical and practical considerations. J Mol Biol **183**: 1–12

MacLean N, Gregory SP, Fravell RA (1983): Eukaryotic genes: Their structure, activity and regulation. Butter-Worths, London

Olson MV, Hood L, Cantor C, Botstein D (1990): A common language for physical mapping of the human genome. Science **245**: 1434–1435

Pardue ML, Gall JG (1970): Chromosomal localization of mouse satellite DNA. Science **168**: 1356–1358

Patterson D, Van Keuren M, Drabkin H, Watkins PC, Gusella JF, Scoggin CH (1985): Molecular analysis of chromosome 21 using somatic cell hybrids. Ann NY Acad Sci **450**: 109–120

Sambrook J, Fritsch EF, Maniatis T (1989): Molecular Cloning, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY

Scoggin CH, Patterson D (1982): Down's syndrome as a model disease. Arch Int Med **142**: 462–464

Siddique T, Figlewitz DA, Pericak-Vance MA, Haines JL, Rouleau G, Jeffers AJ, Sapp P, Hung WY, Bebout J, McKenna-Yasek D, Deng G, Horvitz HR, Gusella JF, Brown RH Jr, Roses AD, Collaborators (1991): Linkage of a gene causing familial lateral sclerosis of chromosome 21 and evidence of genetic-locus heterogeneity. New Engl J Med **324**: 1381–1384

Sinclair JH, Brown DD (1971): Retention of common nucleotide sequences in the ribosomal deoxyribonucleic acid of eukaryotes and some of their physical characteristics. Biochemistry **10**: 2761–2769

Singer MF, Skowronski J (1985): Making sense out of lines: long interspersed repeat sequences in mammalian genomes. Trends Biochem Sci **10**: 119–122

Slightom JL, Sieu LC (1992): Direct sequencing of baculovirus genomic DNA: Sequence determination of the engineered respiratory syncytial virus chimeric FG gene. BioTechniques **13**: 94–105

St George-Hyslop PH, Tanzi RE, Polinsky RJ, Haines JL, Nee RG, Watkins PC, Myers RH, Feldman RG, Pollen D, Drachman D, Bruni A, Foncin JF, Salmon D, Fromment P, Amaducci L, Sorbi S, Placentini S, Stewart GD, Hobbs WJ, Conneally PM, Gusella JF (1987): The genetic defect causing familial Alzheimer's disease maps on chromosome 21. Science **235**: 885–890

Tang X, Tashiro H, Eki T, Murakami Y, Soeda E, Sakakura T, Watkins PC, Yokoyama K (1992): Generation of 19 STS markers that can be anchored at specific sites on human chromosome 21. Genomics **14**: 185–187

Williams JF (1989): Optimization strategies for the polymerase chain reaction. BioTechniques **7**: 762–768