Volume 10 Number 10 1982 Nucleic Acids Research

Nucleotide sequence definition of a major human repeated DNA, the Hind III 1.9 kb family

Laura Manuelidis

Yale University School of Medicine, 333 Cedar Street, New Haven, CT 06510, USA

Received 3 February 1982; Revised and Accepted 30 April 1982

ABSTRACT

A human Hind III I.9 kb repeated DNA fragment was isolated and cloned in pBR322. A cloned member that hybridized predominantly to the 1.9 kb Hind III band in a digest of whole human DNA was chosen for sequencing. It is an 1894 bp fragment that shows no significant internal repeats. Few pCG residues are observed in the sequence and there are numerous stop codons. Detailed sequence comparisons confirm this is a novel class of repeats that is not related to previously characterized human satellite DNAs or Alu sequences. At least a portion of the sequence described is conserved in evolution.

INTRODUCTION

A 1.9 kb band is visible in Hind III digests of human male and female DNA and by densitometry, accounts for $\geq$ 0.25% of the genome; by hybridization experiments its sequence has been shown to be distinct from previously identified human repeated DNAs including the simple DNA satellites, the Eco RI 340 bp dimer repeats and the Alu sequence family (1). The complete sequence of a cloned member of this 1.9 kb family is here reported. An accompanying paper describes minor sequence variations in the whole genomic array that are detectable in hybridization studies using this cloned DNA as a probe (2). Such studies clearly link the sequence described here to several primate Kpn I bands observed by other investigators (3, 4). The 1.9 kb sequence also is likely to be related to repeated sequences found in cloning studies of the 3' side of the βglobin gene (5) and to those identified in X chromosome cell hybrid experiments (6).

The 1.9 kb sequence shows no internal repeats of statistical significance, and in accord with previous hybridization studies, it is not homologous to any of the previously described human repeated DNA families. Related Kpn I repeats have not yet been sequenced and thus cannot be compared directly for homologies. At least part of the 1.9 kb sequence is conserved

© IRL Press Limited, 1 Falconberg Court, London W1V 5FG, U.K.

0305-1048/82/1010-3211$ 2.00/0

in evolution (1, 2, 3). The "function" of this sequence is unknown.

MATERIALS AND METHODS

The Hind III band visualized in a digest of whole human DNA was excised and eluted from a preparative 0.8% agarose gel, cloned in λ Charon 27 and subcloned in pBR322 (2). One recombinant clone that in Southern blot experiments hybridized predominantly to the 1.9 kb Hind III band in digests of human nuclear DNA (1) was chosen for further sequence studies. Plasmid DNA isolated from E. coli as described (7) was purified by equilibrium CsCl-ethidium bromide centrifugation. The DNA insert was mapped using restriction enzymes and standard agarose gel electrophoresis (8, 9). Appropriate restriction fragments were 5' or 3' end labelled as described in detail (9, 10, 11) using a series of 20%, 8% and 4% gels. Most of the sequence was analyzed at least twice (Fig. 1) and complementary strand analysis or sequence analysis in a second experiment showed unambiguous confirming data. Restriction enzyme map sites consistent with fragments obtained prior to sequencing were confirmed in the sequence.

The sequence was analyzed extensively using the sequence analysis system of the NIH MOLGEN SUMEX-AIM computer facility at Stanford University. The homology, symmetry and dyad-symmetry searches in this system are based on an updated version of the Korn-Queen algorithm; details of the parameters used and their statistical significance were generously provided and can be obtained from the MOLGEN group at Stanford (J. Clayton, P. Friedland, L. Kedes, and D. Brutlag). The system has numerous sequence files available for comparative searches.

RESULTS AND DISCUSSION

Mapping and Sequencing Strategy for the Hind III 1.9 kb Insert

The enzymes used for sequencing of end labelled insert fragments are depicted in Fig. 1. Bgl II/Hind III digests were used to sequence positions 1-700 since fragments of 400 bp and 1500 bp were well resolved from 4362 bp plasmid DNA. Similarly, Xba I/Hind III digests yielded the sequence from positions 1200-1900. Rsa I/Eco RII digests were found most useful for sequencing central fragments of the insert. Complementary strand analyses were also done using Rsa I/Bgl II digests or Rsa I/Sau IIIA (positions 1424-1667) digests. Hinf I, Hinc II, Kpn I and Taq I cleavage studies on whole plasmid or end labelled fragments were also used to determine the orientation of the insert in pBR322 and to confirm restriction sites obtained in

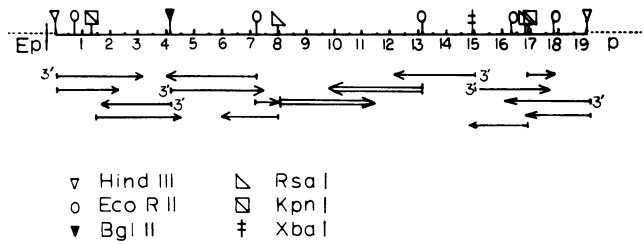


Figure 1. Diagrammatic summary of restriction gel fragments analyzed with major enzymes of use in DNA sequencing. 3' represents 3' end labelled fragments used for sequencing, all other fragments were 5' labelled. Arrows show direction of sequencing from each labelled end. Double line represents identically labelled fragments sequenced twice. Most of the sequence analysis contains complementary strand confirmation. Ep represents the end of the vector arm to position 29 (the Hind III insert site) and p represents plasmid from position 29 onwards; insert scale is 1900 bp.

the sequence.

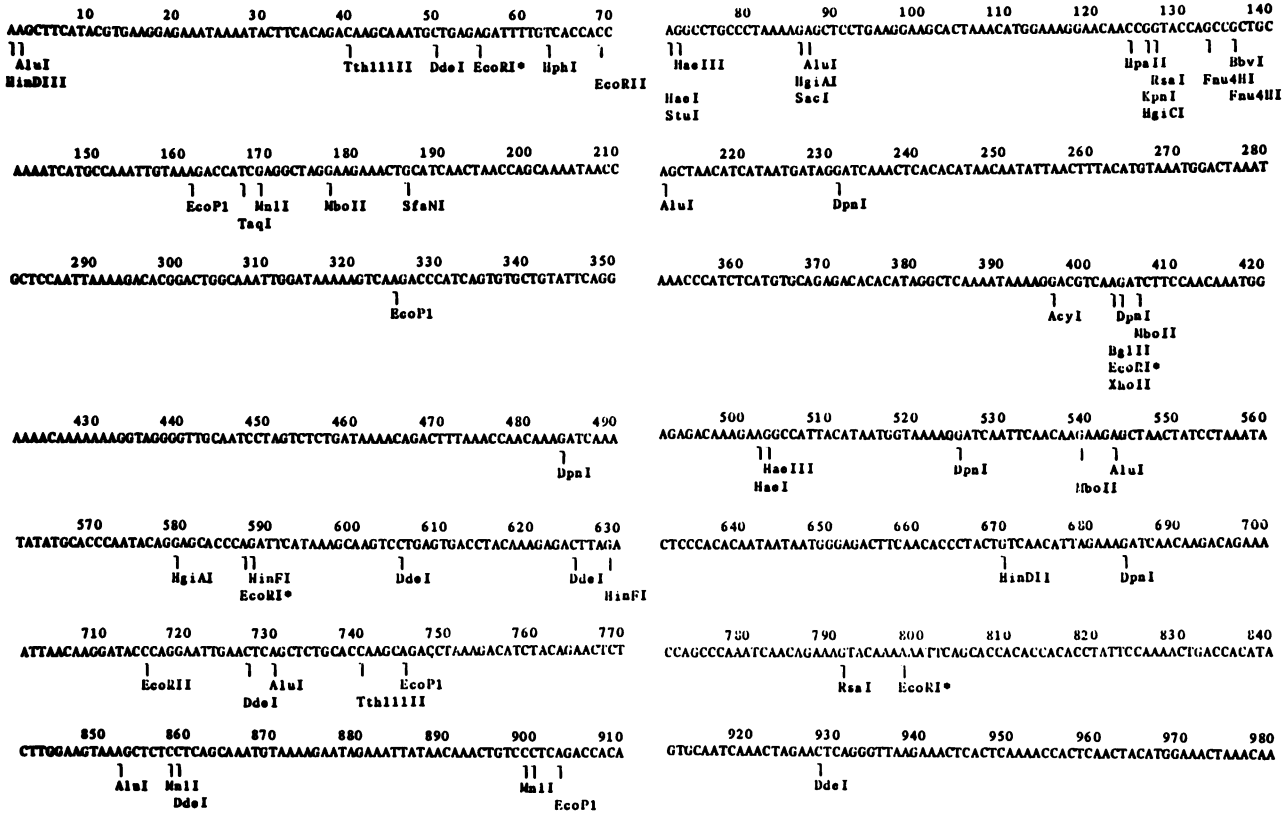
Restriction Sites of the Hind III 1.9 kb Insert

A complete restriction map is shown with the obtained DNA sequence of the 1.9kb insert (Fig. 2). The sequence determinations confirmed our initial restriction site studies indicating no Pst I, Ava I, Bgl I, Bam HI, Xho I, Ava II, Bst E II or Eco RI sites in the cloned insert. The Rsa I, Eco RII, Bgl II and Xba I sites were confirmed by sequencing. Restriction studies using Hinf I and Hinc II digestion on the cloned plasmid DNA, and Alu I, Hae III and Taq I on end labelled fragments, were compatible with the sequence obtained. Studies with whole genomic DNA have also confirmed the absence of Pst I, Bam HI, and Bst EII sites within the majority of uncloned members of this family. Hind III, Xba I, Hae III, Rsa I, Eco RII and Hinc II sites were observed at the same positions in the genomic family with minor variations; a single base change in one of the Eco Rl* sites (at position 1642 or 1652) is likely to be represented as Eco RI sites in a fair proportion of uncloned members of this family (2).

Nucleotide Composition

There were 42.6% Adenine (A), 21.9% Cytosine (C), 16.5% Guanine (G), and 19.1% Thymine (T) residues (reading 5' to 3'). Taken together this yielded a base composition of 38.4% GC and 61.7% AT for the double stranded molecule. Whole human DNA bands at a density of 1.697 in neutral CsCl (unpublished data) which is equivalent to average GC content of 39-40%. Thus this DNA is somewhat AT rich. We have never been able to purify or significantly resolve fragments containing the Hind III 1.9 kb band in high mole-

Arrows mark recognition sites, not cutting sites.



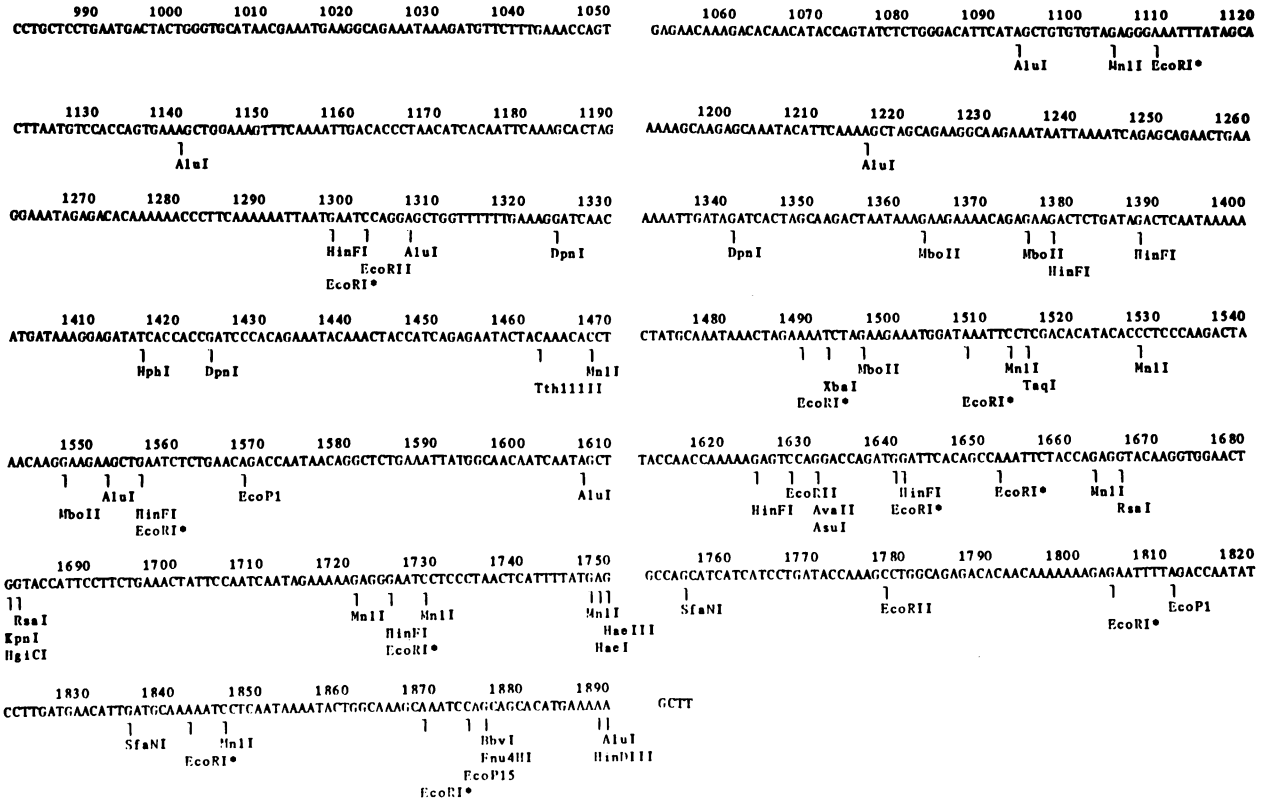


Figure 2. Complete nucleotide sequence and restriction enzyme positions of the Hind III 1.9 kb insert. It is shown as oriented in pBR322 (see Fig. 1). Eco RI* = PPATQQ.Dpn I = MboI.

cular weight human DNA using CsCl-Hoechst or Ag-$CsSO_4$ gradients (12) possibly because the repeat is integrated with other sequences in the genome.

Analysis of the dinucleotide sequence distribution was generally in keeping with the percent expected in a random sequence, with one striking exception. The pCG ratio (which is the same in both strands) was quite low and showed only 9 pCG pairs or 0.5%, as compared to an expected percent of 3.6% in total DNA. One of these pCG sites is at the Msp I (Hpa II) site (position 127). In preliminary studies using the restriction enzymes Msp I and Hpa II to differentiate methylated from non-methylated cytosines, there appears to be a small degree of methylation at this site in whole human DNA (data not shown). Only one other pCG site of the 9 in the total sequence has a methyl C residue that potentially can be studied with a restriction enzyme; at position 134 a simple inversion of CG could yield a Hha I site. Generally, the degree of potential methylation sites in the 1.9 insert is considerably lower than that seen in mouse satellite DNA (11, 13). Although pCG methylation has been considered an important feature in repeated DNAs and the inactivation of coding sequences (14, 15, 16), the low percent of pCG residues here suggest C methylation may not be an important modifying feature in the 1.9 kb sequence. The low representation of pCG sites also could coincide with special protein binding affinities of this sequence, especially if it is adjacent to active gene coding regions.

Lack of Internal Repeats

Since the 1.9 kb Hind III sequence is repeated, one might suspect that the sequence is built upon a simpler sequence, as for example that seen in the 340 bp Eco RI human dimer, or the 232 bp mouse satellite sequences (9, 11, 13). We therefore searched the sequence for internal homologies that were statistically significant (see Methods). The best homology was the match between 175-212 and 1545-1580 in the 1.9 kb insert (30 of 40 bases matched). The expected value (E) for this match as compared to a random DNA is 0.118, and showed a one base loop out (insertion-deletion) at 6 positions. A statistically significant homology would have zero base loop outs, and an expected value E of 0.05 or less. We further reviewed all the matches of 75% homology and found no repeated sequence that was obvious in these matches; there was also no consistent spacing pattern discernable in the homologies obtained. As a control of the computer program we compared its expected value for the two repeats of the human Eco RI 340 bp dimer, which using a different analysis were considered to be highly significant statistically (9). The computer also read this as a highly significant

repeat with an expect value E = 0.000.

Regions of Symmetry and Dyad Symmetry in the 1.9 kb Fragment

Several symmetries were noted in the sequence, for example at position 430. However, computer analysis of all symmetries showed there were none of statistical significance. Furthermore, there were no dyad symmetries where the total free energy of base pairing was -10 K cal; -15 K cal or less has been considered to be the limit for the stable formation of RNA hairpins under physiological conditions (17).

Translation Analysis of the 1.9 kb Sequence

The translation products in all three reading frames were examined. There were numerous stop codons. The longest reading frame (in any frame, without stop codons) was 251 amino acids long, starting at position 1139. More stop codons were observed in a search of the inverse complement of the sequence (the largest translation array was 82 amino acids in length). More detailed translation analyses (and homology and dyad symmetry analyses) are available on request.

Comparison with other Sequences

The Hind III 1.9 kb sequence was compared with the human 340 bp Eco RI dimer, in both orientations, using the sequence homology program. No statistically significant homologies were obtained. It is curious that 5 of the 12 matches (41.7%) with $\geq$ 75% homology to the 1.9 kb insert oriented as shown in Fig. 2, matched a region of the dimer sequence (ATAGAAACTAGACAGAATAAT) at positions 683, 1192, 1338, 1477 and 1485 in the 1.9 kb sequence. It is not known if this sequence region may be of importance in cross-over events or evolution of repeated DNAs. No evidence that the 1.9 kb fragment was closely related to the Eco RI dimer family was obtained from these sequence analyses, and this is in accord with our hybridization studies showing no homology between these two repeated DNAs (1). These results contrast to hybridization studies of related "alpha" satellites by other investigators (3). The best homology between these two different repeats within the same species was 76.9% (20 of 26 bases) with an unimpressive E value of 0.995 (1.9 kb positions 683-708 with Eco RI dimer positions 327-351). Analysis of the inverse complement for homologies also showed no significant similarities between these two repeats. Analysis of dyad symmetries between the human Hind III 1.9 kb fragment and the Eco RI 340 bp dimer did however show 7 matches where the pairing yielded $<$ -15.0 K cal, possibly suggesting that under less stringent hybridization conditions, some positive hybridization between the two different sequences might be obtained. Alternatively, high

molecular weight "alphoid" DNA, isolated as a satellite from cesium gradients, could also contain other repeated DNA sequences in addition to the 340 dimer family that contributed to the positive hybridization results obtained by others (3). In summary, the comparative data here does not lend any positive evidence to support the notion that the shorter Eco RI 340 bp human dimer sequence is related to the longer 1.9 kb human repeat.

We previously found that the 1.9 kb clone hybridized to an "interspersed" mouse repeated DNA band (1,18). Thus it was of interest to find if this sequence bore any resemblance to the common repeated mouse DNA satellite. Again no statistically significant sequence homologies were noted between the human 1.9 kb repeat and mouse satellite, and this was in accord with previous hybridization studies (1). The best homology between these two sequences matched positions 126-157 of the mouse sequence with 1472-1504 of the 1.9 kb sequence (75%, 27 of 36 bp, E = 0.154) as shown:

```
                    *  * * *  *      * *  *  *
1.9 kb 1472: TATG CAA ATAAACTAGAAAATC TAGAAGAAATG  1504
Mouse satellite  126: TATGGCAAGAAAA CT GAAAATCATGGAA AA TG  157
```

This homology between two widely separated species was statistically better than any homology between the human Eco RI 340 bp dimer and the human 1.9 kb repeat. "Alu family" sequences, including a polymerase III in vitro transcription unit (19) were also compared since they are "interspersed" in the human genome. Alu sequences are notably more GC rich than the 1.9 kb sequence depicted here; Alu sequences showed less significant homologies with the 1.9 kb repeat than mouse satellite DNA (sequences compared as shown in Fig. 2 or as inverse complement). Tandemly repeated human simple satellite sequences (12, 20) did not show any obvious relation to the 1.9 kb repeat.

In summary, the 1.9 kb repeated sequence is here defined as a novel human repeated DNA. It lacks any internal repeated sequence subsets, and thus cannot be classified as a tandem repeat based on its internal sequence. In the following paper we demonstrate that larger fragments containing this sequence do not show strict tandem repetition of the 1.9 kb sequence, i.e. the end of the molecule is not followed directly by the beginning of the sequence or vice versa. The "function" of this conserved repeat is unknown. It is of interest that high resolution chromosome hybridization studies directly show at least a proportion of this sequence is localized on the chromosome arms in a discrete or banded configuration (1 and in preparation). It is possible that this sequence, or clusters of this sequence, may have a role in the definition of discrete chromosomal and nuclear domains that are distinct from those occupied by the predominant centromeric

repeats.

ACKNOWLEDGEMENTS

J. Steitz generously supplied very high specific activity γ 32P ATP used in these studies. This work was supported by NIH Grant CA15044.

REFERENCES

1. Manuelidis, L. (1982). In: Genome Evolution and Phenotypioc Variation G.A. Dover and R.B. Flavell, eds. pp. 263-285 Academic Press, New York
2. Manuelidis, L. and Biro, A. (1982). Nucleic Acids Res (in press)
3. Maio, J.J., Brown, F.L., and Musich, P.R. (1981). Chromosoma (Berl.) 83, 103-125
4. Singer, M.F. (1981). International Rev. Cytology (in press)
5. Kaufman, R.E., Kretschmer, P.J., Adams, .W., Coon, H.C., Anderson, W.F. and Nienhuis, A.W. (1980). Proc. Natl. Acad. Sci. (U.S.A.) 77, 4229-4233
6. Schmeckpeper, B.J., Willard, H.F. and Smith, K.D. (1981). Nucleic Acids Res. 9, 1853-1872
7. Clewell, D.B. and Helinski, D.R. (1970). Biochem. 9, 4428-4440
8. Manuelidis, L. (1976). Nucleic Acids Res. 3, 3063-3076
9. Wu J.C. and Manuelidis, L. (1980). J. Mol. Biol. 142, 363-386
10. Maxam, A. and Gilbert, W. (1979). Methods Enzymol. 65, 504-561
11. Manuelidis, L. (1981). FEBS Lett. 129, 25-28
12. Manuelidis, L. (1978). Chromosoma (Berl.) 66, 1-21
13. Horz, W. and Altenburger, W. (1981). Nucleic Acids Res. 9, 683-695
14. Spencer, J.H., Harbers, K., and Duhamel-Maestracci, N. (1981). Advances in Enzyme Regulation 19, 453-470
15. Razin, A. and Riggs, A.D. (1980). Science 210, 604-610
16. Tantravahi, U., Guntaka, R., Erlanger, B. and Miller, O.J. (1981). Proc. Natl. Acad. Sci. U.S.A. 78, 489-493
17. Tinoco, I., Borer, P.N., Dengler, B., Levine, M.D., Uhlenbec, O.C., Crothers, D.M. and Gralla, J. (1973). Nature New Biol. 246, 40-41
18. Manuelidis, L. (1980). Nucleic Acids Res. 8, 3247-3258
19. Pan, J., Elder, J.T., Duncan, C.H., and Weissman, S.M. (1981). Nucleic Acids Res. 9, 1151-1170
20. Frommer, M., Prosser, J., Tkachuk, D., Reisner, A.H. and Vincent, P.C. (1982). Nucleic Acids Res. 10, 547-563