Biochemical Genetics, Vol. 24, Nos. 1/2, 1986

Characteristics of Site Variation Among Clones of the 340-Base Pair, Tandemly Repeated *Eco*R1 Family of Human DNA

N. Burr Furlong,[1] Koenraad Marien,[1] Ben Flook,[1]and Jim White[1]

Received 25 Feb. 1985—Final 9 Aug. 1985

Twenty-four clones of Eco*R1-restricted, 340-base pair (bp) DNA derived from human DNA have been sequenced and compared to a published consensus sequence for this family. No two clones were found to have identical sequences; the clones studied differed from the consensus sequence in as few as 1 or as many as 41 sites. On the average in these clones, a 5.2% divergence from exact homology was found, with 1 of 10 of the site variations being "nonrandom," i.e., cases in which five or more clones had the same nucleotide substitution at that site (viz., 53, 124, 126, 138, 152, and 157). At site 157, for example, 16 of the 24 clones differed from the reference sequence. Positions and their respective changes, as compared to the consensus sequence, are summarized. Variations are discussed with relation to possible functions for these sequences.*

KEY WORDS: human DNA: alphoid sequence; tandem repeats; nucleosomes; *Eco*R1 restriction.

INTRODUCTION

Restriction analyses of DNA from human cells have shown "ladders" characteristic of tandemly repeated DNAs for *Eco*R1, *Hae*III, *Hin*f1 (Manu-

This research was supported in part by Center for Disease Control Grant 1 K01 OH 0035-01, and K.M. was awarded a fellowship from NIOSH Educational Resource Center Grant 5-T15-OH07085.

[1] The Graduate School of Biomedical Sciences and Department of Biochemistry System Cancer Center, The University of Texas Health Science Center, P.O. Box 20334, Houston, Texas 77025.

elidis, 1976), *Hind*III (Maio *et al.*, 1981a) *Kpn*1 (Maio *Et al.,* 1981b), and *Bam*H1 (Willard *et al.*, 1983). The *Eco*R1 family has been shown to be concentrated in centromeric regions of certain chromosomes and to be largely homologous with a major component of the alpha satellite of African green monkey DNA. In human cells the *Eco*R1 family represents over 1% of the total DNA and is present in over 100,000 copies per cell.

The major band obtained from *Eco*R1-digested human placental DNA was reported to consist of 340-basepair (bp) material for which a consensus sequence was published (v.s.). The observation that individual clones differed from this consensus led us to examine the nature of base variations among the members of this family. The results of our analysis of the nucleotide sequences of 24 clones of DNA from this source are reported in this paper. We found that six positions differed from the consensus sequence at frequencies greater than could be expected from a random distribution of variants.

MATERIALS AND METHODS

DNA Extraction. DNA was extracted by homogenizing human placental tissue in a Servall Omnimixer with 15 vol per weight tissue of TNE (10 mM Tris–HC1, *p*H 8.0, 150 mM NaCl, and 10 mM EDTA). The homogenate was centrifuged at 2000*g* for 5 min and the pellet was resuspended in 40 pellet volumes of TNE and treated with proteinase K (0.2 mg/ml) and sodium dodecyl sulfate (0.5%). After incubation overnight at 37°C, the solution was extracted twice with equal volumes of TNE-saturated phenol, chloroform, and isoamyl alcohol at a ratio of 25:24:1 (PCI). Then 3 M ammonium acetate was added to 0.3 M, followed by 2 vol of 95% ethanol. After 30 min at −20°C, the precipitate was spun down and washed with 1 volume of 70% ethanol at −20°C and vacuum dried. The precipitate was resuspended in 20 residue vol of 10 mM Tris–HC1 (*p*H 7.6), 1 mM EDTA, (TE); NaCl was added to 0.15 M, and RNase to 50 μg/ml. After 1 hr at 37°C, the solution was extracted with an equal volume of PCI, precipitated, and washed as above. The vacuum-dried DNA was resuspended in a small volume of TE and stored at 4°C. Enzymes were from BRL, Inc.

Eco*R1 Restriction and Electrophoresis of Human DNA.* Lyphozyme *Eco*R1 was prepared as indicated by the BRL protocol to give a concentration of 7 U/μl. Human placental DNA was redissolved in TE, made 0.1 M in NaCl, and incubated at 37°C for 2 hr with a fivefold excess of restriction enzyme. The reaction was stopped with 0.1 vol of 50% glycerol, 0.1 M EDTA, and 0.1% bromphenol blue. After heating at 60°C for 5 min, the solution was loaded onto a 1% agarose horizontal gel run at 200 V for 10 min and then at 150 V until the dye had gone seven-eights the distance toward the anode. The gel was stained with a 1-ppm ethidium bromide solution for 20 min and viewed under uv to mark the 340-bp DNA band with India ink.

Isolation of 340-bp DNA from Agarose Gel. The 340-bp band was excised from the gel, mashed, placed in a sealed container with excess extraction solution (0.5 M NaCl, 0.05 M Tris–HC1, *p*H 7.3, 0.01 M EDTA), shaken, and incubated at 37°C overnight. The solution was clarified through a glass filter and then washed several times with DNA extraction buffer. To precipitate the DNA from the filtrate, 2 vol of 95% ethanol was added and chilled to −80°C, after which the solution was centrifuged at 4°C for 10 min at 5000*g*. The pellet was washed with 70% ethanol at −80°C and recentrifuged. This pellet was then dissolved and stored in TE buffer at 4°C.

Ligation, Transfection, and Sequencing of the 340-bp Unit. Ligation of the 340-bp DNA strand into M13 and transfection of the M13 into *E. scherichia coli* were done as described by Messing (1983) using the M13 mp 10 and mp 18 strains grown in JM 101 and JM 105 *E. coli* strains, respectively. Sequencing of these cloned fragments was done by the dideoxy method (Sanger *et al.*, 1977).

Sequence Comparisons. Sequences were entered and stored in standard DNA data-bank format using an IBM personal computer. The location and type of variations that occurred between a given cloned sequence and the consensus sequence were easily scored by aligning them on the screen to the closest match and printing the screen. Restriction sites and matrix comparisons were determined using the DNA analysis program written by L. L. Domier.

Testing for the Detection of Ambiguities. In order to test that ambiguities would actually be detected by these procedures, DNAs from 20 clones which had been C-tracked to allow the choice of similarly oriented strands were combined and an aliquot of the mixture was sequenced.

To check for reading errors, all sequences were read by at least two individuals and then reviewed to resolve discrepancies. To evaluate the reproducibility in the sequencing procedure, two clones were sequenced twice and compared. Clones sequenced twice and read twice were found to require fewer than 10 rechecks.

RESULTS

Figure 1 is a listing of the previously published consensus sequence (v.s.) for the human placental, *Eco*R1-restricted, 340-bp sequence family. Using the same numbering system as that used by Manuelidis (1976), those positions which were found to be different in the 24 clones sequenced in the present

```
A. Variations Occurring Between Positions 1 and 59 (11 Clones):
        1        10        20        30        40        50       59
        AATTCTCAGTAACTTCCTTGTGTTGTGTGTATTCAACTCACAGAGTTGAACGATCCTTT
     G   1          2       1 1    1  1       2      1   1    2
     A    1      1   1   2   2                         2      11  1
     T                  1    1                         11  25
     C             1           1     1 1             21    11
B. Variations Occurring Between Positions 60 to 84 (18 Clones):
        60        70            84
        ACACAGAGCAGACTTGAAACACTCT
     G  1411    1  1    1
     A                        2
     T   211   11   3   2      1
     C        11 1
C. Variations Occurring Between Positions 85 to 250 (24 Clones):
        85             100       110       120       130       140
        TTTTGTGGAATTTGCAAGTGGAGATTTCAGCCGCTTTGAGGTCAATGGTAGAATAG
     G  13          2 1 2 2           4  1        2        2
     A  1221                         3  31   1 51    31      a
     T               2      11     1   11      11      1  1
     C    1  1 11    1  11  1 1 2 2            1 5 2 3113     1
        141      150       160       170       180       190
        GAAATATCTTCCTATAGAAACTAGACAGAATGATTCTCAGAAACTCCTTTGTGAT
     G       1     b 1 1  1 2   1    1   11 1        4111
     A            121   c                          1 31
     T               2 1            1        1    1  42
     C           1   1121    2 1      13    3          121 3  2
        196           210       220       230       240       250
        GTGTGCGTTCAACTCACAGAGTTTAACCTTTCTTTTCATAGAGCAGTTAGGAAAC
     G                  11         2   1 1         1
     A      1                                   1        1
     T    1  2    111  2  1        2           1   2    4
     C    11  1    1        11       1          2 1         1
D. Variations Occurring Between Positions 251 and 340 (14 Clones):
        251      260       270       280       290            305
        ACTCTGTTTGTAAAGTCTGCAAGTGGATATTCAGACCTCTTTGAGGCCTTCGTTG
     G              2
     A                2        1            3 21 1
     T       1 1           1           42 1        11  2
     C      1    12   1      1       1       1 3
        306           320       330       340
        GAAACGGGATTTCTTCATATTATGCTAGACAGAAG
     G                  11  1 2
     A                   1  1
     T                  1                2
     C        1      2   1 2      1
```

Fig. 1. Base differences occurring in clones of human 340-bp DNA. a = 14; b = 11; c = 16. Site numbers refer to positions under the leftmost digit of the number.

study were tabulated as positions of variation. Apparent insertions or omissions were not scored since these invariably involved sites which were not clearly distinguishable on the film. The changes that occurred were found to be dispersed throughout all the clones sequenced, not just in a select few.

The sequence data of clones resulting from "minus" strand ligations were converted to the inverse complement and renumbered to facilitate direct comparison with the positive strands. Some 353 differences from the consensus sequence were scored of 6846 nucleotide positions surveyed. This is an average of 5.2 differences in each 100 sites sequenced, or about 18 differences in a 340-bp clone. The probability that the same site would be different in any two clones is approximately 0.0027; for three clones, 0.00014. The occurrence of a difference at the same site in more than four clones is considered too improbable to justify an assumption of randomness. There are 37 nonrandom and 316 random variations scored in the 24 clones sequenced under this definition. The actual rate of occurrence of nonrandom variations in these clones is probably higher than this if we assume that the data from regions where fewer than 24 clones could be sequenced are representative. For example, the four G-for-C substitutions at position 61 (data from 18 clones) and the four T-for-C changes at position 282 (14 clones) would extrapolate to greater than four occurrences if data from 24 clones in these regions were available.

Extrapolated values of the number of base variations with respect to the reference sequence were calculated for each region of 10 sites along the 340-bp sequence. Figure 2 is a representation of these data plotted along

```
                                (Totals)
   (7  15  17   9  20  33  21   9  16   8  12  17  25  25   2  42   8)

40                                                              *
                        *                                       *
30                      *                                       *
                        *                           *   *       *
20                  *   *   *                       *   *       *
        *   *       *   *   *       *           *   *   *       *
10      *   *   *   *   *   *   *   *   *   *   *   *   *       *   *
  1 *   *   *   *   * 51*   *   *   *   *101*   *   *   *    151*   *
  I---I---I---I---I---I---I---I---I---I---I---I---I---I---I---I---I---I
  I---I---I---I---I---I---I---I---I---I---I---I---I---I---I---I---I---I
 171*   *   *   *   *221*   *   *   *   *271*   *   *       *321*   *
10  *   *   *               *           *       *   *           *
        *                               *       *               *
20      *                                       *               *
        *                                       *
30                                              *

  (11  23   9   7   7   7   8   6   7  14   5  29   9   0   5  19   5)
                                (Totals)
```

Fig. 2. Distribution of variations along each dimer.

corresponding regions of each of the two dimers that make up the 340-bp sequence. Variations with respect to the published sequence were used for this comparison.

DISCUSSION

Various repeated sequences contained within DNA have been hypothesized to be functional in (a) determining higher orders of chromosome structure, (b) determining transcriptionally active or inactive regions, (c) determining recombinant events, and/or (d) phasing the folding of DNA in nucleosome formation. The latter was suggested as a function for tandemly repeated DNA by Maio *et al.* (1977) and was supported by Strauss and Varshavsky (1984), who isolated a protein which binds strongly to the 172-bp repeat of African green monkey alpha satellite DNA [which Wu and Manuelidis (1980) have shown to possess 65% homology with the 171-bp dimer of human 340-bp DNA]. DNAse footprinting detected three specific binding sites, each of 7-bp extent.

In the numbering system of the present paper, the positions and sequences corresponding to the location of these sites in the dimers of human DNA are as follows: site I (TTAATTC) occurs at positions 28–34 (TGTATTC) and 199–205 (TGCGTTC); site II (AAATATC), at 142–148 (identical and 313–318 (GATTTC; note one base missing); and site III (GATATTT), at 105–111 (GAGATTT) and 276–282 (GATATTC). In the monkey satellite DNA, sites II and III are inverse complements, which by their separation in tandem repeats of about 145 bases, suggested a function in nucleosome formation. The sequences in homologous regions of human DNA do not exhibit the same precise complementation but, perhaps, could bind similar proteins since their homology differs by no more than the difference between site I and site II. In this regard, it is interesting to note that a single base change in the 7-bp sequences beginning at positions 105 and 279 will make them identical to site III and that these are located at or near the positions which correspond to those of the monkey satellite. There are 17 other 7-bp sites where a change of two bases will result in a sequence identical to one of the three sites where protein binding occurred in satellite DNA. Most of these are close to the sites already defined.

An analysis of base variations among the clones we have sequenced in the regions corresponding to satellite binding sites reveals that 4 of the 14 clones sequenced at 282 have the C-to-T change, making this region exactly match the site III sequence (GATATTT). Other regions which correspond to these three binding sites in our clones show few variations. Our clones have a total of 23 base changes within the six regions listed and only in the 4 cases at site 282 does the change make the 7-bp sequence homologous with a site in monkey satellite DNA. There were 763 total nucleotides sequenced in these six regions; thus, the variation rate for this subset of the total nucleotides sequenced is 0.030, which is somewhat lower than the average variability (.052).

Banerji *et al.* (1983) identified 5′-GTGGTTT-3′ as a putative enhancer core sequence for stimulating immunoglobin gene expression in B lymphocyte-derived mouse cells. Similar sequences, involving TGG(AAA/TTT), have been identified as host cell-specific, cis-acting, transcriptional activators in various viral genomes (Weiher *et al.,* 1983) and homologues have been identified in human DNA (Conrad and Botchan, 1982). The DNA family investigated in the present paper has been associated with heterochromatin and would not be expected to be transcribed. There are, in fact, from two to eight stop codons in all reading frames of the consensus sequence. However, the ability of enhancer elements to influence transcription several kilobases downstream leaves the possibility open that such effects could be exerted by nontranscribed DNA. Thus, it may be of interest that there are two regions of the 340-bp DNA analyzed in the present paper that contain sequences similar to these enhancers: 89–95 is GTGGAAT and 273–279 is GTGGATA. There are several associations between potential enhancer homologues and the protein binding sites described above. The sequence beginning at 273 overlaps the site III binding region described in the last paragraph and a single base change in the sequence at 27–33 (a region overlapping site I binding) of T to G at position 30 would make this region homologous with the putative enhancer sequence.

Our data on variations of nucleotides in the two regions homologous with a possible enhancer sequence reveal that there are only 4 base changes in the 266 nucleotides analyzed in these regions of the 24 clones. This limited sample suggests that these regions may be more highly conserved than the average, for which we would have expected about 14 random changes in 266 nucleotides.

Sites 138, 152, and 157 have variations with respect to the consensus sequence that occur in 45 to 67% of the clones, indicating that there could be several major forms. It is obvious from Fig. 2 that there are more variations in the first 171 bases of the 340-bp sequence than in the last 169. It is also evident that variation along the sequence is nonuniform. For example, in the eight positions including 150 to 157, there are 42 variations among 24 clones, whereas the 21 sites between 299 and 321 show but 1 difference in 14 clones. Relatively few differences are found in two corresponding dimer regions, 11–20 (181–290) and 141–150 (311–321), whereas relatively many differences are found in three regions, 11–20 (181–190), 111–120 (281–290), and 151–160 (321–330). The first dimer shows regions of variability that are not reflected in the second dimer at positions 11–20, 111–120, and 151–160.

Further work is being directed to obtaining additional sequence data to improve the sampling statistics and aid in the classification of regions into those that are more variable or more conserved.

REFERENCES

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell* **33:**729.

Conrad, S. C., and Botchan, M. E. (1982). Isolation and characterization of human DNA fragments with sequence homologies and with the simian virus 40 regulatory region. *Mol. Cell. Biol.* **2:**949.

Maio, J. J., Brown, F. L., and Musich, P. R. (1977). Subunit structure of chromatin and the organization of eukaryoyic highly repetitive DNA. *J Mol. Biol.* **117:**637.

Maio, J. J., Brown, F. L., and Musich, P. R. (1981a). Toward a molecular paleontology of primate genomes. I. *Chromosoma* **83:**103.

Maio, J. J., Brown, F. L., McKenna, W. G., and Musich, P. R. (1981b). Toward a molecular paleontology of primate genomes. II. *Chromosoma* **83:**127.

Manuelidis, L. (1976). Repeating restriction fragments of human DNA. *Nucleic Acids Res.* **3:**3063.

Messing, J. (1983). New M13 vectors for cloning. In Wu, R., Grossman, L., and Moldave, K. (eds.), *Methods in Enzymology, Vol. 101, Part C,* Harcourt Brace Jovanovich, New York, pp. 20–78.

Sanger, F., Nicklen, S., and Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci.* **74:**5463.

Strauss, F., and Varshavsky, A. (1984). A protein binds to a satellite DNA repeat at three specific sites that would be brought into mutual proximity by DNA folding in the nucleosome. *Cell* **37:**889.

Weiher, H., Koenig, M., and Gruss, P. (1983). Multiple point mutations affecting the simian virus 40 enhancer. *Science* **219:**626.

Willard, H.F., Smith, K.D., and Sutherland, J. (1983). Isolation and characterization of a major tandem repeat family from the human X chromosome. *Nucleic Acids Res.* **11:**2017.

Wu, J. C., and Manuelidis, L. (1980). Sequence definition and organization of a human repeated DNA. *J. Mol. Biol.* **142:**363.