The EMBO Journal vol.13 no.13 pp.3203–3210, 1994

Complex recombination events at the hypermutable minisatellite CEB1 (D2S90)

Jérôme Buard[1,2] and Gilles Vergnaud[1,2,3]

[3]Laboratoire de Génétique Moléculaire, Centre d'Etudes du Bouchet, BP 3, 91710 Vert le Petit and [1]Laboratoire de Génétique des tumeurs, INSERM CJF9201, Institut Curie, 26 rue d'Ulm, 75231 Paris, France
[2]Present address: Laboratoire de Génétique des Espèces, Institut de Biologie, 9, Quai Moncousu, 44035 Nantes Cédex 01, France

Communicated by J.Weissenbach

Some minisatellite structures are the site of high rates of DNA recombination in non-pathological situations, with an excess of motif insertion events and a locus-dependent sex-specific mutation bias. We previously reported the cloning of the hypermutable minisatellite locus CEB1 (D2S90), remarkable for its 13% mutation rate in the male germline (compared to ~0.4% in female). We have sought to analyse the mechanisms underlying the addition or deletion of motifs at this locus using the minisatellite variant repeat mapping technique. This is possible with a high precision due to the extreme sequence polymorphism seen between different motifs. No crossing-over event was observed among 38 informative neomutations. Four of the 19 informative mutant alleles with an addition of motifs are interallelic events, the others are intra-allelic. Overall, the insertion and deletion mutations are spread along the alleles, although the subset of interallelic events shows clustering towards the analysed end. The apparently complex recombination events observed can all be interpreted as a succession of elementary duplications–deletions of inter- as well as intra-chromosomal origin, suggesting a model in which sister chromatid as well as conversion-like exchanges are involved in these mutation processes.

Key words: hypermutability/minisatellite/mutation/recombination/tandem repeat

Introduction

A number of recombination events are responsible for the evolution and plasticity of genomes, and also of many pathologies. These include general (homologous) recombination, translocation, gene conversion, sister chromatid exchange (SCE), insertion of transposable elements and variation in tandemly repeated DNA sequences that are not yet part of genes, but might be at the origin of new coding sequences (Ohno, 1985). The natural instability of tandem repeats makes them interesting structures to reveal and analyse some of the mechanisms underlying genome evolution in higher eukaryotes.

Tandem repeats are usually classified according to arbitrary morphological features which *a posteriori* do appear to reflect biological differences. The shorter tandem repeats, called microsatellites, have 1–5 bp long elementary units spanning not more than a few tens of nucleotides. Minisatellites have units usually in the range of 10–100 nucleotides, and the overall structure size range is 0.5–100 kb. Some such structures are polymorphic genes (Simon *et al.*, 1991), or are directly associated with genes and may be involved in their regulation (Krontiris *et al.*, 1993; Lucassen *et al.*, 1993). Tandem repeats are often highly polymorphic in the population due to variations in the number of repeats and they are instrumental in the construction of genetic maps for mammals. Unstable trinucleotide repeats are causally involved in fragile sites and certain human genetic disorders.

In spite of their similar structure, minisatellites and microsatellites appear to be strikingly different in many respects. To begin with, minisatellites have units with internal variations between repeats (Jeffreys *et al.*, 1990). Secondly, in humans, microsatellites are spread all over the genome (Weissenbach *et al.*, 1992), whereas minisatellites appear to cluster towards telomeres (Royle *et al.*, 1988; Vergnaud *et al.*, 1991; NIH/CEPH Collaborative Mapping Group, 1992; Vergnaud *et al.*, 1993). Thirdly, they differ clearly by their relative mutation rate in somatic versus germline conditions: mutation events occur at microsatellite loci in lymphoblastoid cell lines (manifested sometimes as three alleles) (Weber and Wong, 1993), although the germline mutation rate of the thousands of microsatellites isolated so far is <1% (Weissenbach *et al.*, 1992; Weber and Wong, 1993). In comparison, minisatellite CEB1 with a 13% mutation rate in the male germline has shown no instance of mosaicism among the CEPH DNA samples analysed (Vergnaud *et al.*, 1991). Finally, the mutation mechanisms operating predominantly at micro- and minisatellites appear to be different. (CA)*n* microsatellite tracts are destabilized in yeast strains mutated for DNA mismatch repair genes (Strand *et al.*, 1993), suggesting that DNA polymerase slippage, rather than recombination events, plays an important part in the production of microsatellite mutations and showing that the apparent increase in mutation rate is not due to an increase in the number of mutation events, but to the absence of their repair. The validity of the model has been demonstrated by the subsequent finding that the abnormally high mutation rate observed on a genome-wide scale at (CA)*n* repeats in some tumours (Aaltonen *et al.*, 1993; Ionov *et al.*, 1993; Thibodeau *et al.*, 1993) is caused by defective mismatch repair (Fishel *et al.*, 1993; Leach *et al.*, 1993; Parsons *et al.*, 1993). Interestingly, the frequency of larger events—deletions or insertions of >4 bases—is apparently unaffected in the yeast model, reminiscent of the reduced efficiency of the homologous *Escherichia coli* mutHLS system for detecting heterologies >4 bp (Parker and Marinus, 1992). Alternatively, other mechanisms may be responsible for these larger events. This view is further supported by the demonstration that

interallelic recombination is responsible for an important proportion of germline mutation events at the hypermutable minisatellites MS32, MS31A and MS205 (Jeffreys *et al.*, 1991, 1994).

These observations suggest that a combination of mechanisms operate to generate the variability of tandem repeats, among which are replication slippage and recombination. It is likely that the respective role of these different factors is modulated by the tandem repeat sequence itself, but also by some, as yet poorly understood, epigenetic features. Hypermutable minisatellites may then represent a tool to study these aspects of genetic recombination in mammals.

The situation of minisatellites with respect to mutations and mutation rates is very heterogeneous, with monomorphic loci at one end and at the other end extremely unstable loci such as MS1 (Jeffreys *et al.*, 1988) and CEB1 (Vergnaud *et al.*, 1991) with average mutation rates per gamete of 5 and 7%, respectively. So far, ~10 human hypermutable minisatellite loci, with a sex average mutation rate >1%/gamete, have been isolated (Jeffreys *et al.*, 1988; Vergnaud *et al.*, 1991; our unpublished results). The mutation events are not reciprocal, with an excess of insertion events, and a parental effect is observed in the origin of mutations: MS1 mutants are equally of paternal or maternal origin; CEB1 mutants are almost exclusively of paternal origin. These two features, very high mutation rate and paternal bias, suggest that the mutation processes operating at different hypermutable minisatellite loci may be of a different nature. However, we show here that recombination is responsible for some, if not all, germline mutation events at CEB1, as seen for MS32 (Jeffreys *et al.*, 1994), and propose a model in which recombination is initiated by staggered nicks resulting in protruding single-strand DNA ends.

Results

General features of CEB1 (locus D2S90): sequence and internal heterogeneity

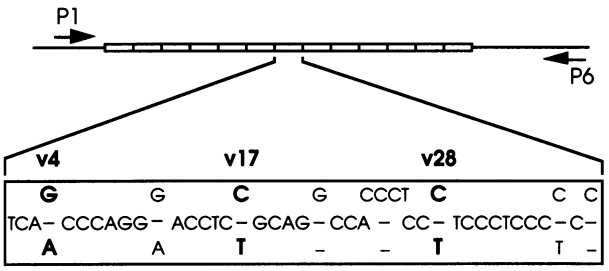
The minisatellite CEB1 (Vergnaud *et al.*, 1991) has a 13% average mutation rate in male meiosis; the screening has now been extended to the 61 CEPH reference pedigrees (565 informative male meioses) in which 73 children inherited a new allele from their father versus two from their mother. Two-thirds of these mutations are increases in the number of repetitions.

The repeated unit of CEB1 is 37–43 nucleotides long with a 70% GC content and a strand disymmetry. There is a high degree of internal variation, with eight polymorphisms found among 20 motifs sequenced from a single allele (Figure 1). Three of the polymorphisms are insertion–deletion events. One of them, the CCCT insertion, is part of an 'internal microsatellite' with three or four CCCT motifs.

The length of the repeat unit, together with the average size allele of 2–3 kb, ensures that the majority of even single-motif insertion–deletion events are detected by Southern blotting.

Variability of the two ends of the tandem array

The internal map of first typed motifs from 46 unrelated alleles and of the apparently last motifs from 30 alleles has been generated in the course of this study. Fourteen different states are observed at the first typed motif, with an allele distribution predicting a heterozygosity of >80%. Eight states are observed at the last motif, with a heterozygosity prediction of 85% (data available on request).



Name	5' Sequence 3'	Conc (nM)
TAG	TCATGCGTCCATGGTCCGGA	1000
P1	ggtctagAGCTCTGCTGAGTCAGAGTCAGCCAG	1000
P6	AAACTGTAATCTGGAGTTGGTCTGGCGATC	1000
v4-G(11G)	(*TAG*)-TGCGGAGGT**C**CCTGGG**C**	4
v4-G(11A)	(*TAG*)-TGCGGAGGT **T**CCTGGG**C**	1
v4-A(11G)	(*TAG*)-TGCGGAGGT**C**CCTGGG**T**	5
v4-A(11A)	(*TAG*)-TGCGGAGGT **T**CCTGGG**T**	5
v17-C	(*TAG*)-AGGGAGGGTGGCCTGC**G**	5
v17-T	(*TAG*)-AGGGAGGGTGGCCTGC**A**	5
v28-C	(*TAG*)-GAGGGGGGAGGGA**G**GGT	10
v28-T	(*TAG*)-GAGGGGGGAGGGA**A**GGT	20

Fig. 1. Sequence of a CEB1 repeat and primers used for MVR-PCR typing. P1 and P6 are flanking primers used to amplify the alleles. P1 on one side overlaps on the first motif (8 nucleotides) and P6 on the other side is located ~250 bp from the last motif. The sequencing of 20 motifs from a single allele revealed eight nucleotidic variations: three of them are insertions/deletions and five are transitions. The three variations (v4, v17 and v28) are selected for the MVR-PCR typings. Interference of other polymorphisms with the v4 variation is greatly reduced by partially degenerating v4 primers at the 11th position. In MVR-PCR experiments, primers P1 and TAG are used with either primers v4-G(11G) and v4-G(11A) or v4-A(11G) and v4-A(11A) or v-17C or v-17T or v-28C or v-28T.

Typing of mutant alleles

Three evenly spaced nucleotidic variations (v4, v17 and v28; Figure 1) were chosen for their informativity and their relative location along the repeat to type alleles in minisatellite variant repeat mapping–polymerase chain reaction (MVR-PCR) experiments. The typing of each variant is done in different tubes, each loaded on one lane. The typing of the three variant sites results in six lanes for each sample. Figure 2 shows an example of MVR mapping applied to a single CEB1 allele. For each polymorphism there are three possible statuses [e.g. G or A or 'null' noted 0; the 'null' status being most probably due to interference of nearby polymorphisms with primer extension (Tamaki *et al.*, 1992)]. The reading of the internal map is similar to the reading of a sequencing gel. In practice, 18 of the 27 potential haplotypes were encountered among 66 unrelated alleles (2813 motifs). Fifty mutant alleles were analysed. In nine cases, the

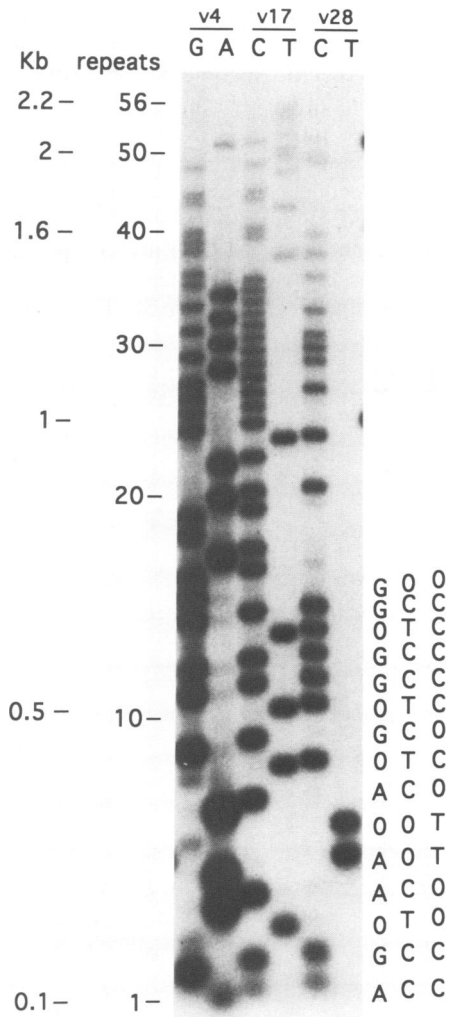


Fig. 2. MVR-PCR adapted to CEB1 and typing of individual alleles. Three nucleotidic variations between repeats are used: v4, v17 and v28. For each polymorphism, two primers generate two complementary ladders. When considered simultaneously, the six ladders allow the coding of each repeat within an allele (in this case the 2.2 kb allele of CEPH individual 144707) as a triplet.

rearrangement presumably occurred beyond the range of the analysis (2.5 kb). Owing to the high discrimination power between repeats, we could deduce the nature of the rearrangement by comparing the internal map of the new allele with the parental ones for 38 mutations out of 41 (Figure 3). No crossing-over event involving the disjunction of sequences flanking the mutation was found.

Insertion events. Among the 23 size increases, 19 typings were informative enough to distinguish 15 intra-allelic from four interallelic events. The 15 intra-allelic events can be classified into two groups: in 12 events (numbered 1–12 in Figure 3A), the typing for the original progenitor allele can be recovered from the mutant allele by the imaginary deletion of some contiguous motifs, but this is not possible in cases 13–15. Cases 1–5 are perfect duplications. In cases 6–10, the duplication has been rearranged, and once again the secondary rearrangement falls in these two categories: 6 and 7 contain a perfect duplication within the duplication, 8, 9, 10, 11 and 12 are more complex events from which the original matrix cannot be reconstituted by a simple deletion. Cases 11 and 12 are instances of triplication attempts. Each of the four cases of interallelic events (numbered A–D in Figure 3B) are of a different type. Case A is a perfect insertion of two units within the recipient allele. In B, the insertion has lost one unit. In C, the insertion has been duplicated and one motif within the duplication has been duplicated. D is an interesting case with generation of a direct repeat flanking the inserted motifs. The 41 insertion and deletion events taken together are evenly distributed along the progenitor alleles (Figure 4). Considered separately, the position of the four interallelic events in the first tenth of the minisatellite (Figure 3) may be reminiscent of what was observed for D1S8 (Jeffreys *et al.*, 1990, 1991, 1994). However, the most distal part has not been explored as extensively as the proximal part because MVR-PCR was only run from the flanking primer P1. We do not know the proportion of interallelic events among the nine events for which the rearrangement has not been detected so that we cannot at present exclude a symmetrical clustering of interallelic events at the other extremity of the array.

Deletion events. Deletion events are significantly less frequent than insertion events in a ratio of 2/1. However, the total number of motifs lost in the mutation events detected within the CEPH families is not significantly different from the total gain of motifs (i.e. the average number of motifs lost in a deletion event is higher than the average number gained in an insertion). Eighteen deletion events have been analysed in this study. None is associated with a crossing-over event. The events appear to be regularly spaced along the alleles (Figure 4). Most can be interpreted as the clear deletion of a contiguous set of motifs. In one case, however, a deletion of 15 motifs is clearly accompanied by an insertion of two motifs of intra- or interallelic origin (data not shown).

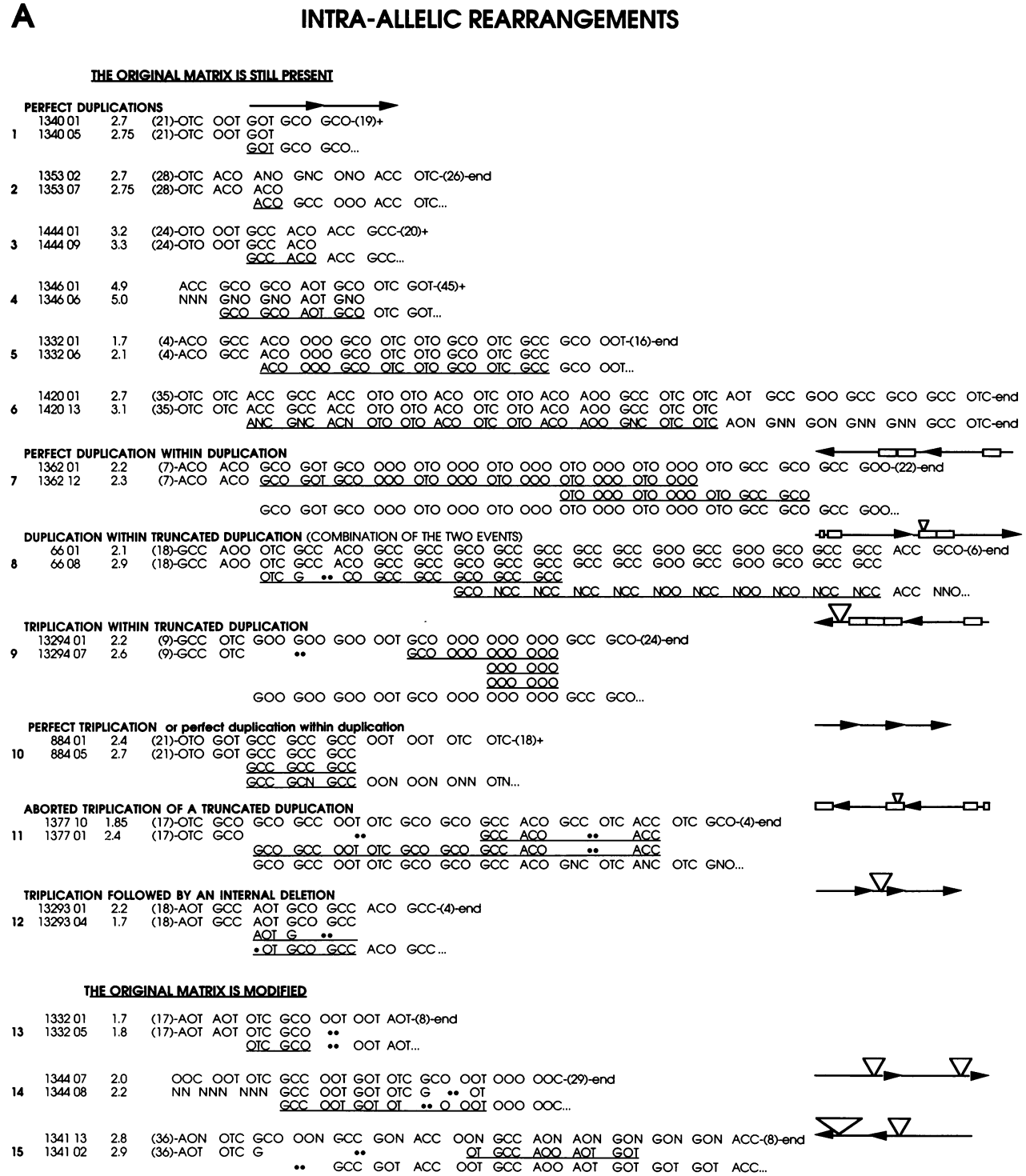
Discussion

The heterogeneity of CEB1 motifs: a new tool for forensics?

The present report analyses mutation events at the minisatellite locus CEB1 (D2S90), characterized by the highest mutation rate observed so far (7% sex-averaged) and a very strong paternal bias (13% in male meiosis) with 2/3 insertion events. We demonstrate here that this hypermutability is not associated with a high internal homogeneity, but that on the contrary the degree of variation among the CEB1 repeat units is remarkably high (Figure 1) with eight insertion–deletion or transition polymorphisms found among 20 motifs sequenced from a single allele. This is in sharp contrast with what has been reported for the three hypermutable loci at which MVR-PCR has been adapted so far: MS32 (Tamaki *et al.*, 1993), MS31A (Neil and Jeffreys, 1993) and MS205 (Armour *et al.*, 1993). Interestingly, both ends of CEB1 alleles appear to be similarly variable. The observation of such variability of CEB1 motifs may provide a new tool for forensic analyses, based on the sequencing of a few motifs from the ends of CEB1 alleles. Many very degraded DNA samples used in forensics still yield fragments of up to 80–100 nucleotides.

A model for the generation of the interallelic germline recombination events at CEB1

The high internal heterogeneity makes possible the development of internal maps with polymorphism information every 13 nucleotides. From these maps, the position and nature of the mutation events can be deduced. The absence of crossing over among the mutation events strongly suggests that models usually invoked to explain homologous recombination features are not appropriate in the present situation. The additional observation that deletion events are significantly less frequent than insertion events is further evidence that the mutation events are not reciprocal. The mutation events may thus be classified as illegitimate, site-specific recombination events. At first glance, the rearrangements observed can be classified into two major groups: intra- (75%) and inter- (25%) allelic events. However, it is intriguing to find, within these two groups, similar secondary events, i.e. duplications, either perfect, or deleted, or with an internal duplication. The neomutation events analysed here are comprised of two basic events, insertions and deletions, which can be combined in various ways (Figure 3). The accumulation



B **INTER-ALLELIC EXCHANGES**

INSERTION WITHOUT ANY MODIFICATION

	1424 01	4.1	ACC GOT GOT GOT GCO-(26)+
	1424 01	1.2	goo gco oto gco gco oto oto-(12)-end
A	1424 10	1.3	goo GOT GOT gco oto gco gco oto oto...

DELETION WITHIN THE INSERTED GROUP

	1346 01	4.9	ACC GCO GCO AOT GCO OTC GOT OCO-(44)+
	1346 01	5.1	ooc gcc gcc gcc-(43)+
B	1346 07	5.3	ooc gcc GCO AOT •• OTC gcc gcc-(43)+

DUPLICATION OF THE INSERTED GROUP (AND DUPLICATION WITHIN THE DUPLICATION)

	13293 01	2.2	AOT GCC OTC GCC GCC OOT OTC OTO OTO GOT-(37)-end
	13293 01	1.55	gcc gcc aoo acc got oot-(23)-end
C	13293 06	1.9	nnn gcc aoo acc GCC OOT OTC OTO
			GCC OOT OTC
			OTC OTO got oot-(23)-end

DUPLICATION OF THE FLANKING RECEIVER GROUP ON THE OTHER SIDE OF THE INSERTION

	1582 01	3.1	AOO GCO GOO GOO ACO GON ACO GCC GCC GCC GOC-(39)+
	1582 01	4.7	acc oot got ooc gco got gcc oto oto-(52)+
D	1582 07	4.8	acc oot got ooc gco got ACO GCC GCC gco got gcc oto oto...

Fig. 3. MVR-PCR typing of insertion events. Nineteen insertion events have been analysed in detail, of which 15 (numbered 1–15) are intra-allelic (A) and four (numbered A–D) are interallelic (B). The progenitor of the mutant alleles is indicated above. Each line indicates the CEPH DNA sample identification (e.g. kindred 1340, individual 01), the size in kilobases of the allele typed (e.g. 2.7 kb), the number of repeats before the first listed [e.g. (21), and none when nothing is indicated]. 'N's indicate ambiguous typings. Motifs inserted are underlined in the mutant allele and the different copies of duplicated groups are aligned; therefore, mutant alleles are presented on more than one line. Black points show possible deletion events. (**A**) The figures on the right help reconstitute the possible intra-allelic events. The first arrow represents the template group of motifs and points towards the second arrow, which represents the new group. The orientation of the arrows is solely intended to suggest a reading direction for an easier understanding. Event 14, in which the original matrix is modified, illustrates the fact that both elements of a duplication may be the target of secondary events. Secondary rearrangements within a group are illustrated by a box for a subgroup of motifs and a triangle for a deletion. For example, a duplication within the rearrangement (case 7) is shown as two contiguous boxes on the second arrow and only one box on the first arrow. (**B**) For interallelic events, both paternal alleles are shown starting from the first motif. The receiver allele is in lower case and the donor allele in upper case.

in one place of a succession of elementary deletion and duplication events, once a mutation has been initiated, indicates that the biochemical machinery repairing these DNA lesions may be identical for both types of events, and that perhaps the underlying mechanisms are also similar. The evidence that gene conversion-like mechanisms occur, and the similarity of intra- and interallelic secondary rearrangements, suggests that at least some of the intra-allelic events are initiated as SCEs. Among the four interallelic events, case D (Figure 3B), where the two motifs before the insertion are found duplicated after the insertion, is not easily predicted by the model proposed in Jeffreys *et al.* (1994). In this model, derived from the double-strand break recombination model (Szostak *et al.*, 1983; Mézard and Nicolas, 1994), an initially blunt double-strand break is processed into single-strand overhangs by exonucleases. In contrast, event D, combined with the search for a minimum number of mechanisms, leads to the proposal of the existence of the intermediate shown in Figure 5. Two staggered single-strand nicks generate 3′ protruding single strands, as a primary event. The double-strand break is then repaired from the homologue (or perhaps from the sister chromatid) by filling in with one or both strands. In that option (interallelic exchange and filling in with both strands; depicted in Figure 5), three types of gene conversion-like mechanisms are predicted: 1–2 are similar insertion events of a perfect number of motifs (cases A, B and C, with or without secondary events, similar to those seen in intra-allelic events); 3 and 4 are insertion events of an imperfect number of motifs associated, respectively, on the receiving allele, with a duplication flanking the insertion or with a deletion of motifs. In the absence of secondary deletion events, the model further predicts that the size of the insertion should

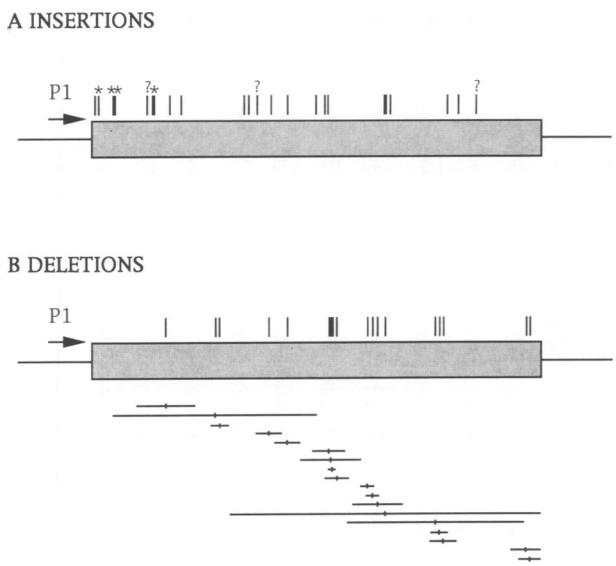


Fig. 4. Relative distribution of the mutations. The intra-allelic mutation events (insertions or deletions) are evenly distributed along the tandem repeat. The four interallelic events are clustered near the P1 end of the minisatellite. (**A**) Insertion events: each vertical bar indicates the relative position of the first repeat of an inserted element along the CEB1 alleles. Stars show the locations of the four increases due to interallelic events and question marks point out the positions of the three increases for which the mechanism was not elucidated. (**B**) Deletion events: each vertical bar indicates the middle of a deleted portion. Below, the relative size of the missing part is indicated with a horizontal bar.

be at least equal to the size of the deletion of the receiving allele. Case D may be interpreted as a type 3 if sequencing can confirm that an imperfect number of motifs has been inserted. Type 4 is not observed among the four

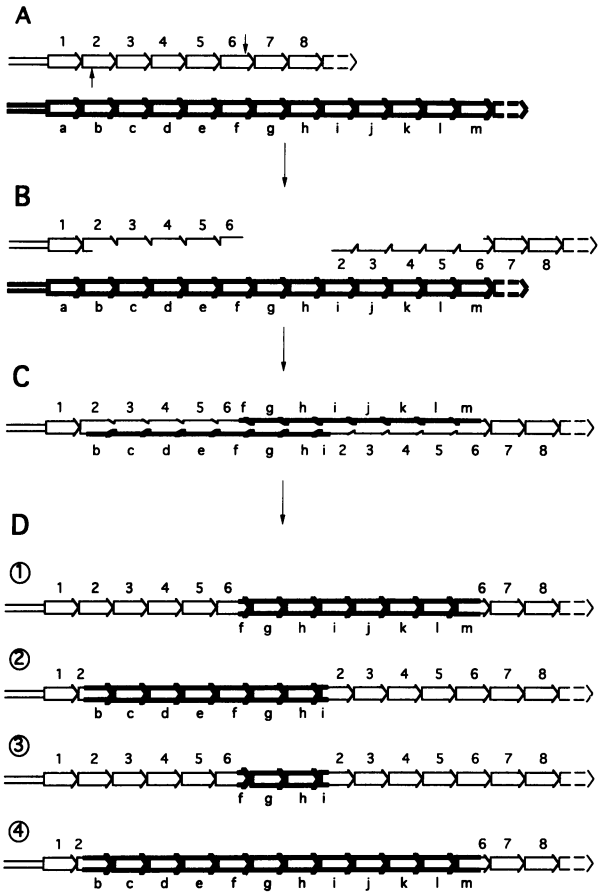


Fig. 5. Model for the generation of inter- and intra-allelic insertion events. (**A**) 3′ protruding single-strand DNA ends are generated by staggered single-strand nicks. (**B** and **C**) The break is repaired by SCE, or from the homologous allele paired in the flanking region. In this drawing, the gap is filled in on both strands by transfer from the donor allele. In that case, two heteroduplex regions are generated when the homologous allele is used (alternatively, if the gap is filled in on only one strand and then the second strand is filled in by copying, there will be only one region of heteroduplex DNA, and in D, only two classes of events: duplication of the flanking motifs or not). The fact that we do not observe new mosaic motifs in the vicinity of the insertions suggests that the mismatches on each side of the insertion are repaired according to one of the two strands (co-repair). In the hypothesis that the two heteroduplex regions are repaired independently, four different products are predicted. (**D**) 1 and 2 are the product of opposite choices for mismatch repair in the heteroduplexes. The resulting events are simple insertion events (or perfect duplications in the case of a SCE). 3 and 4 are the product of identical choices for mismatch repair in the heteroduplexes; 3 results in a duplication of motifs flanking the insertion, while 4 results in a deletion of >3 motifs on the receiving alleles. In 1 and 2, a whole number of motifs is inserted. In 3 and 4, an imperfect number of motifs is inserted. In 4, the insertion (in the absence of secondary deletion events) is predicted to comprise a number of motifs at least equal to the number of motifs deleted on the receiving allele.

interallelic events reported here, but interestingly mutant 42 in Jeffreys *et al.* (1994) is an interallelic event with an insertion of five motifs associated with the loss of five motifs on the receiving allele. If the validity of the model of mutation initiation by protruding ends and its predictions is validated by these future studies, then the question of whether this is a general phenomenon (Nicholls, 1994) responsible for chromosomal abnormalities will be opened. In particular, one may imagine that the production of staggered single-strand nicks, normally restricted to germline or early somatic development, is active in somatic tissues in some pathologies such as Bloom's syndrome (Groden and German, 1992). The observed behaviour of the yeast strain produced using the hypermutable minisatellite MS1 (i.e. showing insertion as well as deletion mutants; Agurell *et al.*, 1994) suggests that it may be possible to develop appropriate experimental models enabling the cloning of the genes responsible for these rearrangements.

The clustering of the four interallelic insertions toward one end of the minisatellite may be the result of some 'in register' DNA–DNA pairing requirement. This is suggested by the striking observation that in all four cases of interallelic exchange (Figure 3), the insertion point on the receiver allele can be aligned (starting from the flanking primer P1) with the beginning of the insertion from the donor allele. It is striking to observe that the same rule is strictly observed for the other hypermutable loci analysed to date (Jeffreys *et al.*, 1994). This requirement is much more easily satisfied within the first few motifs because of the overall size difference usually observed between alleles, thus making pairing throughout the alleles physically impossible. On the contrary, such an alignment requirement will be satisfied equally well all along the alleles in SCEs. This hypothesis does not require the presence of *cis*-acting sequences invoked in other studies (Armour and Jeffreys, 1992; Jeffreys *et al.*, 1994) to explain clustering and predicts that individuals with combinations of very similar alleles might not show such a strong polarity of interallelic mutation events.

The model presented explains the interallelic and intra-allelic insertion events by a common mechanism. Deletions may be the results of the reciprocal events with 5′ single-strand protruding ends, which would be trimmed to generate 3′ protruding ends and would then be processed as proposed above. However, we would expect to observe deletion events having interallelic insertions, which we have not observed unambiguously so far. In addition, it would not explain the 2/1 ratio observed for the number of insertion versus deletion events, unless we postulate that the repair of breaks with 5′ protruding ends, involving additional steps, is less efficient than the repair of 3′ protruding ends, and more frequently subsequently lost. Alternatively, some deletions may result from other one-step intramolecular events.

General considerations and future prospects

Hypermutable tandem repeats are proving to be interesting models to analyse some aspects of DNA recombination in mammals. Some microsatellites with trinucleotide repeat units are responsible for fragile sites or diseases showing increased severity in subsequent transmission (reviewed in Wieringa, 1994), apparently when alleles suddenly cross a size frontier and join the minisatellite allele size class. However, this evolution is probably somatically induced in a particular window of early development, as shown in Imbert *et al.* (1993) and Reyniers *et al.* (1993). It is interesting to note that both hypermutable mouse minisatellites reported so far with, respectively, a penta- and a tetranucleotide repeat unit, and a very high internal homogeneity, show evidence for somatic mutation during early development (Kelly *et al.*, 1989, 1991; Gibbs *et al.*, 1993). This is in agreement with the view that sequence

homogeneity and motif length are important parameters conditioning the predominant mode and timing of evolution of a given tandem repeat, and suggests that at least some of the mutations seen at human hypermutable minisatellite are early somatic events.

The remarkable absence of unequal cross-overs can be explained in view of the size variations among alleles combined with the unexpectedly high degree of divergence between CEB1 repeats. It will prevent the occurrence of the minimum stretch of perfect sequence identity usually found even at abnormal cross-over points (Rouyer *et al.*, 1987). This requirement is much less likely to be satisfied between alleles of polymorphic tandem repeats, making them cold spots for inter-chromosomal crossing over.

The study conducted here was based on mutation events observed among a pool of 60 progenitors. Consequently, the mutation frequency and the rearrangements observed are probably representative of an average behaviour, but do not reflect individual characteristics. Some alleles, or combinations of alleles, may show significantly higher mutation rates, different proportions of intra- and inter-allelic events, a different proportion of each of the rearrangements described, or a different distribution of events along the allele. Also, the role we suggest for SCEs in the generation of mutations at CEB1 implies that it will be important to compare somatic versus germline stability in normal individuals as well as in pathological conditions (such as in Bloom's syndrome patients; Groden and German, 1992) or in tissues exposed to ionizing radiation (Dubrova *et al.*, 1993). These questions will now be addressed by the analysis of a high number of isolated alleles from a few individuals and direct sequencing of recombination junctions. Already, the analysis of a high number of mutants from CEB1 and MS32 is beginning to unravel the general rules and the locus (and/or individual)-specific features of minisatellite evolution.

Materials and methods

Sequencing of CEB1 repeats

Flanking regions and first repeat sequences were obtained from *Alu*I and *Pvu*II CEB1 subclones. CEB1 motifs are cut by *Hae*III and *Pst*I. Internal repeats were thus sequenced from *Hae*III and from *Pst*I puc19 subclones. The Sequenase 2.0 version kit (USB) was used after alkali denaturation of the double-stranded templates as described in Wong *et al.* (1990).

MVR-PCR adapted to CEB1 and typing of individual alleles

Principle of MVR-PCR. MVR-PCR was developed by Jeffreys and colleagues (Jeffreys *et al.*, 1991) to take advantage of the existence of variant repeats within minisatellite alleles (Jeffreys *et al.*, 1990) in order to generate internal maps of minisatellite alleles by a simple PCR assay. The procedure provides maps of up to >2 kb of tandem array. To develop this assay for a new locus, flanking sequence data as well as characterized internal variants are needed. Three primers are used to determine the status of each motif along an allele for one variant. Two of them are non-specific and are used for each reaction: a flanking primer and the *TAG*. The status is determined by using a primer made of part of the minisatellite repeat sequence followed by a non-minisatellite extension identical to the *TAG* primer. This primer, used at a low concentration, generates a set of amplification products starting from the adequate motifs and extending in the flanking sequence. In the following PCR cycles, amplification will occur essentially between the flanking primer (P1) and the *TAG*, both present at normal concentration. The amplification products are then separated on an agarose gel, transferred by Southern blotting on a nylon membrane and hybridized to the minisatellite probe.

Setting up the MVR-PCR assay for CEB1. Three evenly spaced nucleotidic variations (v4, v17 and v28; Figure 1) were chosen for their informativity and their relative location along the repeat to type CEB1 alleles in MVR-PCR experiments. They provide partial sequence information every 13 nucleotides within each motif. Figure 1 indicates the sets of primers used. Nucleotides can be A or G at v4, C or T at v17, and C or T at v28. For each variation, two sets of 17 base long primers differing only at their last 3′ base (or close to it, e.g. see v28) were designed (Figure 1). For instance, the two primers v17-C and v17-T are used to assay polymorphism v17. In addition to the flanking primers P1 and to the *TAG* primer, six different primers are necessary to type the three variant sites v4, v17 and v28. In practice, things are complicated by interference from neighbouring variants. For this reason, v4 is typed using a mixture of primers to neutralize the interference from the variant in position 11 (Figure 1). v28 is typed by two primers extending 3 nucleotides 3′ from the variant site. Primers ending up 3′ at the v28 variant position were detecting too many 'null' motifs, presumably because of an as yet unidentified variant nearby in 5′ (data not shown). Consequently, CCCT insertions 3′ of v28 will contribute some of the 'null' motifs detected by these initial primers. The same 20 base long *TAG* sequence was added at the 5′ end of all these 17-mers. The 37-mer oligonucleotides were synthesized with an Applied Biosystems 381A synthesizer. The *TAG* primer used is the same as in Jeffreys *et al.* (1991).

Isolation of individual alleles. Because the variations between motifs can be of the insertion–deletion type, the MVR typing of diploid DNA samples is often impossible to interpret (the two ladders do not superpose). For this reason, alleles were amplified, separated and typed individually. This has the advantage of clarity and removes ambiguities linked to *a posteriori* haplotype reconstructions in mutant alleles. P1 and P6 are the flanking primers used to amplify the alleles. Genomic DNA (200 ng) was first amplified with the flanking amplimers P1 and P6 at a final concentration of 1 μM each in a total volume of 15 μl using the buffer described in Jeffreys *et al.* (1990). Reactions were cycled for 10 s at 96°C, 15 s at 68°C and 6 min at 70°C for 25 cycles, ended by 15 s at 68°C and 10 min at 70°C for two cycles in a Perkin-Elmer GeneAmp PCR System 9600. One-tenth of the amplification product was electrophoresed, denatured by soaking the gel in 0.4 N NaOH, 1.5 M NaCl for 2 × 20 min, transferred by capillary blotting onto a Hybond-N^+ (Amersham) membrane and hybridized with the ^{32}P-labelled (random priming; Feinberg and Vogelstein, 1984) minisatellite probe to estimate the size and quantities of the two alleles on the autoradiogram. The remaining (9/10) product was then electrophoresed and each allele was recovered from the 1% agarose (Seakem LE, FMC Bioproducts) gel as described in Heery *et al.* (1990). The elution product was diluted in water to a concentration of 0.1–0.5 pg/μl; 1 μl of this dilution of the amplified allele was then submitted to MVR-PCR.

MVR-PCR typing. In MVR-PCR typing experiments primers P1 and TAG were used with either primers v4-G(11G) and v4-G(11A) or v4-A(11G) and v4-A(11A) or v-17C or v-17T or v-28C or v-28T at the final concentration indicated in Figure 1. Buffer and cycling conditions were the same as described for amplifying the entire alleles, except that the number of cycles was reduced to 20. MVR-PCR products were electrophoresed through a 40 cm long 1% agarose (Seakem LE or I.D.NA from FMC bioproducts) gel in Tris–acetate EDTA (TAE 1 × buffer; Sambrook *et al.*, 1989), with buffer recirculation and cooling, at 10 V/cm until the bromophenol blue has run 33 cm. DNA was denatured in the gel, transferred and hybridized to the CEB1 probe as described above. Filters were washed twice for 5 min in 1 × SSC (Sambrook *et al.*, 1989), 0.1% SDS at 65°C. Autoradiography was for 16 h at room temperature.

Acknowledgements

We wish to thank Dr Gilles Thomas, in whose laboratory most of this work was done, for his interest and support, Professor Jean Dausset and Dr Howard Cann for supplying the CEPH DNA samples, and Dr Agnès Marcadet for making a set of Southern blots for the extended CEPH families panel. We thank Dr Christine Pourcel for her criticisms during the preparation of the manuscript and Professor Alec Jeffreys for comments on an early version. We thank Sylvaine Monfouilloux and Patrick Foley for comments on a late version. This work is supported by the Eurogem project and by a grant from the French genome project (GREG).

References

Aaltonen,L.A., Peltomäki,P., Leach,F.S., Sistonen,P., Pylkkänen,L., Mecklin,J.-P., Järvinen,H., Powell,S.M., Jen,J., Hamilton,S.R.,

Petersen,G.M., Kinzler,K.W., Vogelstein,B. and de la Chapelle,A. (1993) *Science*, **260**, 812–816.
Agurell,E., Cederberg,H., Hedenskog,M. and Rannug,U. (1994) *Mol. Gen. Genet.*, **242**, 137–144.
Armour,J.A.L. and Jeffreys,A.J. (1992) *Curr. Opin. Genet. Dev.*, **2**, 850–856.
Armour,J.A.L., Harris,P.C. and Jeffreys,A.J. (1993) *Hum. Mol. Genet.*, **2**, 1137–1145.
Dubrova,Y.E., Jeffreys,A.J. and Malashenko,A.M. (1993) *Nature Genet.*, **5**, 92–94.
Feinberg,A.P. and Vogelstein,B. (1984) *Anal. Biochem.*, **137**, 266–267.
Fishel,R., Lescoe,M.K., Rao,M.R.S., Copeland,N.G., Jenkins,N.A., Garber,J., Kane,M. and Kolodner,R. (1993) *Cell*, **75**, 1027–1038.
Gibbs,M., Collick,A., Kelly,R.G. and Jeffreys,A.J. (1993) *Genomics*, **17**, 121–128.
Groden,J. and German,J. (1992) *Hum. Genet.*, **90**, 360–367.
Heery,D.M., Gannon,F. and Powell,R. (1990) *Trends Genet.*, **6**, 173.
Imbert,G., Kretz,C., Johnson,K. and Mandel,J.-L. (1993) *Nature Genet.*, **4**, 72–76.
Ionov,Y., Peinado,M.A., Malkhosyan,S., Shibata,D. and Perucho,M. (1993) *Nature*, **363**, 558–561.
Jeffreys,A.J., Royle,N.J., Wilson,V. and Wong,Z. (1988) *Nature*, **332**, 278–281.
Jeffreys,A.J., Neumann,R. and Wilson,V. (1990) *Cell*, **60**, 473–485.
Jeffreys,A.J., MacLeod,A., Tamaki,K., Neil,D.L. and Monckton,D.G. (1991) *Nature*, **354**, 204–209.
Jeffreys,A.J., Tamaki,K., MacLeod,A., Monckton,D.G., Neil,D.L. and Armour,J.A.L. (1994) *Nature Genet.*, **6**, 136–145.
Kelly,R., Bulfield,G., Collick,A., Gibbs,M. and Jeffreys,A.J. (1989) *Genomics*, **5**, 844–856.
Kelly,R., Gibbs,M., Collick,A. and Jeffreys,A.J. (1991) *Proc. R. Soc. London Ser. B*, **245**, 235–245.
Krontiris,T.G., Devlin,B., Karp,D.D., Robert,N.J. and Risch,N. (1993) *N. Engl. J. Med.*, **329**, 517–523.
Leach,F.S. *et al.* (1993) *Cell*, **75**, 1215–1225.
Lucassen,A.M., Julier,C., Beressi,J.-P., Boitard,C., Froguel,P., Lathrop,M. and Bell,J.I. (1993) *Nature Genet.*, **4**, 305–310.
Mézard,C. and Nicolas,A. (1994) *Mol. Cell. Biol.*, **14**, 1278–1292.
Neil,D.L. and Jeffreys,A.J. (1993) *Hum. Mol. Genet.*, **2**, 1129–1135.
Nicholls,R.D. (1994) *Hum. Mol. Genet.*, **3**, 9–11.
NIH/CEPH Collaborative Mapping Group (1992) *Science*, **258**, 67–83.
Ohno,S. (1985) *Trends Genet.*, **1**, 196–200.
Parker,B.O. and Marinus,M.G. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 1730–1734.
Parsons,R. *et al.* (1993) *Cell*, **75**, 1227–1236.
Reyniers,E. *et al.* (1993) *Nature Genet.*, **4**, 143–146.
Rouyer,F., Simmler,M.C., Page,D.C. and Weissenbach,J. (1987) *Cell*, **51**, 417–425.
Royle,N.J., Clarkson,R.E., Wong,Z. and Jeffreys,A.J. (1988) *Genomics*, **3**, 352–360.
Sambrook,J., Fritsch,E.F. and Maniatis,T. (1989) *Molecular Cloning: A Laboratory Manual.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
Simon,M., Phillips,M. and Green,H. (1991) *Genomics*, **9**, 576–580.
Strand,M., Prolla,T.A., Liskay,R.M. and Petes,T.D. (1993) *Nature*, **365**, 274–276.
Szostak,J.W., Orr-Weaver,T.L., Rothstein,R.J. and Stahl,F.W. (1983) *Cell*, **33**, 25–35.
Tamaki,K., Monckton,D.G., MacLeod,A., Neil,D.L., Allen,M. and Jeffreys,A.J. (1992) *Hum. Mol. Genet.*, **1**, 401–406.
Tamaki,K., Monckton,D.G., MacLeod,A., Allen,M. and Jeffreys,A.J. (1993) *Hum. Mol. Genet.*, **2**, 1629–1632.
Thibodeau,S.N., Bren,G. and Schaid,D. (1993) *Science*, **260**, 816–819.
Vergnaud,G., Mariat,D., Apiou,F., Aurias,A., Lathrop,M. and Lauthier,V. (1991) *Genomics*, **11**, 135–144.
Vergnaud,G., Gauguier,D., Schott,J.-J., Lepetit,D., Lauthier,V., Mariat,D. and Buard,J. (1993) *DNA Fingerprinting: State of the Science.* Birkhäuser Verlag, Basel, Switzerland, pp. 47–57.
Weber,J.L. and Wong,C. (1993) *Hum. Mol. Genet.*, **2**, 1123–1128.
Weissenbach,J., Gyapay,G., Dib,C., Vignal,A., Morissette,J., Millasseau,P., Vaysseix, G. and Lathrop, M. (1992) *Nature*, **359**, 794–801.
Wieringa,B. (1994) *Hum. Mol. Genet.*, **3**, 1–7.
Wong,W.M., Au,D.M.Y., Lam,V.M.S., Tam,J.W.O. and Cheng,L.Y.L. (1990) *Nucleic Acids Res.*, **18**, 5573.

Received on March 8, 1994; revised on April 22, 1994