Volume 15 Number 6 1987 Nucleic Acids Research

Characterization of a human 'midisatellite' sequence

Yusuke Nakamura[1], Cecile Julier[1], Roger Wolff[1], Tom Holm[1], Peter O'Connell[1], Mark Leppert[1] and Ray White[1,2]

[1]Howard Hughes Medical Institute, University of Utah Medical Center and [2]Department of Human Genetics, University of Utah Medical School, Salt Lake City, UT 84132, USA

Received December 2, 1986; Revised and Accepted February 23, 1987

ABSTRACT

We have examined the structure and DNA sequence of a human genomic locus that consists of a large hypervariable region made up of repeats of a simple sequence. With several restriction enzymes, the locus shows many restriction fragments that vary quantitatively as well as qualitatively. Other restriction enzymes produce only a single, high-molecular-weight fragment at this locus. Almost all of the fragments are revealed with a simple sequence probe. Southern transfers of the high-molecular-weight restriction fragments produced by the restriction enzymes NotI and SfiI, resolved by pulsed-field gel electrophoresis, gave at most two fragments, demonstrated to be allelic, showing that the majority of the restriction fragments seen in the complex patterns are at a single locus. The estimated size of the region homologous to the probe varied from 250 to 500 kilobases. DNA sequencing indicated that the region consists of tandem repeats of a 40-base-pair sequence. Some homology was detected to the tandem repeating units of the insulin gene and the zetaglobin pseudogene hypervariable regions, and to the "minisatellite" DNA at the myoglobin locus.

INTRODUCTION

It is a general feature of organization in the eukaryotic genome that single-copy DNAs are interspersed with repetitive DNA sequences (1). In the human, several kinds of repetitive DNA families, each with a unique length and DNA sequence of repeating unit, a different copy number per genome and a different organization in the genome, have been already reported (2-6). For example, the AluI sequences are ubiquitous (3 X 10^5 copies per genome) and interspersed throughout the human genome. They occur with an average spacing of some 10kb and are known as short interspersed sequences (SINEs)(3). In contrast, elements of the KpnI family, fewer in number, are interspersed at longer distances from each other and are called LINEs (long interspersed

© IRL Press Limited, Oxford, England.

sequences)(4). Another group of repetitive sequences, the alphoid DNA family that is homologous to the alpha component in the African green monkey (5), has been reported to have 55,000 copies per haploid genome (6). This family has been divided into many subgroups, each of which is considered to be chromosome-specific and is estimated to exist in 100 - 2000 copies per haploid genome(6-8).

The function of repetitive sequences has been widely discussed (7-13), but remains unclear. However, the alphoid family is suspected of having an important role in chromosomal structure or function because of its unique genomic organization (6,7,14,15).

Very short clusters of tandem repeats of a simple sequence have also been characterized in the human genome; at some loci, the number of copies of the short repeat varies within the population (16-18). Polymorphic loci of this kind are designated "minisatellites" or hypervariable regions. These "minisatellite" loci as well as some "satellite" DNA loci have been suspected as hot spots of homologous recombination (14,16).

In this paper, we describe yet another form of repetitive sequence: "midisatellite" DNA, which consists of some 250 - 500 kilobases of repetitive DNA that is clustered at a single locus and shows a highly polymorphic pattern in the population.

MATERIALS AND METHODS

DNA sources and cosmid library

Genomic DNAs were isolated from lymphocytes or lymphoblast cell lines as described elsewhere (19). The human genomic cosmid library was a gift from Drs. Y.F.Lau and Y.W.Kan (20). Oligonucleotides were provided by Dr. R. Gesteland.

Pulsed-field gel electrophoresis

To a suspension of 1.2 x 10^8 lymphoblast cells in 1.5 ml of PBS (0.14M NaCl, 5mM KCl and 20mM phosphate [pH7.5]), 1.5 ml of 1% low-melting-point agarose (FMC Bioproducts) containing 10mM Tris.HCl (pH 7.4), 100mM NaCl, and 100mM EDTA at 42^{o} C was added. The gently mixed solution was poured onto a glass plate. After solidification, the agarose was cut into pieces of approximately 45 μl; each piece was expected to include 10μg of human high-

molecular-weight DNA. These pieces were incubated 1-2 days at 37°C in 10mM Tris HCl (pH 8.0), 1% (w/v) sodium lauroyl sarcosine, and 0.45mM EDTA including 1 mg/ml of ProteinaseK (Sigma). The agarose pieces were washed twice in TE (10mM Tris.HCl [pH7.4], 1mM EDTA), and three times in TE-4 (10mM TrisHCl [pH 7.4], 0.1mM EDTA) for 2 hr. Each 45 µl of agarose was digested overnight in 250 µl of reaction mixture under the conditions specified by the manufacturers of the enzymes (EcoRI: Boehringer Mannheim; SfiI and NotI: New England BioLabs), with the addition of 5 mM spermidine. The agarose plugs were loaded into a 1.5 % agarose gel, and the electrophoresis was done at 13°C, 10V/cm under the following conditions: EcoRI--30 sec pulse time for 20 hr; SfiI--40 sec pulse time for 24 hr; NotI--40 sec pulse time for 24 hr (21). For molecular standards we used ligated lambda phage DNA.

Southern transfer and colony hybridization

After electrophoresis, DNAs were transferred to nylon membranes according to (22). Colony transfer was done according to (23). Hybridization was carried out in 50% formamide, 5 x SSC (1 x SSC = 0.15 M NaCl and 0.015 M Na Citrate), 50 mM $NaPO_4$, 1 x Denhardt's solution (0.02% bovine serum alubumin, 0.02% polyvinyl-pyrrolidone and 0.02% Ficoll), 100 mcg/ml of salmon sperm DNA, and 1 x 10^6 (Southern blot) or 1 x 10^5 (colony hybridization) cpm/ml of ^{32}-P dCTP-labeled probe(24). Washing was done twice for 30 min in a solution of 0.1 % SSC, 0.1 % SDS at 65°C.

DNA sequences

In order to identify highly polymorphic DNA marker loci, we screened a human cosmid library with a 14-base oligonucleotide (ACAGGGGTGAGGGG) derived from the insulin gene (16) according to a protocol described elsewhere (25). This oligonucleotide is the consensus sequence that, tandemly repeated, is responsible for multiallelic polymorphism at the insulin locus (16). One of the positive cosmid clones was designated YNI10. A plasmid, pYNI10.4, was obtained by incorporating the 2.0 kilobase HindIII fragment of cosmid YNI10 into the HindIII site of pBR322.

We used the 2.0 kb HindIII fragment from pYNI10.4 as a probe to screen a cosmid library. One of the positive clones was

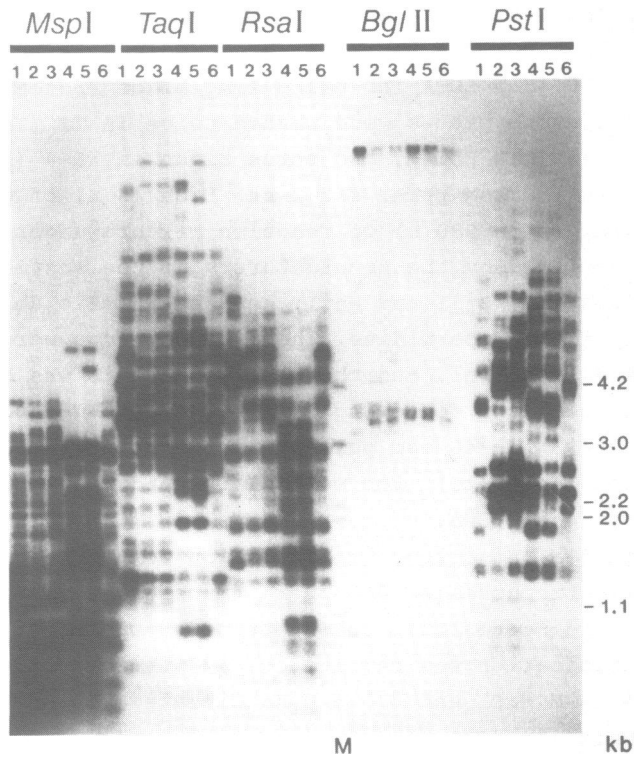

Figure 1. Southern blot of genomic DNA hybridized with pYNI10.4. DNAs were isolated from six unrelated individuals (here designated 1-6; M=marker lane). Exposure time was 4 hours. (From White et al., in Cold Spring Harbor Symposium on Quantitative Biology 51, The Molecular Biology of Homo Sapiens. Used by permission).

designated THIZ41; the 0.7 kb and 0.5 kb MspI fragments from this cosmid were also isolated and subcloned into M13 mp8 and mp9. The sequencing reactions were carried out with ^{32}P-dCTP (New England Nuclear) according to Sanger (26). A 17-base primer was purchased from Boehringer Mannheim, and deoxy- and dideoxy-nucleotides were supplied by P-L Biochemicals.

RESULTS

Southern transfer of pYN10.4

Southern transfers of human DNA from six unrelated individuals, hybridized with pYNI10.4, are shown in Fig. 1. These

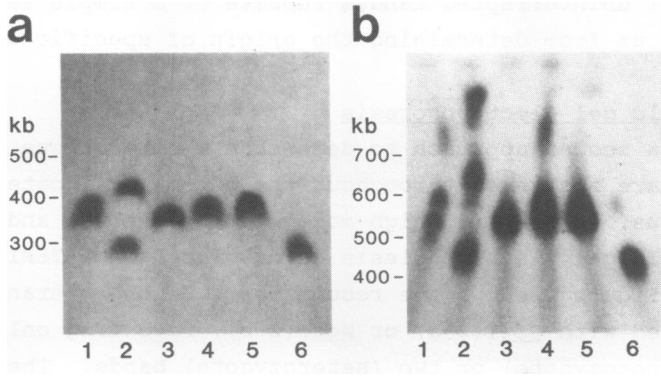

Figure 2. Southern blot of high-molecular-weight DNAs hybridized with pYNI10.4. DNAs were digested with SfiI(a) or NotI(b); 4 hours' exposure.

patterns, the same as those obtained with the whole cosmid as a probe, were visible with an intensifying screen after only four hours' exposure at room temperature. DNAs digested with MspI, TaqI, RsaI, or PstI showed a complex and polymorphic pattern of bands. Digestion with these enzymes also showed a number of fragments of constant size. The sum of the molecular sizes of these bands is more than 100kb; furthermore, many of the fragments of constant length are present in multiple copies and show quantitative variation. On the other hand, no polymorphism could be detected when DNAs were digested with HindIII or BamHI (data not shown), because the size of the resulting fragments was at the limiting mobility of the gel under our conventional electrophoretic conditions. EcoRI gave the same patterns as BamHI or HindIII, except for the presence of several additional, smaller bands like those produced by digestion with BglII.

In an attempt to determine if this complex pattern of fragments reflected a number of loci with high sequence homology scattered throughout the genome, we screened the human genomic cosmid library with the 2.0 kb HindIII fragment insert of pYNI10.4. We obtained 80 positive clones per haploid genome. When digested with BamHI and EcoRI, only five of the 80 cosmids showed the presence of a restriction site, suggesting the

presence of uninterrupted tandem repeats of a simple sequence and preventing us from determining the origin of specific cosmid clones.

Pulsed-field gel electrophoresis

In a second approach to determine whether these repetitive sequences are scattered throughout the genome or located at a single locus, we prepared high-molecular-weight DNA and performed pulsed-field gel electropholesis as described in MATERIALS AND METHODS. Figure 2 shows the results of a Southern transfer of DNA digested with SfiI(2-a) or NotI(2-b), revealing only one (apparent homozygote) or two (heterozygote) bands. These high molecular weight bands also show polymorphism. The estimated size range of the fragments was 250 to 500 kb for SfiI and 450 to 700 kb for NotI.

DNA sequences at the midisatellite locus

We determined the DNA sequences of three fragments derived from two different cosmids. One was the 2.0 kb HindIII fragment of cosmid YNI10, which has many HindIII or BamHI recognition sites; the other two, which corresponded to the very dark 0.6 and 0.8 kb MspI bands in Figure 1, were isolated from cosmid THIZ41 which has no HindIII, BamHI or EcoRI recognition sites in human sequences. Figure 3 shows DNA sequences from these three fragments, involving almost 1400 base pairs of DNA. Suprisingly, all of the sequences we determined consisted of repeats, with some internal variation, of a simple sequence. The size of the repeating units varied from 37 to 43 base pairs, with an average length of 40 base pairs. There are some substitutions among the repeating units, but each has 67.5-95% homology to the consensus sequence, which we consider to be CCTGGGGGTGTGGGTGCTGCTCCAGGCTGTCAGATGCTCA. Within the consensus sequence, two parts (TGCTGC and TCAGATGCT) are well conserved (89-100%), and between these regions there is a variable region(AGGCTG) which maintains only 61-81% homology with the consensus sequence. Figure 4 compares the sequence at the I10.4 locus with the sequences of the repeating units of the insulin gene (19), the zetaglobin pseudogene (17) and the myoglobin "minisatellite" sequence (18). These tandem sequences have some homology to the repeating units of the other loci; GTGGG, for

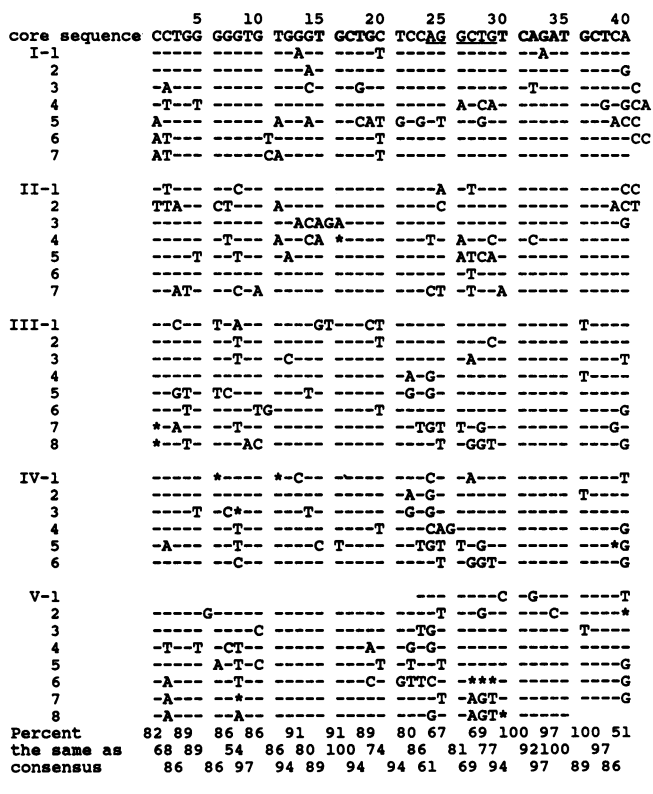

Figure 3. DNA sequence of three fragments of the "midisatellite" family. I and II: partial sequences of 0.6kb of the MspI fragment of cosmid THIZ41; III and IV:partial sequences of 0.8kb of the MspI fragment of cosmid THIZ41; V: partial sequence of 2.2kb of the HindIII fragment (=pYNI10.4) of cosmid YNI10. Core sequence: the DNA sequence most often found in sequenced regions. The sequence underlined is the variable portion. -:identical to the core sequence. *: deletion of the sequence.

example, is common to all four kinds of sequences. We named the I10.4 locus a "midisatellite" because it is similar to the "minisatellite" loci in containing tandem repeats of a simple sequence, but considerably larger. To confirm that the core sequence was included within all bands shown in Figure 1, we synthesized a 40-base-long oligonucleotide corresponding to the consensus sequence and used it as a probe for Southern transfers. The oligonucleotide gave the same result as that seen in Fig. 1.

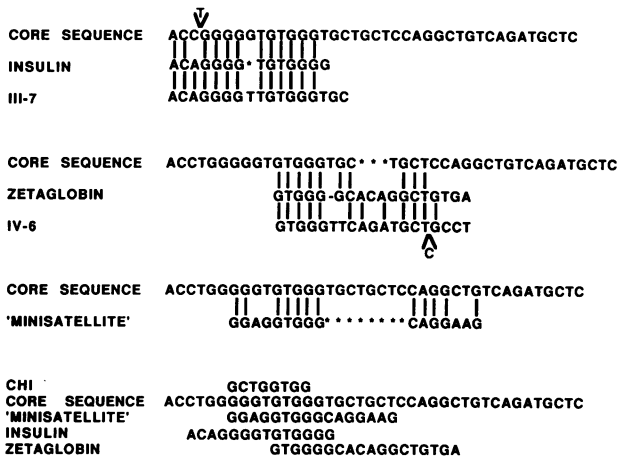

Figure 4. Comparison of DNA sequences between the repeating unit of the "midisatellite" and other repeat sequences, including lines III-7 and IV-6 from fig. 3. See text for explanation.

Is this locus inherited in a Mendelian fashion?

We examined the inheritance of this locus, using DNAs isolated from one pedigree. Figure 5 shows the results of Southern blotting after pulsed-field gel electrophoresis. The locus was inherited in Mendelian fashion, and the sizes of the alleles within the tested sibships were estimated to be 325kb and 350kb.

DISCUSSION

We have described here a repetitive DNA sequence "midisatellite" which shows a restriction fragment band pattern that is highly polymorphic, both qualitatively and quantitatively. This is the first report of a VNTR (Variable Number of Tandem Repeat) locus (25) of this molecular size (250 - 500 kb). We suspect that all or most of this locus consists of repeats of a simple sequence, because hybridization with a 40-base core oligonucleotide reveals the same complex pattern in a Southern transfer as hybridization with the whole cosmid DNA. Other sequences, however, could be interspersed within these repeats.

Jeffreys et al. (18,27) suggested that a "minisatellite"

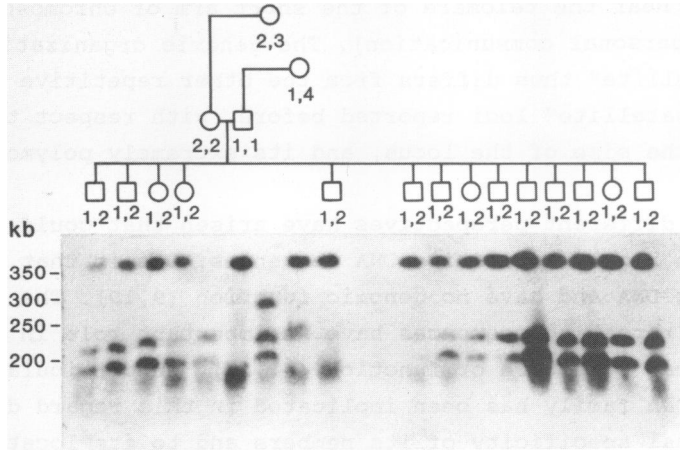

Figure 5. Southern blot of the high-molecular-weight DNAs with pYNI10.4; 5 hours' exposure. DNAs were isolated from individuals of one kindred and digested with EcoRI. The genotypes of individuals in the three-generation family are shown directly below their symbols in the pedigree.

repeat might be a signal for recombination, because of its highly polymorphic "fingerprint" pattern, its sequence similarity to the chi sequence of lambda phage, and the high mutability observed at one of the "minisatellite" loci. Waye and Willard have also suggested homologous unequal crossing-over between or within high-order repeats in the alpha satellite of chromosome 17(7,14).

We have demonstrated that the core sequences of the "midisatellite" locus bear some homology to those of the repetitive sequence of the insulin gene (16) and the zetaglobin pseudogene (17). In particular, we suspect that the sequence GTGGG, which is common within at least four different kinds of repeating units and is similar to the lambda chi sequence, may have a role in recombination. However, any mechanism that results in the amplification or diminution of tandem repeats at this locus could produce the same result as recombination.

The repetitive DNA families such as the AluI family, the KpnI family, or the HinfI family are dispersed throughout the genome (2,4,12). Only the alphoid family is believed to be confined to specific genomic regions (5,15,28-30). The "midisatellite" family reported here appears to be a cluster at

one locus near the telomere of the short arm of chromosome 1 (M.Litt, personal communication). The genomic organization of the "midisatellite" thus differs from the other repetitive families or "minisatellite" loci reported before, with respect to copy number, the size of the locus, and its extremely polymorphic pattern.

Two different perspectives have arisen that could account for human highly repetitive DNA sequences. One is that they are parasitic DNA and have no genomic function (9,10). The other is that the repeating sequences have an important role in chromosomal structure or function (31,32). In particular,the alphoid DNA family has been implicated in this regard due to chromosomal specificity of its members and to its location in centromeres (5,15,28-30). We have no direct evidence that the "midisatellite" DNA family reported here has a significant genomic function. However, because the same fragment sizes are conserved during evolution in spite of showing a highly polymorphic pattern, and because this "midisatellite" is found at only one locus, we speculate that it may have an important chromosomal function, such as the recognition of a homologous region at meiotic pairing.

ACKNOWLEDGMENTS

We are grateful to Esther Fujimoto for expert technical assistance, and to Ruth Foltz for editorial work in the preparation of this manuscript. R.W. is an investigator of the Howard Hughes Medical Institute.

REFERENCES

1. Davidson, E. H., Galau, G. A., Angerer, R. C. and Britten, R. J. (1975) Chromosoma 51, 253-259
2. Wu, J. C. and Manuelidis, L. (1980) J. Mol. Biol. 142, 363-386
3. Houck, C. M., Rinehart, F. P. and Schmid, C. W. (1979) J. Mol. Biol. 132, 289-306
4. Shafit-Zagardo, S., Maio, J. J. and Brown, F. L. (1982) Nucleic Acids Res. 10, 3175-3193
5. Manuelidis, L. and Wu, J. C. (1978) Nature 276, 92-94
6. Devilee, P., Slagboon, P., Cornelisse, C. J. and Pearson, P. L. (1986) Nucleic Acid Reseach 14, 2059-2072
7. Waye, J. S. and Willard, H. F. (1986) Nucleic Acid Res. 14, 6915-6927
8. Wolfe, J., Darling, S. M., Erickson, R. P., Craig, I. W.,

Buckle, V. J., Rigby, P. W., Willard, H. F. and Goodfellow, P. N. (1985) J. Mol. Biol. 182, 477-485
9. Doorlite. I. W. and Sapienza, C. (1980) Nature 284, 601-603
10. Orgel, L. E. and Crick, F. H. (1980) Nature 284, 604-607
11. Haynes, S. R., Toomey, T. P., Leinwand, L. and Jelinek, W. R. (1981) Molec. Cell. Biol. 1, 573-583
12. Shimizu, Y., Yoshida, K., Ren, C-S., Fujinaga, K., Rajagopalan, S. and Chinnadurai, G. (1983) Nature 302, 587-590
13. Furano, A. V., Somerville, C. C., Tsichlis, P. N. and D'Ambrosio, E. (1986) Nucleic Acids Res. 14, 3717-3727
14. Waye, J. S. and Willard, H. F. (1986) Mol. Cell. Biol. 6, 3156-3165
15. Devilee, P., Cremer, T., Slagboom, P., Bakker, E., School, H. P., Hager, H. D., Stevenson, A. F. G., Cornelisse, C. J. and Pearson P. L. (1986) Cytogenet. Cell. Genet. 41, 193-201
16. Bell, G. I., Selby, M. J., and Rutter, W. J. (1982) Nature 295, 31-35
17. Proudfoot, N. J., Gil, A. and Maniatis, T. (1982) Cell 31, 553-563
18. Jeffreys, A. J., Wilson, V. and Thein, S. L. (1985) Nature 314, 67-73
19. Bell, G., Karam, J., and Rutter, W. (1981) Proc. Natl. Acad. Sci. U.S.A. 78, 5759-5763
20. Lau Y-F. and Kan Y. W. (1983) Proc. Natl. Acad. Sci. U.S.A. 80, 5225-5229
21. Carle, G. F. and Olson, M. V. (1984) Nucleic Acids Research 12, 5647-5664
22. Reed, K. C. and Mann. D. A. (1985) Nucleic Acids Research 13, 7207-7221
23. Grunstein, M. and Hogness, D. S.(1975) Proc. Natl. Acad. Sci. U.S.A. 72, 3961-3965
24. Feinberg, A. P. and Vogelstein, B. (1983) Anal. Biochem. 132, 6-13
25. Nakamura, Y., Leppert, M., O'Connell, P., Wolff, R., Holm, T., Culver, M., Martin, C. Kumlin, E. Hoff, M. and White, R. (1987) Science (in press)
26. Sanger, F., Coulson, A. R., Barrell, B. G., Smith, A. J. H., and Roe, B. A. (1980) Am. J. Physiol. 238, G467-G477
27. Wong, Z., Wilson, V., Jeffreys, A. J. and Thein, S. L. (1986) Nucleic Acid Res. 14, 4605-4616
28. Yang, T. P., Hansen, S. K., Oishi, K. K., Ryder, O. A., and Hamkalo, B. A. (1982) Proc. Natl. Acad. Sci. USA 79, 6593-6597
29. Willard, H. F., Smith, K. D. and Sutherland, J. (1983) Nucleic Acid Res. 11, 2017-2033
30. Jabs, E. W., Wolf, S. F. and Migeon, B. R. (1984) Proc. Natl. Acad. Sci. USA 81, 4884-4888
31. Manuelidis, L. (1982) In: Gene Evolution, eds. Dover, G. A. and Flavell, R. B., Academic Press, New York, p.263-285
32. Wu. K. C., Strauss, F. and Varshavsky, A. (1983) J. Mol. Biol. 170, 93-117