Sequences flanking the repeat arrays of human minisatellites: association with tandem and dispersed repeat elements

John A.L.Armour, Zilla Wong, Victoria Wilson, Nicola J.Royle and Alec J.Jeffreys

Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, UK

Received April 26, 1989; Accepted June 1, 1989 EMBL accession nos X14856–X14867, X14953

ABSTRACT

We present DNA sequences flanking cloned hypervariable human minisatellites. In addition to providing confirmatory evidence that minisatellites cluster with other tandem repeats, these flanking sequences contain a high frequency of interspersed repetitive elements. These elements include a retroviral LTR-like sequence, from which one of the minisatellites appears to have expanded, and a recently described short interspersed repeat. We present our own findings concerning this element, in particular that those examples studied do not show significant evolutionary conservation, despite suggestions that the element may have a *cis*-acting function.

INTRODUCTION

Hypervariable minisatellite regions of the human genome [1–3] have greatly accelerated progress in analyses of individual identity[4], family relationship [5] and linkage mapping [6]. The usefulness of these loci would be greatest if they were dispersed randomly in the human genome. Initial analysis in large kindreds, based on simultaneous detection of many hypervariable loci by DNA fingerprinting, showed that the majority of the loci detected by the 'core' probes 33.6 and 33.15 appeared to show independent assortment in the offspring [7]. At the resolution afforded by segregation analysis, therefore, the loci detected appeared to be well dispersed in the human genome.

However, more recent studies using *in situ* hybridisation with highly variable cloned minisatellites showed that the most variable minisatellites are preferentially located near the telomeres of the human autosomes [8]. Similarly, inspection of published linkage maps which include highly variable minisatellite or VNTR markers [6,9,10] shows that while some minisatellites do occur at interstitial locations, the majority cluster towards the ends of the linkage maps. Furthermore, two examples of pairs of proterminal minisatellites showing very close physical linkage have been observed [8], suggesting that these subtelomeric chromosomal regions may be rich in variable tandem-repetitive sequences.

In this paper we describe the DNA sequences flanking six of the most variable minisatellites we have isolated. Analysis of these, together with one previously reported [11], extends the evidence for the clustering of minisatellites with each other, particularly of proterminal minisatellites, and shows that they tend in turn to be associated with dispersed repetitive elements. One of these elements has only recently been reported [12], and our own findings on this element are presented.

© IRL Press

Table 1. Properties of minisatellite loci detected by locus-specific clones

		per cent	chromosomal	bp DNA sequenced	
probe	locus	heterozygosity	location	5′	3′
pλg3	D7S22	97.4	7q36-qter	438	313
pMS1	D1S7	99.4	1p33-p35	434	1284
pMS31	D7S21	98.0 (A)	7p22-pter	12	399
		50 (B)		8	(12)
pMS32	D1S8	97.5	1q42-q43	207	429
pMS43	D12S11	95.9 (A)	12q24.3-qter	14	780
		30 (B)		(780)	373
pMS51	D11S97	77	11q13	113	194
pMS228	D17S134	94 (A)	17p13-pter	–	–
		85 (B)		400	1057

Note: in pMS228 only DNA flanking 228B has been sequenced (see figure 1). Restriction mapping suggests that only about 500bp immediately flanking 228A have thereby been omitted.

MATERIALS AND METHODS

Minisatellite clones The cloning and characterisation of pλg3 has already been described [11]. pMS1,pMS31,pMS32 and pMS43 are plasmid subclones (into pUC13 [13]) of the minisatellite-containing *Sau*3AI inserts from the λMS series already described [2,8]. λMS228 was similarly isolated from a library of large (5–20 kb) *Sau*3AI fragments of human DNA cloned into λL47.1 and screened by hybridisation with the DNA fingerprinting probe 33.15 [14].The *Sau*3AI insert was subcloned into the *Bam*HI site of pUC13 to give pMS228.

pMS51 [15] was isolated as a positively hybridising clone from a library of size-selected 2.5–4kb *Eco*RI/*Sau*3AI human DNA fragments cloned into pAT153 [16] and screened with DNA fingerprinting probe 33.15. Further characterisation of pMS51 and pMS228 is described in 'Results' and summarised in Table 1.

Hybridisation of filters and washing to high stringency was performed as described [2].

In situ hybridisation was performed as described [8].

DNA sequencing was performed using the dideoxynucleotide chain termination method [17] with Klenow fragment, modified T7 polymerase [18] or DNA polymerase from *Thermus aquaticus* [19]. These sequence data will appear in the EMBL/GENBANK/DDBJ sequence databases under the accession numbers X14856–X14867 and X14953.

Computer sequence handling. In database searches, the EMBL database (version 14) was scanned using a FASTN program incorporating the algorithm described by Lipman and Pearson [20]. Other analyses were performed using the programs developed at the University of Wisconsin [21].

RESULTS

Characterisation of pMS51

pMS51 was used to probe *Hae*III-digested DNA from 40 unrelated individuals of French and Mormon origin taken from the CEPH panel of families. Under high stringency conditions [2] it detected a variable locus with alleles ranging in size from 1.7 to 5 kb. At least 9 different alleles could be distinguished in the individuals surveyed, and the population heterozygosity at this locus is estimated at 77%. The probe also cross-hybridises to a monomorphic 0.8 kb fragment of unknown origin. Hybridisation with DNA from

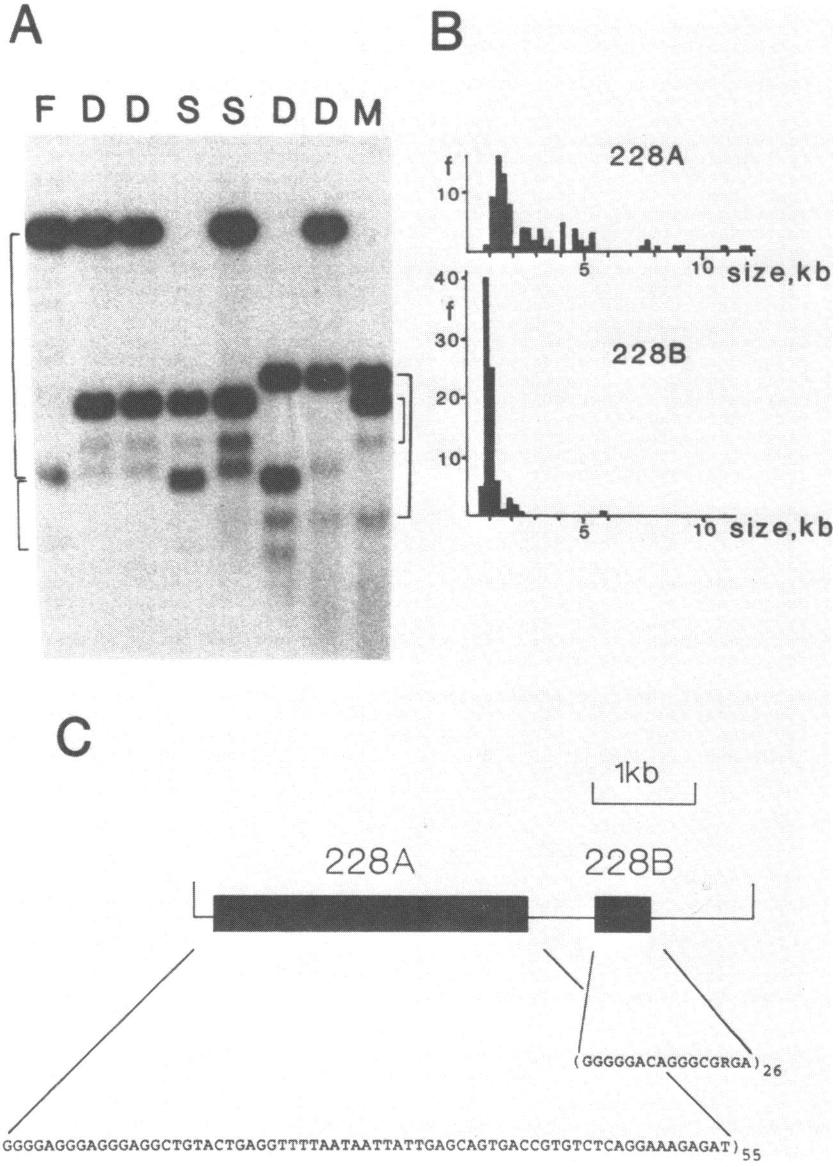

Figure 1. Characterisation of pMS228.(A) DNA from a family group [F=father,M=mother,S=son,D=daughter] digested with *Alu*I and probed with pMS228.Both strongly and weakly hybridising alleles are seen;the strong and faint alleles which cosegregate as linked pairs among the children are shown bracketed in the parents. (B) *Alu*I allele size distributions of 228A and 228B. The allele sizes were determined in 89 (228A) and 48 (228B) unrelated individuals from the CEPH kindreds (see 'Results'). In the histograms, the alleles are allocated to 0.25kb size classes; since each size class may contain many resolvable alleles, the number of alleles at each locus is much more than the number of classes shown.(C)Structure of plasmid clone pMS228.Boxed regions represent the minisatellite arrays, below which the corresponding repeat unit sequences are shown.228A detects the more strongly hybridising bands and 228B the fainter bands on high stringency hybridisation of human DNA.

```
 (A)
               .         .         .         .         .         .         .
     1 tttttttgtagagacactatgttgcccaggatggccttgaactcctggcccaagcaatcctcctacctcag
    73 cctgccaaagtgtttggattacaggagtgcaccactcctggcctgattaccttcattttacagatagggaaa
   145 ctgacagccggggaggaggccatgagcacatggtgccggtggccaaggagcttggaactctttccctgcccc
   217 tggttctctgtcccttcccctccccgacctggcacagctgttttcctgctgcttccacacccacagctccc
   289 tggcacatgcactgctgagagcccaggggcacaggtctgttcttagcatggttcttacagcacaggcagcaa
   361 agaagcatatgcaacccatgagggccaggaccacccaatctgggctcccatgcatgagatatccatatggta
   433 tttGGTAGATAGxxxxxxxxxxxxxxxxTGGATAGGacccactgactcacagcaaggggtgatttctcaatt
   505 tgctggggtggcggggctggaatccagctctccttttcactttgtttggaaacgtgagttggtttttcaaa
   577 aatttttatgtctcaaactatcaaattctgttaaaaacatcaaggctcatcacagaaaagcatgacatacaa
   649 tcattactgttgttctttttttagtgtaagcatgaaggaaagggaggccattggtaaaaccacctgacacta
   721 ataaaatggaaacttaaaatcaatggcaaatgagcgtgattaattattaacattcttaatggatgtagtacc
   793 tggtatggtactttgtagtattaggtactcgatacgtgtttatggaaatggcaaaggaacaatcctaacaat
   865 agtgactgttgattgaatttatattctgtgccaagcactatgctagagaccttacatcattattccgtctag
   937 tcccccagcagccctctagaggaggtacgattatcccttatttacagacaaggaatcagaggctcagaga
  1009 ggttaagtgttttgctcaaggtcacacagaaaatgggagtgggggtaggggtagagttgagaggcaaagcag
  1081 tttgttaactcccaggcccagtaaccgggtcgctggcctctgaggctctgggaccatttgtgtccagtgtta
  1153 tgaccaacgccagagctatctatttgcccaccttcctccctcccctgcagtctgcgctcctcagaccaccgg
  1225 caggtgaacagcctgatgcagaccgaggagtgcccacctatgctagacgccacacagcagcagcagcagcag
  1297 gtggcagcctgtgagggccgggcctttgacaacgagcaggacggcgtcacctacagctactccttcttccac
  1369 ttctgcctggtgctggcctcactgcacgtcatgatgacgctcaccaactggtacaagtgcgtagctggtggg
  1441 gcatggacagagcccggaggtgcagggtgcaggtccacacatctctcctgcggcccttcactgggtttcct
  1513 gatgtctgcttaaggcttgggggtggcggtgtgtatcaaggaccctccatagagcctgccccttccctgcta
  1585 cagtttccccaacagtgccttcatctgctcagctcccgcaccttcactaggagactgaaccggggaaggcac
  1657 agcagctggggctggacctgtacagggccaggcgggggtttacaaatcaggggtcatggggtcagatttggg
  1729 cttcatggttggagggttaggatc

               .         .         .         .         .         .         .
  (B)1 gatccccaCAGCACGCCGTCCCCACACCTGCACCCCGTCCCCACACCTGCACGCCGTCCCCACACCTGCACC
    73 CCGTCCCCACACCTGCACCCCGTCCCCACACCTGCACCCCGTCCCCACACCGGCACACCGTCCCCACACgcc
   145 catccggccGGCAGTGTCTGTGGGAGGTAxxxxxxxxxxxxxxxxxTGTCTGTGGGAGGTGGACAGccaagg
   217 ccaggtccccttcatggcccctcctgacccccccagcccctcctgagcccaaggcgaaaggcaaaacgccacc
   289 gtccagcgtgcaaagggaagagccaacattgtgcatgcggcttcccctgcctaggttgtgcctgtggctccc
   361 gggccagagccacctggggagccgtgggctctgagggaggcccggcaggaaaccccagcaggccggactgag
   433 tgggggaggggccgtggctcccagaccgggcctcacacaggacttcagcatcttcctcgggtcctccagcca
   505 catccccatgtccaggccactttcccagggcctcagacaaggctggcttcaatccttccagaacattctagg
   577 cttgctcctaactgcaggtggttccgagtggatc

               .         .         .         .         .         .         .
  (C)1 gatcttcgtgtctattcttctaccaataccaccctgtcacgtctattgtatctttctagtaaatcttaaaat
    73 tgggtaacataagctctccatttccagtttctggaaaaatttgtgtagaatttgttgtaaataaatttttgg
   145 tgctgcaaaagaaataccactcaaacataagtttaatttctcagcaaggcaattttacttctctagaaggg
   217 tgcgactcgcagatggagcaatggccagagcacacctgaacaagggaggggaaggggttctgattcctgaca
   289 caggtagcccctactgatgcgtcgttcccgtatcggctagggttggactgcacagtctaagctaattccgat
   361 tggctactttaaagagagcaggggtatgagccagagtggcggggtgagtagtttggtgggaagggtggttAC
   433 AGAACAGGTGACTCAGGATGATTCAGGTxxxxxxxxxxxxxxxTGACTCAGAATGGAGCAGGTGACCAGGGG
   505 aatagacgttaactactgattagaactgttggaaaaggttgtttagtgaaactagggctgaggagaacgagg
   577 aagttcaactttaaaatggagaacaaagaactgaacatactgacatactgattctttgaagagaaatttaga
   649 actcactgtattcaacaaattattatttttgcttttaagtgtctgtggaattcaccggtgatc

               .         .         .         .         .         .         .
  (D)1 gatcagcgaacttcctctcggctcccgatatcctcctcgatacgcactctgccacaacgggcagggtccctt
    73 tcagcgtctcatccacagtgaacgggagttgaggctttcttagcggaggggcTGGAGGGACACAGCAGGAGG
   145 GCAGGAGGGCAAGxxxxxxxxxxxxxxxTGGAGGGACAGAGCAGGAGGGCAGGcctccctgcggtttccgga
   217 tgctacggggtggatcggagtgtggtgttaagcacatctggacacgctctgtccgagacacatagtccccag
   289 gcgacctacagccacagcctgacctcctgaaaatttcccagcttcccacagtcctcaatgtggaaaccagtg
   361 cccaaaggccacctgcccacactggcaccgaattc
```

Figure 2. Sequences flanking human minisatellites. The tandem repeat block is abbreviated to the two outermost repeats (in capitals) separated by x's.(A) pMS1 – novel dispersed repetitive element underlined, Alu element in bold italics (B) pMS31 – the tandem repeat 31B near the beginning of the clone is shown in full in capitals, unlike the major minisatellite ('31A') which is abbreviated as above (C) pMS32 – retroviral LTR underlined, L1 sequence in bold italics (D) pMS51

a panel of human-rodent somatic cell hybrids [2] assigned the locus to human chromosome 11; *in situ* hybridisation localised it further to the interstitial region 11q13 (data not shown).

Characterisation of pMS228

pMS228 contains a 5.7 kb human *Sau*3AI fragment. This insert hybridises to two variable loci in human DNA (Figure 1A). One is detected strongly and is represented by *Alu*I alleles ranging from 0.8 to 12 kb in length, with a heterozygosity of 94% determined from 89

```
(A)            .         .         .         .         .         .         .
1     gatctatacatgtTTACACACATGCCACACACCCTTCCCAAGGCCGTCCCTATACCCAxxxxxxxxxxxxxx
73    xCCCGGGGGCCATCggcgcagtggcctcagcgtccagcctccccgtcctccggagctctctctggtggggc
145   cccagtcctcctcgtcttccggagctctctctctctggtggggccccacctggccgacctgaaggtcaggg
217   catccactctttggggaaacgaagttgggttcgtaggggcagctctggtgttgttggtccttccctccagac
289   accacgtctcttcgtctggctcccaggagaccccagttctatttctatttctcccgcctctgtcatcccctc
361   aaactgcttctgcatagcccccactctgcctctcaggctttaaccccaactgtgagccccacctggctgtcc
433   tggggccactgaactccacgtgtccatcgtggctctccccaacatgccaccacgcaccggctggcattctgt
505   ccctatgaagccctgcctgcctggggactgggaatcagccgctgtcccagcccacccgtcaatcacccatca
577   atcaccggccaatcacccgtcaatcaccgtctcttgccagctttgcatcctaaatatttttggaagagtcc
649   cgtgtccgccacaccccacaagaaagaaagttcaccctactgggctcaacagtaaagatgttaatgaaggga
721   agcttagggatgtgtggttaggggcatagggacagacgggggcagtgcaacctcaggggcggcaacgtgcag
793   ttctcgacaccggcagagtgaggcctcgggcgccactcggaccacggtgcagcttccaacaacctgtggtgc
865   ctccacgggccaagacagcctgcgtgcctccacgggccaagagcacaaggtgggaggagggaaggagctggg
937   ggactggaaagcggtggtcagcatctgagggcccggcacgTGGGGCTAGACGGGGGATGTGGTGAxxxxxxx
1009  xxxxxxxxCGGGGGCCTGGTGATGGGGacaggtgagtctcctgtccctcaccaggagaactgggttctggtt
1081  ctcagcctgtctcccaccgtagagcaacctctgaccgccagacaggacaccaaggtgaaaactttcaagccc
1153  ccaggacgggagagaggtcagggggagaggtcgggggagaggtggcagggagggcagggaggtcagggagtg
1225  gagtaaacgctgggcgcggccacagctgtgccttgtcaggggatgaaggtgaaagacagaacccgcggtgac
1297  acaggctcccctgccccgcggaggagtcaggccatcagaaaggcccctatgtcagccaggcgtggtggttca
1369  cacctggaatcccagcactgtgggaggcagaggcgggcagatc

(B)            .         .         .         .         .         .         .
1     aggccacccccaccaacaacataaccagaggggaaggaagtcaggccccttctcactctgagaagctggtgc
73    ctgggatttaggctgtcacgacgattccacccggccagggcaggcccgaaccggccggaggccacaggaga
145   accaatgagcctggctggactcctgcaaaccttgagggacgccgagattcacattcactgaattttcatgcc
217   acgtgacactcttcttcttttgttccctcccgccctgccccacggagccatttacaaacataaaaccatttt
289   taaaccatttttaaatggtttaaaacctgtttctgtaccaagtggtacagaaataggggccagcccccggtc
361   cagcgccacgagctcctagggccagagtgcaagagaggcACAGGGCGAGAGGGGGxxxxxxxxxxxxxxxC
433   GAGAGGGGGACAGGGtcatcagggtgcttagggtgggctccggggcgtctgcaccaccaggcgcacagccca
505   ggaggtggcaggagtcatccgttctgaaacagccagaggtacaacctcgtcgtccaggcaccggccgagttg
577   ggactcagggtcaaagccaagctgaggcaacgtcgagatggagggtaacagccctcagcctgcacctgccac
649   actgcggaggccccacaggaacaatccgggagggtggggtggtgcctgcctggcagctgcgggggctgggta
721   gggaagggctactccaccctggaggcccagctcacaccaacctcctagcccctgacgtcccaccaggcagct
793   tcacaaggttacaggtcggttccttctccactggattcctcccacatcgggtgacctgaccacacacacggc
865   aggtgcccagcggtggtcccagccccaacatctcaagagcaggacacccgagtggagatactaggtcacagg
937   aatgtctccacacagacattcaggcaggttcgagggaagaagacagcttcccggccactctcccaccacgcc
1009  acacccggtgggctcctctcctcaacctgggcccacattctctcccaggttactcacatcactcagtcatcc
1081  ctcacatcactcacatcgccccatgacatccgcctctgagctgccagcctccctccccagcccctttcttcc
1153  ttccctccctccctccctcctttctttttttttgaaacaggtcacccaggctggagttcagtggcacaatctt
1225  gactcactgcagcctccgcctctggggctcaagccttccaggctcaagcaatcctcagcctcagcctcccaa
1297  gtagttggaaacgtaactgggcaccaccatgcccagctattttttttttttttcagtagagatgaggtctca
1369  ctacattacccaggctggtctcagactcctggtctcaagcaatcttctcaccttggcctcccaaagtgctag
1441  gattacaggtgtgagccactgcgcccgacttccccagcccctttctgacccacagcctgggatc
```

Figure 3. Sequences flanking pMS43 and pMS228B. Alu dispersed repeats are in bold italics, and the minisatellite tandem arrays are abbreviated as in Figure 2. (A) pMS43 – sequencing 3′ of 43A resumes at a *Sma*I site within the last repeat unit, and so only 13 bases are shown in capitals. Underlining shows other short regions of tandem repeats (B) pMS228 – sequence from a 2kb subclone of DNA surrounding 228B; direct repeats flanking the Alu element are underlined.

unrelated individuals of Utah Mormon, French, Venezuelan and Amish origin from the CEPH panel (Figure 1B). The second locus is also highly variable (heterozygosity 85%), with *Alu*I alleles 0.6–5.9 kb long and is detected relatively weakly by pMS228. Family analysis (Figure 1A) showed that these loci were tightly linked (no recombinants in 25 offspring, $\theta < 0.11$ ($p > 0.95$)). Restriction mapping and DNA sequence analysis of pMS228 revealed the presence of two separate tandem-repetitive regions termed 228A and 228B (Figure 1C). Reprobing human DNA with subclones from pMS228 showed that 228A detected the larger, more intensely hybridising locus, and 228B the smaller, more faintly hybridising locus. These minisatellites were localised, using somatic cell hybrids and *in situ* hybridisation, to the terminal G-band 17p13-pter (data not shown). The properties of the loci detected by pMS51 and pMS228, and by the other minisatellites discussed in this report, are summarised in Table 1.

```
    1        10        20        30        40        50        60
    .      ata   tt    .     a   .         .  t      a         .
(a) nnnnnnntattatccccattttacagatgaggaaactgaggcacagagaggttaagtanctt..
           | ||  ||||||||||||||||||||||||||||| ||||||||||||||| |||
(b)       ttnttnnccccattttacagatgaggaaactgaggctcagagaggttaagtaactt..
                       |||  || |||||||| |||||| ||||||||||||||   ||
(c)                    tttnnagntgaggaaantgaggcncagagaggttaagtnnntt..


                70        80        90       100       110
               a.         .         .         .         .
(a)    ..gcccaaggtcncnnancn
         |||||||||| |  | |
(b)    ..gcccaaggtcacacagctagtaagtgggnagaggcaggatttgaacc
         |||||||||||||||||
(c)    ..gcccaaggtcacacagc
```

Figure 4. Consensus sequences of a recently described interspersed repeat element. (a) shows the consensus of Donehower *et al.* [12], and (b) and (c) the sequences arising from our analyses (see Results): (b) is the sequence resulting from a consensus criterion of ≥ 12 matches from the 27 human examples studied, while (c) has a consensus criterion of ≥ 18 matches from 27 sequences studied. Note that while the consensus sequences are nearly identical over the central 55 bases, (c) is less well defined at the 5′ end than (a), but is better defined at the 3′ end. (b) extends a further 28 bases in the 3′ direction, suggesting that this repeat element may be longer than the 70bp element proposed by Donehower *et al.* [12].

DNA sequences flanking minisatellites

DNA sequence data around pλg3 have already been reported [11]. The DNA sequences flanking the other minisatellites in Table 1 are shown in Figures 2 and 3. For clarity and brevity, only the outermost repeat unit on each side of each minisatellite tandem array is shown (in capitals), separated by a series of x's to indicate the omission of most repeats. The extent of each minisatellite array could be clearly defined, and in most cases the

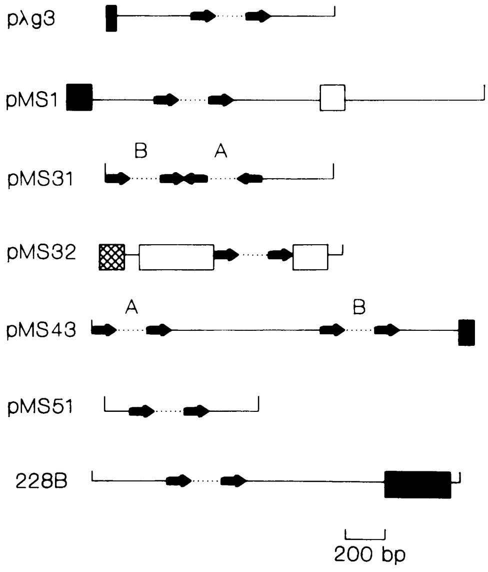

Figure 5. Dispersed repeat elements flanking human minisatellites. The tandem repeat arrays have been abbreviated to a pair of arrows separated by a dotted line. Boxes represent regions of dispersed repeat sequence – Alu elements as filled boxes, L1 element cross-hatched. The open boxes represent (in pMS1) a novel dispersed repeat and (in pMS32) retroviral LTR sequences.

boundary with the flanking sequence could be placed to within one or two bases. In only one of the repeat arrays (pMS51) does there appear to be an integral number of repeat units; hence in most cases the outermost repeat units shown will appear out of register.

Association with tandem repeat sequences

Sequencing confirmed the existence of two distinct minisatellites in pMS43 [8] and pMS228 [this work] previously defined by restriction mapping, and revealed the presence of additional shorter regions of tandem repeat sequence in pMS43 (Figure 3).

A previously unsuspected tandem repeat region was found by sequencing pMS31; this array is separated from the major minisatellite by 12 bases, and consists of 7 repeats of a 19 bp G/C rich sequence which extends from 9 bases into the clone (Figure 2b). This region, designated 31B, is G-rich on the opposite strand to the G-rich strand of the major minisatellite ('31A'). Genomic mapping using a 31B-containing subclone showed that this minisatellite is variable, but with only two *Alw*NI alleles of about 350 and 460 bp detected in DNA from 10 unrelated individuals. The population heterozygosity at this minisatellite is estimated at 50% (data not shown).

Association with dispersed repeat sequences

Sequences from seven minisatellite clones are reviewed here, and four (pλg3,pMS1,pMS43 and pMS228) contain at least part of an Alu dispersed repeat [22] (Figure 5). Moreover, the first 91 bases of pMS32 show a 64% similarity to bases 4572 to 4663 of the L1 element consensus of Singer (cited in [23]), suggesting that this is the start of a 5′ truncated member of the L1 family of dispersed repeats.

Interspersed repeat element in pMS1

Searches of the EMBL sequence database revealed the presence of an element in pMS1 which showed similarity to many human and mammalian sequences. Similar findings have recently been reported for a 70 bp element, initially found near a Hepatitis B virus integration site on chromosome 17 [12]. We constructed an intermediate consensus from the alignment of human sequences detected by pMS1 in the EMBL sequence database. This intermediate consensus was then used to re-scan the database; there were 80 matches judged to be significant by comparison with scans using randomised sequences of identical base composition. 67 of the matches were mammalian sequences, and of the 40 best matches (all mammalian), all but three were human. The best 27 human matches were used to construct the human consensus sequences shown in Figure 4, both of which are similar to the consensus of Donehower *et al.*[12]. Like Donehower *et al.*, we find that the central part of the element is least variable, and that the length of the consensus sequence obtained depends on the stringency of the consensus criteria. Limiting the consensus to bases present in at least 18 out of the 27 sequences studied produces a consensus sequence 60 bp long. A less stringent criterion of 12/27 matches produces a weaker but still significant consensus 103 bp in length.

Donehower *et al.* [12] propose that these elements have been conserved in mammalian evolution, on the basis of sequence similarity of elements found in homologous positions in the human and murine *N*-myc and myoglobin genes. A difficulty with this approach is that the pairs of human and murine genes studied were selected from the database by virtue of similarity to the consensus sequence. A bias might thereby be introduced towards pairs of elements which have been well preserved between the human and murine counterparts. We therefore chose to compare a randomised selection of twenty human elements and the surrounding DNA with the homologous sequences from other mammalian genes, if present in the database.

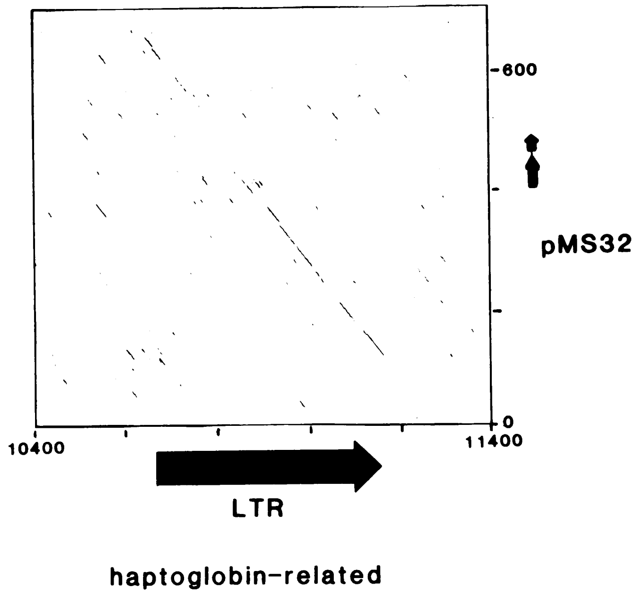

Figure 6. Dot matrix analysis showing sequence similarity between pMS32 and the RTVL-I LTR in the haptoglobin-related gene [24]. This analysis places a point where there are 14 or more matches within a window of 24 bases. The sequence similarity with pMS32 is interrupted by the minisatellite tandem repeats, and so the sequence used here for pMS32 has all but 1½ repeat units removed. The extent of the RTVL-I LTR is shown by the large arrow, while the smaller arrows indicate the position of the remaining repeat units in the truncated pMS32 sequence.

Six regions could be analysed in this way, in the human and rat cytochrome p450, fibrinogen γ-chain and enkephalin genes,and three in the human and mouse myoglobin genes. The element in intron 1 of the cytochrome p450 genes appeared to be better preserved (71.4%) than expected (χ^2 = 4.8, $p < 0.05$) for human and rodent non-coding DNA [24]. However, the element is not significantly better preserved than the surrounding 800 bp of intron DNA (62.5% similarity), and it may be that this region shows higher than expected sequence similarity because of its proximity to the boundary with exon 2. Elements in the 3′ flanking DNA of the human and rat enkephalin genes were not significantly better preserved 65.5%) than the surrounding non-coding DNA (61.8%). An element in intron 6 of the human fibrinogen γ-chain gene did not show sufficient sequence preservation to be definitively identified in the homologous rat sequence. Of three copies of the element studied from the human and mouse myoglobin genes (two in intron 1,one in intron 2) two were no better preserved (40% and 45.5%) than the surrounding intron DNA (56% and 45.5% respectively), and the third was not identifiable in the murine sequence.

A putative retroviral LTR in pMS32

Bases 125−666 of the sequence of pMS32 in figure 2c showed 70% similarity to the 3′ LTR of a retrovirus-like element (RTVL-I) described in the human haptoglobin-related gene [25] (Figure 6). The minisatellite array of pMS32 interrupts this similarity, which resumes at the end of the tandem repeat block; in Figure 6 the tandem repeat array of pMS32 has been abbreviated to 1½ repeat units. No evidence could be found for similarity extending beyond the bounds of the LTR. It appears, then, that the tandem repeat array of pMS32 may have originated by expansion from this point in a diverged member of the LTR family. While sequences from within the RTVL-I appear to be present in more than 100 copies per genome [25], the copy number of isolated RTVL-I LTRs has not, to our knowledge, been established.

DISCUSSION

The sequences flanking human minisatellites presented here provide further examples of the association between minisatellites and other tandem repeats [8]. Previous evidence for a pair of minisatellites in pMS43 was confirmed by sequence data, and two further examples (pMS31 and pMS228) of minisatellite clones containing two closely adjacent tandem repeat arrays were discovered. With the inclusion of the paired minisatellites defined by genomic mapping with pλg3, the four examples of paired minisatellites (pλg3, pMS31, pMS43 and pMS228) now documented are all located in proterminal regions (see Table 1). Genomic mapping has failed to show any evidence for clustering of minisatellites at interstitial locations. The minisatellite 228B (Figure 1,Table 1) combines a high heterozygosity (85%) with small allele sizes within a limited range (0.6−5.9 kb,see Figure 1B); this combination makes it a very useful locus for analysis of minisatellites by the polymerase chain reaction [15].

The flanking sequences also provide evidence that there may be a high frequency of dispersed repeat elements near minisatellites (Figure 5). In less than 6.5 kb of flanking sequence, we have found seven dispersed repeat elements (four Alu elements,one L1 (Kpn),a retroviral LTR and a recently described short interspersed repeat); there is no obvious difference in the frequency of dispersed repeat elements between proterminal and interstitial minisatellites. Although local variations do occur, one would expect an Alu element to occur every 5−6 kb [22] and an L1 every 150kb [26] in human DNA. Thus the frequency of Alu and L1 elements alone suggests an excess of dispersed repeats near minisatellites. Other examples of clustering of dispersed repeats are known, for example in the non-coding DNA of the human tissue plasminogen activator (t-PA) gene [27], in which 28 complete or partial Alu elements and a partial L1 element occur in 36.6 kb, together with a substantial block (570 bp) of a 7bp tandem repeat.

A dispersed repeat sequence located in DNA flanking the minisatellite in pMS1 appears to be a member of the set of repeats recently described by Donehower *et al.*[12]. The consensus sequences we have deduced by comparison of human examples of this element show some differences from the Donehower consensus (Figure 4), and we have detected examples in additional human genes, including the interleukin-2, cytochrome p450, apolipoprotein E, proopiomelanocortin and tissue plasminogen activator genes. We have failed to find evidence for evolutionary conservation of the homologous human/rodent sequences we have compared, but find that homologous human/rodent elements have a level of base substitution not significantly lower than that of adjacent non-coding DNA and consistent with the expected level of substitution between human and rodent non-coding DNA [24].

An example of this element is found between the poly-A tail and an Alu sequence in a processed X-linked GAPDH pseudogene [28], and the presence of such elements in a minisatellite clone and near a processed pseudogene would suggest that the apparent association with coding sequence [12] may be a result of the bias towards coding sequence in the databases. On the other hand, the origin and role of these elements are still uncertain, and the apparent DNA binding activity for these elements found in liver cell lines [12] might suggest that at least some of the elements may have some *cis*-acting function.

We have found putative retroviral LTR sequences either side of the minisatellite in pMS32. The alignment between the sequence of pMS32 and the RTVL-I LTR [25] extends to the boundary with the tandem repeat array, and resumes on the other side (Figure 6), suggesting that the tandem repeat block may have expanded from within a diverged member of the LTR family. A similar example in which a mouse minisatellite appears to have arisen from within a member of an interspersed repeat (MT) family has been found [R.Kelly, G.Bulfield, A.Collick, M.Gibbs and A.Jeffreys,manuscript submitted]. Moreover, a variable tandem repeat region appears to have arisen from within a member of the human *Mst*II dispersed repeat family [29], prompting speculation that the presence of the *Mst*II dispersed repeat may have been instrumental in the generation of the tandem repeat array. These sequence data provide further evidence for the clustering of human minisatellites, and show that they tend in turn to be associated with dispersed repeat elements. This suggests that the more variable minisatellites cluster in atypical regions of the genome, which are rich in both tandem and dispersed repeats, and which have a marked tendency to occur near the ends of human autosomes.

ACKNOWLEDGEMENTS

This work was supported by a grant to A.J.J. from the MRC. J.A.L.A. is an MRC Training Fellow. A.J.J. is a Lister Institute Research Fellow. We thank Robert Kelly for many helpful suggestions, Raymond Dalgleish for help with computing and Profs. H.Cann and J.Dausset for DNA from members of the CEPH families. The minisatellite probes are the subjects of patent applications, and commercial enquiries should be addressed to ICI Diagnostics, Gadbrook Park, Rudheath, Northwich, Cheshire CW9 7RA,UK.

REFERENCES

1. Jeffreys.A.J.,Wilson,V. and Thein,S.L.(1985) *Nature* **314**,67–73.
2. Wong,Z.,Wilson,V.,Patel,I.,Povey,S. and Jeffreys, A.J. (1987) *Ann.Hum.Genet.* **51**,269–288.
3. Nakamura,Y.,Leppert,M.,O'Connell,P.,Wolff,R.,Holm,T., Culver,M., Martin,C.,Fujimoto,E.,Hoff,M.,Kumlin,E. and White,R.(1987) *Science* **235**,1616–1622.
4. Gill,P.,Jeffreys,A.J. and Werrett,D.J. (1985) *Nature* **318**,577–579.
5. Jeffreys,A.J.,Brookfield,J.F.Y. and Semeonoff,R. (1985) *Nature* **317**,818–819.
6. DonisKeller,H.,Green,P.,Helms,C.,Cartinhour,S., Weiffenbach,B.,Stephens,K.,Bowden,D.W.,Smith,D.R., Lander,E.S.,Botstein,D.,Akots,G.,Rediker,K.S., Gravius,T.,Brown,V.A.,Rising,M.B.,Parker,C.,Powers, J.A.,Watt,D.E.,Kauffman,E.R.,Bricker,A.,Phipps,P., Muller-Kahle,H.,Fulton,T.R.,Ng,S.,Schumm,J.W., Braman,J.C.,Knowlton,R.G.,Barker,D.F.,Crooks,S.M., Lincoln,S.E.,Daly,M.J. and Abrahamson,J. (1987) *Cell* **51**,319–337.
7. Jeffreys,A.J.,Wilson,V.,Thein,S.L.,Weatherall,D.J. and Ponder,B.A.J. (1986) *Am.J.Hum.Genet.* **39**,11–24.
8. Royle,N.J.,Clarkson,R.E.,Wong,Z. and Jeffreys,A.J. (1988) *Genomics* **3**,352–360.
9. Nakamura,Y.,Lathrop,M.,O'Connell,P.,Leppert,M.,Lalouel,J-M. and White,R. (1988) *Genomics* **3**,67–71.
10. O'Connell,P.,Lathrop,G.M.,Leppert,M.,Nakamura,Y., Müller,U.,Lalouel,J-M. and White,R. (1988) *Genomics* **3**,367–372.
11. Wong,Z.,Wilson,V.,Jeffreys,A.J. and Thein,S.L. (1986) *Nucleic Acids Res.* **14**,4605–4616.
12. Donehower,L.A.,Slagle,B.L.,Wilde,M.,Darlington,G. and Butel,J.S. (1989) *Nucleic Acids Res.* **17**,699–710.
13. Vieira.J. and Messing,J. (1982) *Gene* **19**,259–268.
14. Jeffreys,A.J.,Wilson,V. and Thein,S.L. (1985) *Nature* **316**,76–79.
15. Jeffreys,A.J.,Wilson,V.,Neumann,R. and Keyte,J. (1988) *Nucleic Acids Res.* **16**,10953–10971.
16. Twigg,A.J.and Sherratt,D. (1980) *Nature* **283**,216–218.
17. Sanger,F.,Nicklen,S. and Coulson,A.R. (1977) *Proc.Nat.Acad.Sci.U.S.A.* **74**,5463–5467.
18. Tabor,S. and Richardson,C.C. (1987) *Proc.Nat.Acad.Sci.U.S.A.* **84**,4767–4771.
19. Peterson,M.G. (1988) *Nucleic Acids Res.* **16**,10915.
20. Lipman,D.J. and Pearson,W.R. (1985) *Science* **227**,1435-1441.

21. Devereux,J.,Haeberli,P. and Smithies,O. (1984) *Nucleic Acids Res.* **12**,387–395.
22. Schmid,C.W. and Jelinek,W.R. (1982) *Science* **216**,1065-1070.
23. Demers,G.W.,Brech,K. and Hardison,R.C. (1986) *Mol.Biol.Evol.* **3**,179–190.
24. Li,W-H.,Tanimura,M. and Sharp,P.M. (1987) *J.Mol.Evol.* **25**,330–342.
25. Maeda,N. (1985) *J.Biol.Chem.* **260**,6698–6709.
26. Singer,M.F. and Skowronski,J. (1985) *Trends Biochem.Sci.* **10**,119–122.
27. Friezner Degen,S.J.,Rajput,B. and Reich,E. (1986) *J.Biol.Chem.* **261**,6972–6985.
28. Hanauer,A. and Mandel,J.L. (1984) *EMBO J.* **3**,2627–2633.
29. Mermer,B., Colb,M. and Krontiris,T.G. (1987) *Proc.Nat.Acad.Sci.U.S.A.* **84**,3320–3324.

This article, submitted on disc, has been automatically converted into this typeset format by the publisher.