Human minisatellite loci composed of interspersed GGA–GGT triplet repeats

JOHN A. L. ARMOUR1, MOIRA CROSIER1, SUE MALCOLM2, JOSEPH C.-T. CHAN2 and ALEC J. JEFFREYS1

1 Department of Genetics, University of Leicester, University Road, Leicester LE1 7RH, U.K.
2 Institute of Child Health, University of London, 30 Guilford Street, London WC1N 1EH, U.K.

SUMMARY
We have isolated two tandemly repeated loci from human DNA which contain long blocks of GGA and GGT trinucleotide repeats. These two repeat unit types, together with other less common variants, are apparently irregularly interspersed along each repeat array. Genotyping methods have been developed for these highly polymorphic loci, including typing by polymerase chain reaction followed by Southern blot hybridization. Linkage analysis in Centre d'Etude du Polymorphisme Humain (CEPH) pedigrees has been used to map the loci to chromosomes 15 and 22. In normal individuals, alleles at these loci can contain thousands of repeats, greatly exceeding repeat copy number at most trinucleotide and other simple repeat loci. No evidence for longer, higher-order repeats was observed among the limited number of repeats sequenced. These loci may represent a transitional state between simple repeat loci and some minisatellites.

1. INTRODUCTION

Blocks of tandemly repeated DNA are abundant in the genomes of humans and other higher eukaryotes; these repeat arrays can display multiallelic polymorphism over a wide variety of size ranges, from short dinucleotide repeat arrays to blocks of satellite up to 5 Mb in length (Willard 1991). In general, there is a correspondence between the repeat unit size and the overall extent of the repeat array; however, some examples have been documented of unusually long dinucleotide (Wilkie & Higgs 1992) and tetranucleotide arrays (Boylan et al. 1990; Brereton et al. 1993; Gibbs et al. 1993). We describe two human loci composed of triplet repeats, which display substantial length polymorphism, and at which alleles can contain more than 1000 repeat units.

2. MATERIALS AND METHODS

(a) Southern blot hybridization

Genomic DNA or polymerase chain reaction (PCR) fragments were blotted onto Hybond-N+ (Amersham). Hybridizations using repeated probes were done in phosphate–SDS buffer (Church & Gilbert 1984) as described, and those using single-copy probes in 0.45 M sodium phosphate pH 7.2, 2% SDS, 1 mM EDTA and 0.5% dried milk powder (Vergnaud 1989). Hybridizations were done overnight at 65°C; for hybridizations to unamplified genomic DNA, alkali-denatured human DNA was added to a final concentration of 15 μg ml−1 as competitor (Wong et al. 1987).

The single-copy probes used for genotyping unamplified genomic DNA were a 1.1 kb SaI fragment from pMS633c (‘A’ in figure 2) at D22S679, and a 550 b.p. PsI fragment from pMS633c (‘B’ in figure 2) at D15S560; the single-copy STS probe was amplified from human genomic DNA using the primers 633SKA (5’TCAACCAAAGGCACTGAGTG-3') and 633SKB (5’GGTAGGCTTCCTCAGTGA3'). For detection of PCR products, the repeat-containing PsI fragment from pMS633c was used; because of the similarity of the repeat arrays, this probe was used to detect PCR products at both loci.

(b) Subcloning and sequencing

Primary subclones pMS632 and pMS633 were constructed by ligation of the Sau3AI insert from the corresponding Charomid clone into the BamHI site of pBluescriptII KS+ (Stratagene). The subclones (pMS632c and pMS633c) containing shorter repeat arrays were prepared by transformation into a rec+ strain of E.coli (C600), in which tandem repeat arrays are unstable and undergo deletions (Kelly et al. 1989). After preparation of plasmid DNA, size selection was used to recover the shortest DNA molecules (containing fewest repeats) from the heterogeneous population of deleted plasmids. Upon retransformation into E.coli XL1-Blue (recA), plasmids with shorter inserts were isolated, and the recombinant clones could now be propagated to give high yield of plasmid DNA. DNA sequence was determined from these clones and fragments isolated from them, by using T7 DNA polymerase (Tabor & Richardson 1987) and single-stranded DNA rescued from pBluescript recombinants. The data have been deposited with the EMBL DNA sequence database (accession numbers X79555–X79558).

(c) PCR conditions

D22S679 was amplified from 200 ng genomic DNA using primers 1834 (5’GGGAACCAGAGAGGTGATCGAGG-
3') and 1835 (5'CAAGCTGCTGACCCTGCCTCAGTC3') in 10 μl reactions using 0.5U Taq polymerase in the buffer described (Jeffreys et al. 1990). Alleles at D15S560 were amplified using primers 1368 (5'CTGGTGAGAAGGGAGCTGGGCGATG3') and 1369 (5'CATTGACATTACACCAGGCCAGAGCC3') using the same conditions. For both loci, the cycle conditions were 95 °C 1 min, 67 °C 1 min, 70 °C 10 min; 20 cycles were used at D22S679, and 18 cycles at D15S560. Products were resolved on 0.8% agarose gels and detected by Southern blot hybridization with a tandemly repeated probe as detailed above.

3. RESULTS

(a) Isolation and subcloning

The Charomid clones cMS632 and cMS633 were isolated from a human genomic library of size-selected (4–9 kb) MboI fragments cloned into Charomid 9–36 (Saito & Stark 1986) and screened by hybridization with DNA fingerprinting probes (Armour et al. 1990). When used as hybridization probes at high stringency on human genomic DNA, inserts from both these clones gave similar profiles composed of many hybridizing fragments, of which many were polymorphic (figure 1a). This suggested that both cloned inserts contained blocks of similar simple sequence repeats present at many locations in the human genome.

The Sau3A inserts were subcloned into the BamHl site of pBluescriptII vectors; like the original Charomid clones, both of these plasmid subclones (pMS632 and pMS633) gave very poor yields of plasmid DNA. Because clones containing substantial blocks of repeated sequences frequently displayed poor growth properties in E. coli, we selected for plasmids containing shortened repeat blocks (see §2). These shortened subclones (pMS632c and pMS633c) were used in subsequent analyses.

(b) Structural and sequence analysis

Structural analysis of pMS632c showed that the tandem repeats were contained within a 1 kb Smal-Sau3A1 fragment (figure 2), from which the DNA sequence was determined; DNA sequences determined from the repeats at the 3' ends of duplicate pMS632c clones were identical at the proximal end of the array. In pMS633c, the tandem repeats were located centrally in the cloned insert (see figure 2); sequence data were therefore obtained from either end of the clone, to give sequence corresponding to the cloned flanking DNA and the extremities of the tandem array. Both repeated arrays consisted of irregularly interspersed GGA and GGT triplet repeats (= TCC—ACG), with additional variants occurring less frequently (figure 2). It is not possible, from the analysis of sequence at the extremities of deleted clones, to exclude the possibility that higher-order repeats may have been lost during subcloning. However, in neither array did inspection of the

Figure 1. Analysis of DNA from three unrelated people at the loci described, (a) Southern blot hybridization of MboI-digested DNA using the entire Sau3A1 insert from pMS632 as probe, revealing multiple polymorphic loci. A similar result was obtained with the whole Sau3A1 insert from pMS633 (data not shown). (b) and (c) show analysis of DNA from the same three people at the individual loci D22S679 and D15S560, using either Southern blot hybridization of MboI-digested DNA with a unique sequence flanking probe (left) or PCR followed by hybridization with a repeat probe (right).
sequence reveal any evidence for higher-order repetition.

(c) Locus-specific genotyping

Locus-specific genotypes could be obtained directly by Southern blot hybridization of Mbol-digested genomic DNA using single-copy flanking probes isolated as restriction fragments from the clones (A and B in figure 2). At D15S560 (MS633), a locus-specific flank ing probe corresponding approximately to fragment B could also be generated as a ‘sequence-tagged site’ (STS) by PCR between primers 633SKA and 633SKB (data not shown: see figure 2 and §2). PCR using specific flanking primers, followed by Southern blot hybridization using a repeat array probe, could also be used to genotype these loci (figure 1b, c). PCR typing at these loci resulted in the additional production of a smear of cross-hybridizing products, perhaps reflecting the cumulative effects of polymerase slippage events. However, this additional material migrates more slowly on agarose gels than the canonical products and may therefore be caused by intermolecular heteroduplexes or other complex structures rather than simple polymerase slippage, which would be expected to give rise to shortened products. Treatment of PCR products with S1 nuclease (Jeffreys et al. 1988) did not reduce the intensity of this smear; the addition of small amounts of a proofreading thermostable (Pfu) DNA polymerase (Barnes 1994) gave a small improvement in the yield, but no improvement in the fidelity (data not shown). The origin of the additional products arising during PCR at these loci remains uncertain.

Amplified alleles from 20 unrelated individuals of Caucasian (CEPH) origin ranged in size between 1.8 kb and 9 kb at D15S560 (MS633) (observed heterozygosity 80%) and 1.6 kb and 3.5 kb at D22S679 (MS632) (observed heterozygosity 80%). Analysis of segregation in CEPH kindreds and of DNA from somatic cell hybrids as described (Armour et al. 1990) was used to map the loci: using the Genethon microsatellite linkage maps (Weissenbach et al. 1992) as reference, D15S560 maps in the interval D15S122-D15S118; analysis of father-mother-child trios in spontaneous deletion cases of Angelman’s syndrome showed biparental inheritance (data not shown), thus mapping the locus outside the regions involved in these deletions (Pembrey et al. 1989). Linkage analysis in CEPH pedigrees suggested that D22S679 is located distal to D22S277 on chromosome 22q.

4. DISCUSSION

We have shown that two human loci composed of long arrays of interspersed trinucleotide (GGA-GGT) repeats are highly polymorphic. Because full-length repeat arrays cannot be conveniently analysed, sequence data have not been obtained for entire alleles at either of these loci; those repeat units which have been sequenced appear to be entirely irregular in their distribution, and inspection of sequencing gels at higher molecular weights similarly reveals no evidence for larger, higher-order repeats. The multilocus polymorphism produced on Southern blot hybridization with repeat-containing probes from pMS632 and pMS633, together with the sequence similarity, suggests that these loci may contribute to (CAC)n DNA fingerprint profiles (Ali et al. 1986).

While it remains possible that higher-order repeats may have been preferentially lost during the production of the shorter subclones pMS632c and pMS633c, these loci appear to be more similar to large, irregular di- and tetranucleotide repeat loci (Boylan et
al. 1990; Wilkie & Higgs 1992; Brereton et al. 1993; Gibbs et al. 1993) than to those minisatellites at which the repeated unit bears the traces of expansion from shorter repeats, such as MS29 (Wong et al. 1990), DXYS82 (Armour et al. 1992) and minisatellite loci cloned using (CAC)\textsubscript{n} probes (Zischler et al. 1992). At the latter loci the higher-order minisatellite repeats, which are still discernible even in rearranged clones, appear to have arisen from a pre-existing array of simple repeats. Therefore it is possible that minisatellites identified with (CAC)\textsubscript{n} probes and the loci reported here simply represent two extremes of a continuum, in which large (but irregular) blocks of triplet repeats provide the raw material for the subsequent emergence of higher-order repeats, and thus for the evolution of the locus into a larger minisatellite structure. Bearing in mind that only alleles which can be resolved on an agarose gel will be scored as distinct, and thus, that these loci may at the sequence level be much more informative than is apparent from this typing method, it is not possible to predict simply whether a stepwise mutational model is appropriate for these loci. The unit of mutation at these loci may indeed be a single triplet repeat, but it is also possible that mutations may frequently change by longer, irregular units containing several triplet repeats.

Extrapolation from the behaviour of loci such as FRAXA, FRAXE and DM might suggest that such very long triplet repeat arrays would be ‘ultravariable’ loci, with very high rates of mutational instability in soma and germline. Indeed, the long mouse tetrancleotide hm-2 (Gibbs et al. 1993) was identified as a locus showing extreme somatic and germline instability. By contrast, like other very long simple repeat arrays described to date in humans, the loci reported here do display substantial levels of polymorphism, but in the small number of transmissions analysed did not display the extreme germline and somatic instability displayed by full mutation alleles at, for example, FRAXA (Nakahori et al. 1991). The relative stability of these very long arrays may result from the irregularity of the repeat array: for different classes of simple repeats there is a correlation between instability and the size of perfect repeat blocks (Weber 1990; Gibbs et al. 1993; Hirst et al. 1994). To date, all examples of dramatic expansions have involved AGC or CCG repeats, so it may be that these motifs alone are commonly subject to very rapid expansions, correlated perhaps with potential for hairpin formation, and therefore that GGA and GGT repeats are not subject to such dramatic expansions. However, it is also possible that the loci described in this report may represent the longer-term result of ‘explosive’ expansions at (phenotypically neutral) triplet repeat loci.

We thank Sue Povey for DNA from somatic cell hybrids, and Esther Signer, Nicola Royle, Celia May, Joanna Coleman, Mark Jobling, Andy Collick and Yuri Dubrova for helpful comments on the manuscript. We are grateful to the U.K. HGMP Resource Centre for computing facilities and oligonucleotide synthesis. This work was made possible by grants from the Wellcome Trust and the M.R.C.-G.T.-C. is supported by Action Research. A.J.J. is an International Research Scholar of the Howard Hughes Medical Institute.

REFERENCES


Wilkie, A. O. M. & Higgs, D. R. 1992 An unusually large


Received 3 May 1995; accepted 14 June 1995.