Characterisation of a novel minisatellite that provides multiple splice donor sites in an interferon-induced transcript

Maria-Grazia Turri, Karen A. Cuin and Andrew C. G. Porter*

Department of Biochemistry, University of Oxford, South Parks Road, Oxford OX1 3QU, UK

Received March 17, 1995; Revised and Accepted May 5, 1995

GenBank accession no. U22970

ABSTRACT

Nucleotide sequence features of the human interferon-inducible gene 6-16 are described and include, within a CpG island, a partially expressed minisatellite consisting of 26 tandemly repeated dodecanucleotides. The repeat unit consensus sequence (CAGGTAAGGGTG) is similar to the mammalian splice donor consensus sequence [(A/C)AGGT(A/G)AGT]. The splice donor site of exon 2, as determined previously, forms part of the most upstream of the repeat units. We show that the two neighbouring repeat units also provide functional splice donor sites effectively extending exon 2 by 12 or 24 nt and inserting four or eight amino acids respectively into the predicted gene product. A similar pattern of differently spliced transcripts is detected in several human cell types. Both the number of repeat units per allele and the nucleotide sequence itself show limited polymorphism within the human population. Similar minisatellites from non-human primates are described and also appear to modulate splicing of a 6-16 transcript. The 6-16 minisatellite is therefore an example of tandemly repeated DNA that has a role in gene expression and may provide a useful *in vivo* system for the analysis of 5′ splice site choice and minisatellite biology.

INTRODUCTION

Minisatellites are arrays of tandemly repeated sequences with repeat units in the range of 8–100 base pairs (bp), each locus being composed of one type of repeat unit (1). Minisatellites are distributed throughout the mammalian genome although, in humans, distribution of hypervariable loci is biased towards telomeres (2). Many, though not all, loci are highly polymorphic with respect to the number of repeat units per allele (3,4). The importance of hypervariable minisatellites as informative genetic markers for the mapping and typing of genomes is well known (5). Mechanisms that lead to minisatellite polymorphisms may include, to varying degrees, replication slippage and recombination in the form of unequal sister chromatid exchange and inter-allelic gene conversion (6). Mechanisms differ between germ line and somatic cells, the former having a bias towards generating alleles of increased size (7). 'Core' sequences within repeat units (3), similar to the chi recombination sequence of bacteriophage lambda, and as yet unidentified sequences outside minisatellites (6,7), have been proposed as promoters of variability.

Minisatellites are often regarded as examples of 'selfish' DNA that play no active role in gene or genome function. This would tend to free them from selective forces and may help to explain the fact that so many loci are polymorphic. Some minisatellites are associated with genes, however, where they may form part of protein coding (8,9) or transcriptional regulatory (10) sequences. We describe here a minisatellite that modulates the splicing pattern of its host gene, the human gene 6-16.

The 6-16 gene is one of many human genes transcribed in response to type I interferon (IFN), is inducible in all human cell types analysed to date (11,12) and is located at chromosome 1p35 (13). No homologous transcript is detectable in mouse cells but one from chimpanzees has been characterised (14). The human gene has the potential to encode a generally hydrophobic protein of 113 amino acids, including a putative N-terminal signal sequence, but its presumed role in the antiviral or antiproliferative effects of IFN is not known. Previously we used gene targeting to create human cell lines in which one or both copies of 6-16 gene have been inactivated (15,16). The latter cells were used to show that the 6-16 gene was not required for the IFN-induced antiviral state, a least with respect to the small set of viruses tested (16).

The 6-16 gene can be used as a natural target sequence for studying gene targeting and extrachromosomal homologous recombination in human cells (16). *Alu* sequences (17), microsatellites (18) and minisatellites (19–22) have been implicated in various types of homologous recombination. We were therefore interested in these or any other sequence features within the 6-16 gene whose influence on homologous recombination might be explored in their normal context and we describe *Alu* sequences, a CpG island and a minisatellite that might be useful in this regard. The 6-16 minisatellite is described in the most detail, however, because it has the novel and unexpected property of providing multiple splice donor sites for its host gene.

* To whom correspondence should be addressed at present address: MRC Clinical Sciences Centre, Royal Postgraduate Medical School, Hammersmith Hospital, Du Cane Road, London W12 ONN, UK

METHODS

General molecular biology

Standard methods for digestion of DNA with restriction enzymes, ligation and cloning were used (23). Southern and Northern blots were as previously described (16). The probe for the Southern analysis was previously described as probe a (15).

Cell lines

The following cell lines were used as a source of DNA and/or RNA: HT1080 (human fibrosarcoma; 24); HT1080+/– (an HT1080 derivative with one silenced 6-16 allele; clone 6-42; reference 15); HT1080–/– (an HT1080 derivative with both 6-16 alleles silenced; clone 1.1; reference 16); HeLa (human epitheloid carcinoma; 25); CaCo (human colon adenocarcinoma; 26); EJ138 (derivative of EJ, human bladder carcinoma; 27); RT112 (human bladder carcinoma; 28); Daudi (human lymphoma; 29); MRC-5 (human foetal lung; 30); GMO3452 (chimpanzee skin fibroblast; 31); Puti (orang-utan B-cell; 32); MLA 144 (gibbon lymphoma; 33); Vero (African green monkey kidney; 34). Growth and treatment with type I IFN (Wellferon or A/D hybrid) was as described (35) except that lymphoblastoid cells were grown in RPMI.

Construction and nucleotide sequencing of plasmids

All sequencing was by the chain termination method (36) using 'sequenase' T7 polymerase and other reagents supplied by USB. Templates were double-stranded and derived as follows. A plasmid carrying the human 6-16 gene subcloned (35) from cosmid 10-3 (12) in the vector pGEM2 (Promega) was linearised with *Kpn*I (intron 1) or *Cla*I (intron 4) and treated with exonuclease Bal31 for various times. The remaining 6-16 DNA upstream of the linearisation site was removed by digestion with *Sma*I (polylinker) and recircularisation. The resulting series of deletion plasmids was sequenced with an SP6 primer. Additional sequence was obtained from subcloned *Bam*HI–*Kpn*I (intron 1) and *Cla*I–*Pvu*II (intron 4) fragments. Sequencing through the minisatellite in both strands was achieved with primers oE2+ (5′-TACCTGCTGCTCTTCACTTG-3′) and oI2– (5′-AAGCACGCCAGACCCTCTAC-3′).

Genomic and RT-PCR products were sequenced with primers T7 and T3 after they were cloned by the method of Marchuk *et al.* (37) into the *Eco*RV site of pBluescript IIKS+ (Stratagene). PCR-generated sequence errors are a possibility when sequencing cloned PCR products; we detected no such errors when the sequences of two independent clones of the same genomic PCR product (Daudi large allele) were compared.

Preparation of nucleic acids

Cytoplasmic RNA was prepared as described previously (35). Genomic DNA from the peripheral blood lymphocytes of a variety of human volunteers, or from cell lines derived from such individuals, was prepared in agarose 'plugs' and kindly provided by Neal Mathias and Chris Tyler-Smith (38). Additional genomic DNA was prepared from some of the cell lines described above by the method of Laird *et al.* (39). The HT1080 6-16 allele was isolated (P. M. England and A.C.G.P., unpublished) from an HT1080 genomic library kindly provided by Elizabeth Fisher, of the Biochemistry Department at St Mary's Hospital, London.

Polymerase chain reactions on genomic DNA (Genomic PCR)

Each reaction (50 μl) contained genomic or 'plug' DNA (0.1–1 μg), primers oE2+ and oI2– (1 μM each), 1 mM $MgCl_2$, nucleotide triphosphates (200 μM each), 1 × reaction buffer (10 mM Tris–HCl, pH 9.0; 50 mM KCl; 0.1 % Triton X-100) and 2.5 U *Taq* Polymerase (Promega). Reactions were heated to 94°C for 2 min then at 94°C (30 s), 60°C (1 min) and 72°C (30 s) for 30 cycles. Aliquots (20 μl) were analysed electrophoretically.

Reverse transcription-polymerase chain reactions (RT-PCR)

Samples (1 μg) of cytoplasmic RNA were reverse transcribed with AMV reverse transcriptase in 20 μl reactions for 30 min at 42°C with the enzyme, buffer, salt, nucleotide triphosphates and primer concentrations specified by the supplier (Promega). PCR reactions (50 μl) contained portions (2.5 μl) of reverse transcription reactions, primers oE2+ (1 μM) and oE3– (5′-ATGAAGGTCAGGGCCTTCCA-3′; 1 μM), 2.5 mM $MgCl_2$, nucleotide triphosphates (200 μM each), 1 × reaction buffer (see above) and 2.5 U *Taq* Polymerase (Promega). Reactions were heated to 94°C for 2 min then for 30 cycles of 94, 65 and 72°C, 30 s at each temperature. Aliquots (20 μl) were analysed by PAGE. For some analyses T4 polynucleotide kinase (USB) was used to end-label primer oE2+ with ^{33}P. Kinase reactions (20 μl) contained 300 ng primer, 30 U kinase and 30 μCi [γ-^{33}P]dATP (Amersham, 1000 Ci/mmol) in kinase buffer (USB) and were incubated at 37°C for 30 min then at 65°C for 5 min. PCR conditions were as above except that small (10 μl) reaction volumes were used and each reaction included a portion (1 μl) of the kinase reaction. Aliquots (3 μl) of the labelled PCR products were analysed by denaturing PAGE.

Polyacrylamide gel electrophoresis (PAGE)

Polyacrylamide gels were prepared and used in 1 × Tris/Borate/EDTA buffer as described (23). The acrylamide:bis-acrylamide ratio was 29:1 by weight. Gels were 3.5% (w/v) acrylamide for analysis of genomic PCR products and, unless stated otherwise, 5% (w/v) acrylamide for the analysis of RT-PCR products. Ladder markers (Gibco/BRL) with concatemers of 10 or 100 bp units were used at 500 ng/lane. A voltage of 120 V was applied across the gels (20 × 20 × 0.1 cm) for 3 h. Gels were then soaked in ethidium bromide (~200 μg/ml) for 20 min and photographed (polaroid 665 film) on a UV transilluminator. Polaroid negatives were reproduced for figures. Standard sequencing gels and autoradiography (22) were used for denaturing PAGE analysis of labelled PCR products.

Heteroduplex analysis

Cloned PCR products were excised from their vector by digestion with *Bam*HI and *Hin*dIII under standard conditions. Aliquots of digestion reactions (~100 ng) were mixed and analysed by PAGE with or without prior heating at 94°C for 2 min followed by cooling, over a period of 2 min, to 20°C.

RESULTS

Sequence features of the human 6-16 gene

The sequence of the promoter (35) and exon (12) regions of 6-16 were reported previously and a half *Alu* sequence in the untranslated portion of exon 5 was noted (12). We sequenced the rest of the gene to generate, in combination with the established sequence data, 6083 nt of uninterrupted sequence from a *Bgl*II site

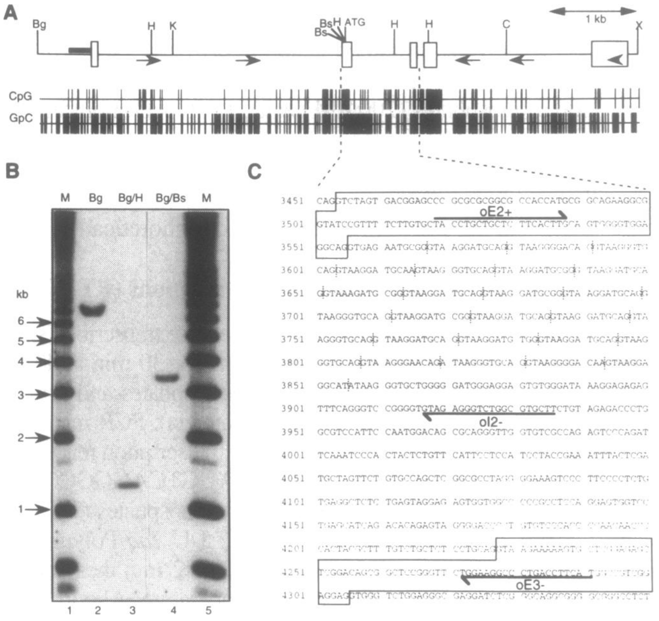

Figure 1. Sequence features of the 6-16 gene. (**A**) Scale map showing positions of exons 1–5 (boxes, left to right), *Alu* repeat sequences (arrows), initiation codon (ATG), probe used for Southern analysis (black bar), CpG and GpC dinucleotides and sites for restriction enzymes *Bgl*II (Bg), *Hae*II (H), *Kpn*I (K), *Bss*HII (Bs), *Cla*I (C) and *Xho*I (X). (**B**) Southern analysis of *Bgl*II (lane 2), *Bgl*II/*Hae*II (lane 3) and *Bgl*II/*Bss*HII (lane 4) digested human placental DNA probed with probe shown in (A) The sizes of radioactive markers (M) are indicated. (**C**) Nucleotide sequence spanning exons 2 and 3 (boxed). Sequences identical (oE2+) or complementary (oI2– and oE3–) to oligonucleotides used for PCR and sequence analyses are shown by half arrows. Dotted lines indicate the positions of potential splice donor sites within the minisatellite.

605 nt upstream of exon 1 to an *Xho*I site 96 nt downstream of exon 5. All clones used to generate this sequence were derived from the cosmid 10-3 (13) which was isolated from a genomic library of human placental DNA. Some sequence features are summarised in Figure 1.

The positions of the dinucleotides CpG and GpC with respect to the previously determined (12) intron/exon structure are shown in Figure 1A and clearly indicate the presence of a CpG island of ~1.5 kb positioned, unusually, in the centre of the gene, spanning exons 2, 3 and 4. That this island is unmethylated was suggested by Southern analysis of *Bgl*II digested human placental DNA (Fig. 1B). The 3.5 kb fragment detected after further digestion with *Bss*HII indicates that at least one of the overlapping *Bss*HII sites in exon 2 is unmethylated. Similarly, a *Hae*II partial digestion product of 3.5 kb was detected indicating that the *Hae*II site close to the *Bss*HII site is also unmethylated. The 5′ end of the gene may also be unmethylated to some extent since the major *Bgl*II–*Hae*II product of 1.3 kb indicates that the *Hae*II site in intron 1 is largely unmethylated.

The 6-16 gene includes five *Alu* repeat sequences the positions and orientations of which are shown in Figure 1A. We found no appreciable stretches of dinucleotide repeats within the gene. The most significant feature, for the purposes of this report, is an array of 26 tandemly repeated dodecanucleotide sequences of consensus sequence CAGGTAAGGATG. The last three nucleotides of exon 2 form the first three nucleotides of the first repeat unit (Fig. 1C). The array falls into the category of tandem repeats known as minisatellites as defined by Tautz (1).

Low variability in minisatellite repeat numbers

We wanted to know if the 6-16 minisatellite is hypervariable. We therefore used the polymerase chain reaction (PCR), with primers oE2+ and oI2– (Fig. 1C), to amplify the 6-16 minisatellite in genomic DNA samples originating from 44 unrelated individuals (Materials and Methods). PCR products were sized by PAGE (Fig. 2), and in certain cases (see below) were cloned and sequenced, so that the number of repeat units in each product could be accurately assessed. The results, summarised in Table 1, showed that the 6-16 minisatellite is not hypervariable. Human genotypes could be divided into four classes. The great majority (37/44) of samples, including the cell line HT1080, were of class I where only a single allele size, estimated to contain 26 repeat units, was detected (e.g. Fig. 2, lane 2). Five samples, including the cell lines Daudi and HeLa, were of class II, giving rise to two products estimated to contain 26 and 27 repeat units (e.g. Fig. 2, lane 3). The remaining classes, each represented by only one sample, appeared to give rise to products with 27 and 29 repeat units (class III; Fig. 2, lane 4) or 26 and 29 repeat units (class IV; Fig. 2, lane 5). The predominant (26/28) Caucasian genotype was homozygous (class I). Of the seven samples with one of the rare, heterozygous genotypes (classes II–IV), at least four were of non-Caucasian origin suggesting that in some ethnic groups heterozygosity may be more common.

Table 1.

Origin of human	Individuals	Genotype			
DNA	tested	I	II	III	IV
Caucasian	28	26	1	1	0
Black African	2	1	1	0	0
Black American	1	0	1	0	0
Amerindian	2	1	1	0	0
Cambodian	2	2	0	0	0
Melanesian	2	2	0	0	0
Mbuti Pygmy	1	1	0	0	0
Biaka Pygmy	2	1[b]	0	0	1
!Kung	1	1	0	0	0
Unknown[a]	3	2	1	0	0
Total	**44**	**37**	**5**	**1**	**1**

[a] Cell lines HT1080, MRC-5 and EJ138.
[b] This classification uncertain.

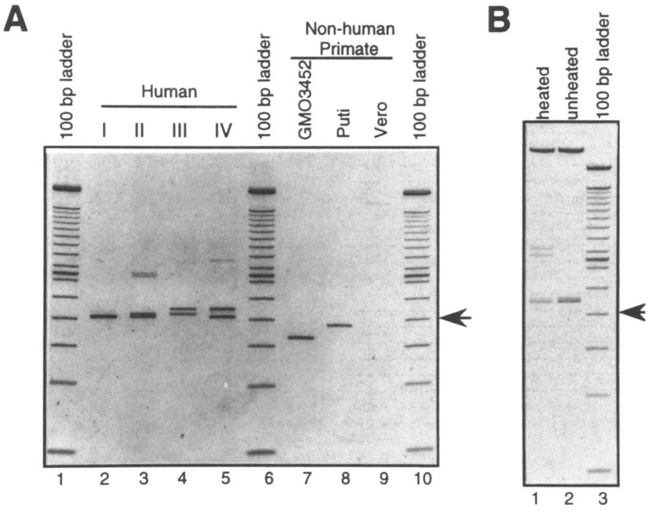

Figure 2. PAGE analysis of genomic PCR products; arrows indicate the position of the 400 bp markers. **(A)** PCR products of genomic DNA from Caucasian (lanes 2 and 4), African (Daudi; lane 3) or Biaka Pygmy (lane 5) cells represent the four classes (I–IV) of products obtained from a total of 44 human DNA samples (see text and Table 1 for details). PCR products of genomic DNA from chimpanzee (GMO34520), orang-utan (Puti) and African green monkey (Vero) cell lines are shown in lanes 7, 8 and 9 respectively. **(B)** Heteroduplex analysis of a mixture of the two cloned PCR products from Daudi cells (class II genotype). Identical samples were analysed with (lane 1) or without (lane 2) denaturation and renaturation.

The low variability at the human 6-16 minisatellite could reflect a conserved function for the 6-16 minisatellite (see below). We were therefore interested to know if variation at the 6-16 minisatellite has been constrained during recent evolution and accordingly amplified the 6-16 minisatellite in three non-human primates. Chimpanzee and orang-utan DNA gave rise to single products with 19 and 23 repeat units respectively (Fig. 2, lanes 7 and 8) suggesting that heterozygosity might be low in these species too. African green monkey DNA yielded two (faint) PCR products with 20 and 23 repeat units (Fig. 2, lane 9). No PCR products were obtained from samples of gibbon or mouse DNA (not shown).

When two allele sizes were detected in a sample, additional DNA species were often observed migrating slowly in the gel (e.g. Fig. 2, lanes 3 and 5). These are likely to be heteroduplexes as was demonstrated by analysis of the cloned PCR products of a class II sample. Thus, when the two products were excised (with some flanking polylinker DNA) from their vectors (Methods), mixed and separated electrophoretically with or without a prior denaturing and reannealing step, only the heated/cooled sample gave rise to extra, slowly migrating products similar to those observed in the original amplification (Fig. 2B).

Minisatellite sequence variations

We sequenced a selection of the genomic PCR products shown in Figure 1 as well as the minisatellite of a 6-16 clone isolated from a bacteriophage library of the human cell line HT1080 (Fig. 3A). All the sequenced 6-16 minisatellites conformed to the consensus sequence CAGGTAAGGATG although an improved consensus sequence for the African green monkey allele would replace the A at position 10 with G. We identified a total of 25 repeat unit variants, arbitrarily naming each with a letter of the alphabet, so that the repeat unit organisation of the various alleles could be compared (Fig. 3B).

The four sequenced human alleles were remarkably similar. The sequence of the HT1080 allele was identical to the allele from cosmid 10-3 (Fig. 1C) . The Daudi small allele was identical to the HT1080 allele except for single base substitutions in repeat units 19 and 21. The Daudi large allele differed from the HT1080 allele only by the addition of an extra repeat. Alleles from non-human primates can be placed in an order of decreasing similarity to the HT1080 allele: chimpanzee, orang-utan, African green monkey. This order agrees with established primate evolutionary trees (40). Repeats at the 3′ end of the array, and perhaps to some extent at the 5′ end, have been conserved between species. Similar observations have been made during comparisons between primates of the hypervariable minisatellite MS1 (41).

Multiple minisatellite splice donors

The minisatellite consensus sequence contains a perfect match with all but the last residue of the mammalian splice donor consensus sequence, (A/C)AGGT(A/G)AGT (42). The previously determined sequence of human (12) and chimpanzee (14) 6-16 cDNA, indicates that the splice donor site in repeat unit 1 is active. We were interested to know if the splice donor sites of other repeat units are ever used. We therefore carried out PCR on reversed transcribed cytoplasmic RNA (RT-PCR) using primers oE2+ and oE3– (Fig. 1C). Products were analysed by PAGE (Fig. 4). As a control, PCR was carried out on the previously cloned and sequenced human 6-16 cDNA which gave rise to a single product consistent with the expected size (101 bp; Fig. 4A, B and C, lane 2). In contrast, RT-PCR of RNA from HeLa and HT1080 cells treated with IFN generated not only the 101 bp product but also several products of slower mobility (Fig. 4A, lanes 4, 7 and 9). This apparent complexity of products did not depend on the expression of both 6-16 alleles since it was not reduced in an HT1080 clone expressing only a single 6-16 allele (Fig. 4, lanes 7 and 9). Most of the products depended on IFN treatment for maximum expression (Fig. 4, lanes 3, 6 and 8) and were undetectable in an HT1080 clone in which both 6-16 genes have been inactivated (lanes 10 and 11) indicating that they were genuinely derived from the 6-16 transcript. Sometimes large products (apparently >400 bp) were detected but they did not satisfy these criteria and were ignored.

We analysed a variety of human and non-human primate cell lines in a similar way; the results (IFN-treated cells only) are shown in Figure 4B. The patterns of products obtained from the RNA of different human cell types and from the chimpanzee cells were very similar to each other (Fig. 4B, lanes 3–8). The patterns of products obtained from the orang-utan and gibbon RNA were also similar to each other but slightly different to the patterns from human cells (Fig. 4B, lanes 9 and 10). African green monkey cells were different again with few indications of products >101 bp. We believe that most of these differences reflect sequence variations between heteroduplexes rather than different splicing patterns (see below). Although the relative intensities of products depended on magnesium ion concentration in the PCR reaction (not shown), under the conditions employed in Figure 4 the most abundant product from human and chimpanzee RNA was the second fastest migrating product, whilst that for the remaining non-human primates was the fastest migrating product.

A

	Human (HT1080)	Human (Daudi, small allele)	Human (Daudi, large allele)	Chimpanzee (GMO3452)	Orang-utan (Puti)	African green monkey (Vero,small allele)
	CAGGTAAGGATG	CAGGTAAGGATG	CAGGTAAGGATG	CAGGTAAGGATG	CAGGTAAGGATG	CAGGTAAGGATG
1	-----G--A---	-----G--A---	-----G--A---	-----G--A---	-----G--A---	-----G--A---
2	-G----------	-G----------	-G----------	-G----------	-G----------	TG-------G--
3	---------GGA	---------GGA	---------GGA	---------GGA	---------GG-	-----G---G--
4	---------G--	---------G--	---------G--	---------G--	T--------G--	-C----------
5	------------	------------	------------	------------	------------	-----G---G--
6	--A------G--	--A------G--	--A------G--	--A------G--	-GA----C-G--	-C-------G--
7	------------	------------	------------	-------A----	-------A----	-----G---G--
8	-G----------	-G----------	-G----------	-G----------	-G----------	-C-------G--
9	-------A----	-------A----	-------A----	------------	-C----------	-----G---G--
10	-G----------	-G----------	-G----------	-G----------	-G----------	-C-------G--
11	------------	------------	------------	------------	-------A----	-C-------G--
12	-G----------	-G----------	-G----------	TG----------	TG----------	------------
13	---------G--	---------G--	---------G--	------------	-G-A--------	-C----------
14	------------	------------	------------	---------G--	------------	-G-------T--
15	-G----------	-G----------	-G----------	---------GAA	-G----------	---------G--
16	------------	------------	------------	---A-----G--	-------A----	---------GAA
17	---------G--	---------G--	---------G--	---------GGA	TG----------	---A-----G--
18	------------	------------	------------	--A-------G-	---------G--	---------GGA
19	------------	---------G--	TG----------	--TA-----G--	TG-------GAA	--A-------G-
20	TG----------	TG----------	------------		---A-----G--	--AA-----G--
21	---------G--	--C------G--	TG----------		---------GGA	
22	---------GAA	---------GAA	---------G--		--A------AGG	
23	---A-----G--	---A-----G--	---------GAA		--TA-----G--	
24	---------GGA	---------GGA	---A-----G--			
25	--A-------G-	--A-------G-	---------GGA			
26	--TA-----G--	--TA-----G--	--A-------G-			
27			--TA-----G--			

B

```
      HT1080  abcdefebgbebdebedeehdijckl
              IIIIIIIIIIIIIIIIII I IIIII
Daudi, small  abcdefebgbebdebededhmijckl

      HT1080  abcdefebgbebdebede.ehdijckl
              IIIIIIIIIIIIIIIIII IIIIIIII
Daudi, large  abcdefebgbebdebedehehdijckl

      HT1080  abcdefebgbebdebedeeh.dijckl
              IIIIII IIII      II IIIIII
  Chimpanzee  abcdef..gbeb......ehedijckl

              HT1080  abcdefebgbebdebedeehdijckl
                      II    I I  II    II II I
          Orang-utan  abnoepgbqbghreb...ghdsjctl

              HT1080  abcdefebgbebdebedeehdijckl
                      I          I      IIIII
African green monkey  a..uvqvwvwvwweqx....dijcky
```

C

```
a TTC ACT TGC AGT GGG GTG GAG GCA GGT AAG AAA AAG TGC TCG GAG AGC TCG
  phe thr cys ser gly val glu ala gly lys lys lys cys ser glu ser ser

b TTC ACT TGC AGT GGG GTG GAG GCA GGT GAG AAT GCG GGT AAG AAA AAG TGC TCG GAG AGC TCG
  phe thr cys ser gly val glu ala gly glu asn ala gly lys lys lys cys ser glu ser ser

c TTC ACT TGC AGT GGG GTG GAG GCA GGT GAG AAT GCG GGT AAG GAT GCA GGT AAG AAA AAG TGC TCG GAG AGC TCG
  phe thr cys ser gly val glu ala gly glu asn ala gly lys asp ala gly lys lys lys cys ser glu ser ser
```

D

```
λ33.6   TGGAGG.AAGGGC          MS1     TGGATAGGG...
        II III IIIII                    II  IIII
6-16    TGCAGGTAAGGG.          6-16    AGGTAAGGGTGC


core    GGGCAGGAXG...          core    GGGCAGGAXG
        III  I I I                     III  III
6-16    GGG.TGCAGGTAA          MS1     GGG.TGGATA
```

Figure 3. Analysis of nucleotide sequences of 6-16 minisatellites. (**A**) Sequences of human and non-human primate 6-16 minisatellites obtained from cloned PCR products. Repeat unit numbers are shown to the left. Nucleotides that do or do not differ to consensus sequence (top) are indicated respectively by the correct nucleotide or by a dash. (**B**) Alignment with the HT1080 allele of various 6-16 minisatellite alleles, shown as sequences of repeat unit variants arbitrarily represented by letters. Identities (I) and gaps in the alignment (.) are indicated. (**C**) Nucleotide and amino-acid sequence through splice junctions in cloned (a) 101 bp, (b) 113 bp and (c) 125 bp RT-PCR products derived from HT1080 RNA. The splice junction in the smallest product (dotted line) and the novel residues resulting from alternative splicing in the two larger products (boxes) are indicated. (**D**) Sequence alignments [as in (B)] between isomers of the 6-16 minisatellite repeat unit, the 12 bp subrepeat consensus of the 133.6 minisatellite (3), the 9 bp repeat consensus of the MS1 minisatellite (2) and the chi-like minisatellite 'core' consensus (3).

The three fastest migrating products appeared to differ in size by an amount similar to that expected (12 bp) if they were derived from transcripts generated by splice donor sites in repeat units 1, 2 and 3. This was confirmed by nucleotide sequence analysis of cloned RT-PCR products derived from HT1080 (101, 113 and 125 bp products), orang-utan (101 and 113 bp products) and African green monkey (101 bp product) cells. The nucleotide sequence, and predicted amino acid sequence, for the spliced region of the three HT1080 products is shown in Figure 3C.

The mobility of all but the three fastest migrating products was anomalous, differing with respect to the mobility of molecular weight markers when analysed on different percentage gels (Fig. 4A and C) and suggesting that they are heteroduplexes. This interpretation was reinforced by the experiment shown in Figure 4D where three cloned products of 101, 113 and 125 bp were excised (with some flanking polylinker DNA) from their vectors , mixed and analysed by PAGE with or without a prior denaturing and reannealing step. Only the heated/cooled sample gave rise to extra, slowly migrating products similar to those observed in the original RT-PCR analyses. As further confirmation of our interpretation of the RT-PCR results shown in Figure 4, RT-PCR analyses of the human RNA samples were repeated with one of the primers ^{33}P-labelled. As expected, when the products were analysed on a denaturing gel, three major products of 101, 113 and 125 nt were detected (data not shown).

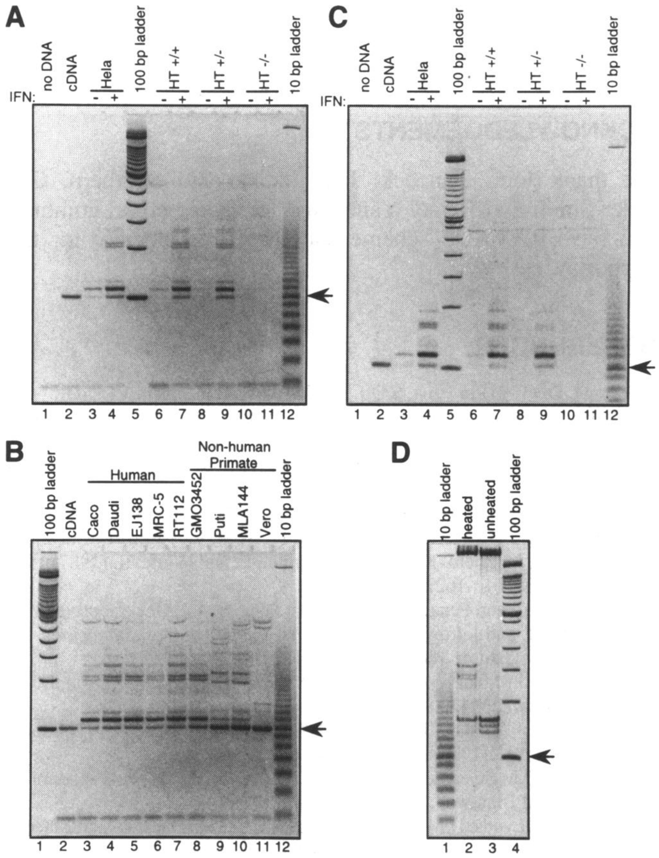

Figure 4. PAGE analysis of RT-PCR products. Arrows indicate the position of the 100 bp markers. (**A**) 5% gel of RT-PCR products from mRNA of HeLa (lanes 3 and 4), HT1080 (lanes 6 and 7), HT1080+/– (lanes 9 and 10) and HT1080–/– (lanes 10 and 11) treated with or without IFN as indicated. Controls (lanes 1 and 2) were PCR reactions with either no DNA or a plasmid (3.7 kb, 10 pg) carrying the previously characterised (12) 6-16 cDNA. (**B**) RT-PCR product of mRNA from a variety of human and non-human primate cells that had been treated with IFN. The cDNA control was as in (A). (**C**) Samples as in A but separated on a 3.5% gel. (**D**) Heteroduplex analysis of a mixture of cloned 101 bp, 113 bp and 125 bp RT-PCR products. Identical samples were analysed with (lane 2) or without (lane 3) denaturation and renaturation.

DISCUSSION

The 6-16 minisatellite as a splice-donor

We have identified and characterised a novel human minisatellite whose repeat units each carry a potential splice donor site. We have used RT-PCR and nucleotide sequencing of the products to show that at least the three most upstream splice sites are used during splicing of the 6-16 transcript. Much of the remaining complexity of the RT-PCR products appears to be the result of heteroduplex formation between the three major RT-PCR products. However, it is possible that additional splice donor sites within the minisatellite are used but give rise to mRNAs that are inefficiently amplified during RT-PCR and therefore not detected. For similar reasons, it is unlikely that the relative intensities of the products accurately reflect the relative abundance of the splicing products from which they derive.

Whatever the exact number and relative abundance of different 6-16 mRNAs, it is clear that the 6-16 minisatellite extends the coding capability of the 6-16 gene. Thus the use of internal splice donor sites introduces a mixture of charged, neutral and polar extra amino acids into the predicted product of the 6-16 gene (Fig. 3) at the junction between the putative signal sequence and the rest of the protein. It is difficult to assess the significance of such changes without knowing the function of the 6-16 gene product. It is possible that this function is aided by, or depends on, the degree of structural diversity generated by the extra residues. Alternatively, some of the 6-16 proteins may be non-functional, but their ability to co-exist with a functional product of the same allele may facilitate the evolution of 6-16 proteins with new or improved function.

The fact that some minisatellites can modify the coding or control sequences of their host genes is relevant to the origin and evolution of tandemly repeated sequences. In theory, there will be a selective force for the maintenance and expansion in the form of tandem arrays of any sequence whose situation in the genome is such that its tandem repetition would be advantageous to gene or genome function. The 6-16 minisatellite and other minisatellites (8–10) provide examples of situations in which tandem repeats are potentially advantageous to a host or nearby gene and may have arisen because of this. It is possible that many other tandemly repeated sequences have arisen from situations of this kind.

The apparent preference of the splicing machinery for the three most upstream of many splice sites indicates that splice site choice is sensitive to sequence context or other positional information. *In vitro* studies of splice donor site competition often detect a preference for the donor site closest to the acceptor site, but this preference can be reversed by deletions and substitutions in the 5′ exon (43,44). Sequences in 6-16 exon 2 may therefore promote the use of sites in repeat units 1–3 in preference to downstream sites. The 6-16 system may be useful for *in vivo* studies of splice donor site choice especially if splicing patterns can be shown to vary in some way. It is possible, for instance, that splice site choice varies with cell type, although there was no evidence for this amongst the small variety of cells we tested.

It would be interesting to know if other genes have their splicing patterns modified by a minisatellite. Minisatellites that are positioned at exon/intron boundaries and carry potential splice donor sites have been discovered in the genes for horse zeta globin (45) and human factor VII (46) but whether these sites are in fact used was not determined. If they are, only a fraction of the predicted splice products are expected to conserve the reading frame of the host gene product.

Variability at the 6-16 minisatellite locus

Simulations (47,48) of tandem repeat sequence variability predict, and observations (3) confirm, a correlation between the amount of sequence homogeneity among repeat units and the degree of variability in repeat number. With low but detectable allele length heterozygosity and only modest repeat unit homogeneity, the 6-16 minisatellite is consistent with this picture. Despite the low variability between human 6-16 alleles, we were able to analyse sufficient variability among primate alleles to detect a tendency for end repeat units, at the 5′ end in particular, to be conserved. Conservation of end repeats was also observed in computer simulated tandem repeat variability (48).

In a previous study of minisatellite variability between primates (41), conservation of end repeat units was observed for the hypervariable minisatellite MS1. This suggests that related mechanisms are responsible for generating variability in the 6-16 and MS1 minisatellites. Although the MS1 locus is similar to the 6-16 locus in its subtelomeric location (chromosome 1p33-35) and in having a short (9 bp) repeat unit, it differs in having a high allele-length heterozygosity (>99%) and a high number of repeat units (~140–2500). It is not clear why these loci have such contrasting stabilities. Both MS1 and the 6-16 minisatellite have sequences showing similarly weak homology to minisatellite chi-like core sequence (Fig. 3) so, unless a critical number of such sequences is required to promote variability, it seems unlikely that differences between these sequences are responsible. Polarities of instability have been observed across some minisatellite arrays suggesting that sequences promoting variability may be located outside repeat arrays (6,7); it may be that the 6-16 and MS1 loci differ in their content of such sequences. A further possibility, especially given the 6-16 minisatellite appears to be partially expressed, is that there may be a selective advantage in preserving the observed 6-16 alleles.

Similarities to the 6-16 minisatellite

In preliminary low stringency Southern analyses, a synthetic probe consisting of ~20 tandemly repeated copies of the degenerate sequence CAGGTAAGGRTG detected a single major fragment of the size expected for the 6-16 gene, and fainter products that did not appear to be polymorphic (M.-G.T. and A.C.G.P. unpublished results). Despite these findings we have noted a similarity between the 6-16 minisatellite and a previously characterised human minisatellite. Thus λ33.6 (3) has repeats of 37 bp, each consisting of three diverged 12 bp units whose consensus sequence (TGGAGGAAGGGC) can be aligned with the 6-16 repeat consensus sequence (Fig. 3). It may be, therefore, that the 6-16 minisatellite is sufficiently similar to some minisatellites to allow it to be detected, with the appropriate multilocus probe, as a relatively non-polymorphic fragment in a DNA fingerprint analysis. Searches of DNA and protein databases revealed a similarity between the 6-16 minisatellite and a 12 bp repeat unit (consensus: CTGGTAATGGTG) present in the coding region of the merozoite surface antigen 2 (MSA-2) gene of *Plasmodium falciparum* (49). Although database searches involving repeat sequences often reveal spurious similarities, the identification of the MSA-2 gene is notable given that various cytokines, including interferon alpha (50), appear to modulate a cellular anti-parasitic response.

6-16 gene sequence features

The most novel sequence feature of the 6-16 gene is clearly the unusual splice donating minisatellite discussed above. Our original purpose in sequencing the 6-16 gene, however, was to identify features whose effect on homologous recombination involving 6-16 might be tested. The 6-16 minisatellite and *Alu* sequences might be interesting in this context since minisatellites (19–22) and *Alu* sequences (17, and references therein) appear to promote homologous recombination in a variety of situations. The effect of the CpG island on homologous recombination might also be interesting to explore although we know of no evidence directly linking CpG islands with recombination. The 6-16 island is of interest in its own right, however, because of its unusual position. Most CpG islands are located at the 5′ end of genes (51). The positioning of the 6-16 island in the centre of the 6-16 gene may indicate that exon 1, which is non-coding, has become part of the 6-16 gene relatively recently in evolution.

ACKNOWLEDGEMENTS

We thank John Hancock, Jane Itzhaki, Ed Southern, Chris Tyler-Smith and Rafael Yáñez and for their helpful comments and the ERASMUS scheme and the Royal Society for their support.

REFERENCES

1 Tautz, D. (1993) In Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (ed.), *DNA Fingerprinting: State of the Science*. Birkhauser Verlag, Basel. pp. 21–28.
2 Royle, N.J., Clarkson, R.E., Wong, Z. and Jeffreys, A.J. (1988) *Genomics*, **3**, 352–360.
3 Jeffreys, A.J., Wilson, V. and Thein, S.L. (1985) *Nature*, **314**, 67–73.
4 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*, **23**, 1616-1622.
5 Jeffreys, A.J. and Pena, S.D.J. (1993) In Pena, S.D.J., Chakraborty, R., Epplen, J.T. and Jeffreys, A.J. (ed.), *DNA Fingerprinting: State of the Science*. Birkhauser Verlag, Basel. pp.1–20.
6 Jeffreys, A.J., MacLeod, A., Tamaki, K., Neil, D.L. and Monckton, D.G. (1991) *Nature*, **354**, 204–209.
7 Jeffreys, A.J., Tamaki, K., MacLeod, A., Monckton, D.G., Neil, D.L. and Armour, J.A.L. (1994) *Nature Genet.*, **6**, 136–145.
8 Swallow, D.M., Gendler, S., Griffiths, B., Corney, G., Taylor-Papadimitriou, J. and Bramwell, M.E. (1987) *Nature*, **328**, 82–84.
9 Kim, H.S., Lyons, H.M., Saitoh, E., Azen, E.A., Smithies, O. and Maeda, N. (1993) *Mammalian Genome*, **4**, 3–14.
10 Trepicchio, W.L. and Krontiris, T.G. (1992) *Nucleic Acids Res.*, **20**, 2427–2434.
11 Kelly, J.M., Gilbert, C.S., Stark, G.R. and Kerr, I.M. (1985) *Eur. J. Biochem.*, **153**, 367–371.
12 Kelly, J.M., Porter, A.C.G., Chernajovsky, Y., Gilbert, C.S., Stark, G.R. and Kerr, I.M. (1986) *EMBO J.*, **5**, 1601–1606.
13 Itzhaki, J.E., Barnett, M.A., MacCarthy, A.B., Buckle, V.J., Brown, W.R.A. and Porter, A.C.G. (1992) *Nature Genet.*, **2**, 283–287.
14 Kato, T., Esumi, M., Yamashita, S., Abe, K., and Shikata, T. (1992) *Virology*, **190**, 856–860.
15 Itzhaki, J.E. and Porter, A.C.G. (1991) *Nucleic Acids Res.*, **19**, 3835–3842.
16 Porter, A.C.G. and Itzhaki, J.E. (1993) *Eur. J. Biochem.*, **218**, 273–281.
17 Rudiger, N.S., Gregersen, N. and Kielland-Brandt, M.C. (1995) *Nucleic Acids Res.*, **23**, 256–260.
18 Wahls, W.P., Wallace, L.J. and Moore, P.D. (1990) *Mol. Cell. Biol.* **10**, 785–793.
19 Kabori, J.A., Strauss, E., Minard, K. and Hood, L. (1986) *Science*, **234**, 173–179.
20 Chandley, A.C., and Mitchell, A.R. (1988) *Cytogenet. Cell Genet.*, **48**, 152–155.
21 Krowczynska, A.M., Rudders, R.A. and Krontiris, T.G. (1990) *Nucleic Acids Res.*, **18**, 1121–1127.
22 Wahls, W.P., Wallace, L.J. and Moore, P.D. (1990) *Cell*, **60**, 95–103.
23 Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbour Laboratory Press, NY.
24 Rasheed, S., Nelson-Rees, W.A., Toth, E.M., Arnstein, P. and Gardner, M.B. (1974) *Cancer*, **33**, 1027–1033.
25 Gey, G.O., Coffman, W.D. and Kubicek, M.T. (1952) *Cancer Res.* **12**, 264.
26 Fogh, J., Wright, W.C. and Loveless, J.D. (1977) *J. Natl. Cancer Inst.* **58**, 209–214.
27 Shih, C., Padhy, L.C., Murray, M. and Weinberg, R.A. (1981) *Nature*, **290**, 261–264.
28 Marshall, C.J., Franks, L.M. and Carbonell, A.W. (1977) *J. Natl. Cancer Inst.*, **58**, 1743–1747.
29 Klein, E., Klein, G., Nadkarni, J.S., Nadkarni, J.J., Wigzell, H. and Clifford, P. (1968) *Cancer Res.*, **28**, 1300–1310.
30 Jacobs, J.P., Jones, C.M. and Baille, J.P. (1970) *Nature*, **227**, 168–170.
31 NIGMS Human Genetic Mutant Cell Repository 1990/1991 Catalog of Cell Lines: Supplement (1991) NIH Publication No. 91-20011.

32 Lawlor, D.A., Warren, E., Ward, F.E. and Parham, P. (1990) *Immunol. Rev.* **113**, 147–185.

33 Kawakami, T.G., Huff, S.D., Buckley, P.M., Dangworth, D.L., Snyder, S.P. and Gilden, R.V. (1972) *Nature*, **235**, 170–171.

34 Simizu, B., Rhim, J.S. and Wiebenga, N.H. (1964) *Proc. Soc. Exp. Med. Biol.*, **125**, 119–123.

35 Porter, A.C.G., Chernajovsky, Y., Dale, T.C., Gilbert, C.S., Stark, G.S. and Kerr, I.M. (1988) *EMBO J.*, **7**, 85–92.

36 Sanger, F. (1981) *Science*, **214**, 1205–1210.

37 Marchuk, D., Drumm, M., Saulino, A. and Collins, F.S. (1990) *Nucleic Acids Res.*, **19**, 1154.

38 Mathias, N., Bayes, M. and Tyler-Smith, C. (1994) *Hum. Mol. Genet.* **3**, 115–123.

39 Laird, P.W., Zijderveld, A., Linders, K., Rudnicki, M.A., Jaenisch, R. and Berns, A. (1991) *Nucleic Acids Res.* **19**, 4293.

40 Koop, B.F., Goodman, M., Xu, P., Chan, K. and Slightom, J.L. (1986) *Nature*, **319**, 234–238.

41 Gray, I.C. and Jeffreys, A.J. (1991) *Proc. R. Soc. Lond. B*, **243**, 241–253.

42 Stephens, R.M. and Schneider, T.D. (1992) *J. Mol. Biol.* **228**, 1124–1136.

43 Reed, R. and Maniatis, T. (1986) *Cell*, **46**, 681–690.

44 Horowitz , D.S. and Krainer, A.R. (1994) *Trends Genet.*, **10**, 100–106.

45 Flint, J., Taylor, A.M. and Clegg, J.B. (1988) *J. Mol. Biol.*, **199**, 427–437.

46 O'Hara, P.J., and Grant, F.J. (1988) *Gene*, **66**, 147–158.

47 Smith, G.P. (1974) *Cold Spring Harb. Symp. Quant. Biol.* **38**, 507–513.

48 Smith, G.P. (1976) *Science*, **191**, 528–535.

49 Smythe, J.A., Coppel, R., Day, K.P., Martin, R.K., Oduola, A.M.J., Kemp, D.J. and Anders, R.F. (1991) *Proc. Natl. Acad. Sci. USA*, **88**, 1751–1755.

50 Orago, A.S. and Facer, C.A. (1991) *Clin. Exp. Imunol.*, **86**, 22–29.

51 Bird, A.P. (1987) *Trends Genet.*, **3**, 342–347.