The EMBO Journal Vol.17 No.12 pp.3495–3502, 1998

Influences of array size and homogeneity on minisatellite mutation

Jérôme Buard[1,2], Agnès Bourdet[1], Jane Yardley[1], Yuri Dubrova[1,3] and Alec J.Jeffreys[1]

[1]Department of Genetics, University of Leicester, Leicester, LE1 7RH, UK and [3]N.I. Vavilov Institute of General Genetics, Russian Academy of Sciences, Moscow, Russia

[2]Corresponding author
e-mail: jtrb1@le.ac.uk

Unstable minisatellites display high frequencies of spontaneous gain and loss of repeats in the human germline. Most length changes arise through complex recombination events including intra-allelic duplications/deletions and inter-allelic transfers of repeats. Definition of the factors modulating instability requires both measurement of mutation rate and detailed analysis of mutant structures at the level of individual alleles. We have measured mutation rates in sperm for a wide range of alleles of the highly unstable human minisatellite CEB1. Instability varies by three orders of magnitude between alleles and increases steadily with the size of the tandem array. Structural analysis of mutant molecules derived from six alleles revealed that it is the rate of intra-allelic rearrangements which increases with array size and that intra-allelic duplication events tend to cluster within homogeneous segments of alleles; both phenomena resemble features of trinucleotide repeat instability. In contrast, inter-allelic transfers occur at a fairly constant rate, irrespective of array length, and show a mild polarity towards one end of the minisatellite, suggesting the possible influence of flanking DNA on these conversion-like events.

Keywords: instability/minisatellite/mutation/sperm/
tandem repeat

Introduction

Hypervariable minisatellites provide the most informative tools for unravelling some mechanisms of tandem repeat turnover. First, the germline-specific instability which drives this hypervariability can produce new length minisatellite alleles at high rates and with a marked bias towards expansions (Jeffreys *et al.*, 1988; Vergnaud *et al.*, 1991). *De novo* mutants can be identified both in pedigrees and by small-pool PCR (SP-PCR) amplification of minisatellite arrays from sperm DNA (Jeffreys *et al.*, 1994; May *et al.*, 1996). Secondly, base composition variation between repeats, an apparently universal property of human minisatellites, makes it possible to gain insights into the mutation process by mapping the succession of variant repeats along the tandem array (minisatellite variant repeat mapping using PCR, MVR-PCR) before and after mutation (Jeffreys *et al.*, 1991). Complex rearrangements, both intra-allelic duplications and polarized inter-allelic transfers of repeats, account for the vast majority of germline expansions at the four unstable GC-rich minisatellites studied to date (Buard and Vergnaud, 1994; Jeffreys *et al.*, 1994; May *et al.*, 1996).

CEB1 is the most unstable human minisatellite yet isolated, with new length alleles being produced almost exclusively by males at an average rate, estimated from pedigree data, of 13% per sperm (Vergnaud *et al.*, 1991). The average mutation rate together with the frequency of intra-allelic events at CEB1 are far more elevated than at the three other minisatellite loci studied MS32, MS31A and MS205. Half of the CEB1 repeat arrays in Caucasians span less than 3 kb and are therefore potentially amenable to SP-PCR analysis. Eight different base variations exist between CEB1 repeats and form the basis for the most discriminatory MVR-PCR system developed to date. This high resolution has allowed us to propose a model for minisatellite instability in which staggered nicks initiate a double-strand break (DSB) in the tandem array (Buard and Vergnaud, 1994).

The definition of factors modulating minisatellite instability is key to our achieving a better understanding of their mutation process. Pedigree mutant data from minisatellite g3 suggest that the mutation rate is affected by flanking nucleotide variation and by the size of the tandem array itself, with higher mutation rates for larger alleles (Andreassen *et al.*, 1996). However, lack of detailed information on instability of individual alleles and on the mutation process of the g3 minisatellite make it impossible to define precisely these structural modifiers of instability. Better progress has been made at minisatellites MS32 and MS205, where small-pool PCR analysis has revealed heterogeneities between sperm mutation rates of individual alleles (Jeffreys *et al.*, 1994; May *et al.*, 1996). A nucleotide transversion in the flanking vicinity of minisatellite MS32 is associated with a 110-fold suppression of mutation rate (Monckton *et al.*, 1994). However, no other structural factors modulating instability have been identified, although 10-fold variations have been observed between mutation rates of unstable MS32 (Jeffreys *et al.*, 1994) and MS205 (May *et al.*, 1996) alleles.

To analyse structural features in minisatellite alleles that can modulate instability, we have developed SP-PCR at CEB1, measured the sperm mutation rates of numerous alleles of different length and investigated the structural basis of mutation as a function of array length and composition.

Results

Allele-specific variation in germline instability

The mutation rate for each of 58 alleles was estimated by SP-PCR analysis of diluted sperm DNA from 38 unrelated

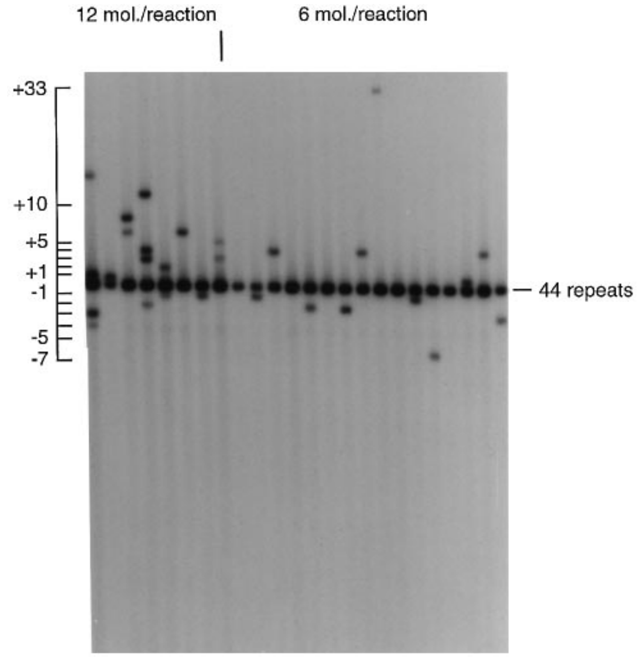


Fig. 1. An example of allele-specific SP-PCR at CEB1. A 44-repeat CEB1 allele (Figure 4, allele F) and mutant molecules were amplified from multiple aliquots of sperm DNA, eight containing 12 molecules on average and 16 containing six molecules on average. Most mutant molecules derived from this allele have gained or lost fewer than six repeats.

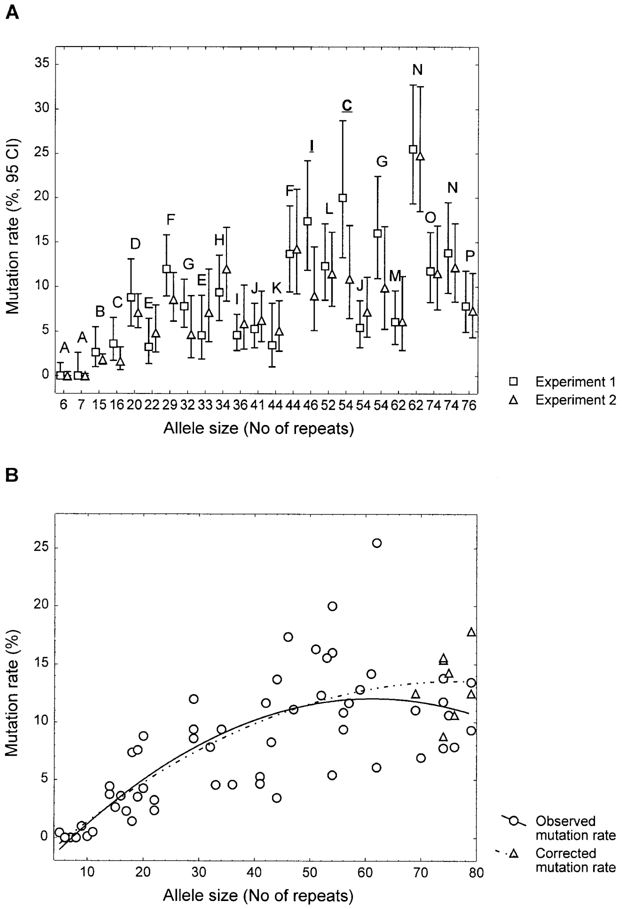


Fig. 2. Sperm mutation rates of CEB1 alleles. (**A**) Reproducibility of SP-PCR estimation of CEB1 mutation rates. For each of 24 alleles from 16 individuals (A–P) and ranging in size from six to 76 repeats, two independent estimates of mutation rates have been obtained, with several weeks between the two experiments. The 95% confidence interval (CI) is indicated for each value. In 22 cases, the two estimates of the mutation rate are concordant, the CI associated with one value overlapping the other value. Two alleles with statistically significant difference between the two estimates (I-46, $P = 0.0146$ and C-54, $P = 0.0204$) are underlined. The probability of encountering at least two significant differences of this magnitude out of 24 independent tests is 0.3392. (**B**) Mutation rate and allele size. Mutation rates were estimated for 58 different CEB1 alleles, using SP-PCR or allele specific SP-PCR. Quadratic fit for experimental (solid line), $y = -0.7077 + 0.7284\times\text{Size} - 0.0062\times\text{Size}^2$, $F(2/55) = 68.33$; $P = 1.22\times10^{-13}$ for the arcsine transformed values of mutation rates. For nine alleles containing 69–79 repeats (circles), +1 and –1 repeat mutants could not be reliably scored and their frequencies were therefore extrapolated from the size distribution of mutants for alleles 30–60 repeats long. New quadratic fit (dashed line), $y = -0.0852 + 0.6536\times\text{Size} - 0.0050\times\text{Size}^2$, $F(2/55) = 71.76$; $P = 9.99\times10^{-14}$.

individuals (two Asian, 18 Caucasian and 18 African) by assuming that each mutant molecule was derived from the progenitor allele closer in size, as seen for CEB1 mutations detected in pedigrees (Vergnaud *et al.*, 1991). This assumption was confirmed by allele-specific SP-PCR for 12 of the 58 alleles, ranging in size from six to 76 repeats, using flanking base substitutional heterozygosities (see Materials and methods for sequence of primers) to selectively amplify one or other allele (Monckton *et al.*, 1993). For 18 sperm donors, only one allele was analysed. In 12 cases, only the shorter allele could be examined because gains and losses of one and two repeats could not be reliably scored for the larger allele (>80 repeats). For the six remaining individuals, analysis was focused on only one allele (six short alleles) by allele-specific SP-PCR. As observed in pedigrees, 80% of CEB1 sperm mutants detected by SP-PCR involve the gain or loss of less than six repeats with a marked bias towards expansions (Figure 1).

SP-PCR estimation of mutation rates was carried out in duplicate for 24 alleles (Figure 2A), with several weeks between the two experiments. In 22 cases (>90%), the two independently estimated values of the mutation rate were concordant, indicating that estimation of CEB1 mutation rates by SP-PCR analysis is robust.

SP-PCR analysis of mutation rate was carried out in 21 individuals for which sperm mutant molecules could be scored for both alleles. The mutation rate per individual varies from <0.1% to 13% between these 21 sperm donors (data not shown). The frequency of mutations scored for each of two alleles in each individual did not correlate ($r = 0.2849$; $P = 0.237$). This independence strongly suggests that CEB1 instability is largely allele-specific and that factors intrinsic to the repeat array or its flanking DNA might modulate the mutation rate.

Effect of array size on CEB1 instability

The relationship between array length and mutation rate for the full set of 58 alleles is shown in Figure 2B. Rates vary by at least three orders of magnitude, from <0.02% to >20%. Instability appears to increase steadily with the size of the tandem array up to 60 repeats and plateaus above this size. This strong relationship indicates that the size of the CEB1 repeat array is the major factor influencing instability. However, mutation rates can vary up to 4- to 5-fold for alleles of the same size (i.e. 44, 54 and

Allele	ori.	no. rpts	μ%	–72	–4	MVR map	+256	+384
1	As.	5	0.42	G	A	USBU	G	G
2	Af.	6	<0.05	A	A	VGVQE	G	A
3	Af.	6	<0.05	A	A	PPOGP	G	A
4	Af.	7	<0.05	G	A	UBPBPP	G	G
5	Af.	8	0.1	G	A	UBPPBPU	G	G
6	C.	9	1.0	G	A	TBIIOQE	A	G
7	Af.	10	0.1	G	A	OGBGSOSG	G	A
8	Af.	11	0.5	G	A	OOOPGOPQF	G	A
9	C.	14	3.7	G	A	GQBMMGLGMBSG	G	G
10	C.	14	4.4	G	A	KSBBIBSBBBBP	A	G

Fig. 3. Structural relationships between 10 short CEB1 alleles. MVR maps and flanking haplotypes defined by two 5′ flanking nucleotide variations (positions –72 and –4) and two 3′ ones (positions +256 and +384) were determined for the 10 short alleles. The number of repeats and the mutation rate (μ%) of each allele are indicated together with the number of repeats of the partner allele and the ethnic origin (ori.: As, Asian; Af, African; C, Caucasian) of the sperm donors. Only two alleles (4 and 5) are clearly related.

62 repeat alleles) implying that factors other than size can influence the level of instability. Below 20 repeats, the mutation rate decreases dramatically to <0.02% below nine repeats, but even for five to 11 repeat alleles, considerable variation of instability is observed. Indeed, two alleles of five and nine repeats show 0.45 and 1.05% mutation rates, respectively, representing a 10- to 20-fold enhancement over the six other short alleles of this size class. These two alleles are Caucasian and Asian, respectively, in origin, whereas the other six are African. Typing of these short alleles by MVR-PCR and for four base substitutional polymorphisms near the repeat array (Figure 3) showed that while two African alleles (numbers 4 and 5) are structurally very similar, the remaining alleles have highly divergent internal structures and occur on different haplotypic backgrounds. This suggests that germline stability at short alleles is most likely a direct consequence of short array length.

A variable rate of intra-allelic duplication

To investigate the structural basis of this variability of instability at CEB1, 138 mutant molecules derived from six different alleles (A–F; 9, 14, 18, 29, 29 and 44 repeats long, respectively) were recovered by size enrichment SP-PCR (SESP-PCR) (Jeffreys and Neuman, 1997) and their internal structure determined by MVR-PCR (Figure 4).

Mutant allele structures could be divided into three broad classes. The first involves intra-allelic duplications and deletions of repeat blocks, sometimes complex but with no evidence of repeat transfer between alleles (e.g. mutants D3 and D32). The second class shows clear evidence for inter-allelic transfer of repeats (e.g. mutant A12). The third class, including 30 of 138 mapped mutants, involves rearrangements so complex that interpretation is not possible; for example, mutant B2 consists of the whole of allele B into which a 12-repeat block has been inserted (BOBBBBBBGBGB) present in neither progenitor allele but including individual O- and G-type repeats probably derived from the other allele.

Both intra-allelic rearrangements and inter-allelic transfers of blocks of repeats lead to size changes for each CEB1 allele but the relative proportion of each of these two kinds of rearrangement varies considerably between alleles. Figure 5 summarizes the frequencies for the different rearrangements for the six alleles analysed, with uninterpretable mutants omitted. For alleles of <18 repeats (alleles A, B and C), intra-allelic rearrangements are scarce. These rearrangements occur at much higher frequency in the larger alleles D–F.

MVR maps of gain mutants from allele D show that more than half (12 of 21) of the intra-allelic duplication events are centred on a string of six identical repeats (BBBBBB) within allele D (Figure 6A). Such a homogeneous stretch of repeats does not exist in allele E; however, four of six intra-allelic duplications in this allele have again occurred within a four-repeat portion formed by the duplication of two contiguous repeats. Similarly, three of 10 intra-allelic duplication events in allele F are clustered in a six-repeat block consisting of a three-repeat dimer. In contrast, no such phenomenon can be observed for the 14-repeat allele B, which harbours a string of four identical repeats.

For most of the 30 deletion events mapped, only one block of contiguous repeats has been lost. More complex events involving the loss of two separate blocks was observed in two instances. In contrast to duplications, for which each event is unique, we observed five cases of two independently recovered deletion events with identical MVR maps. One of these five pairs shows the loss of two distinct blocks of repeats.

A constant rate of inter-allelic transfer

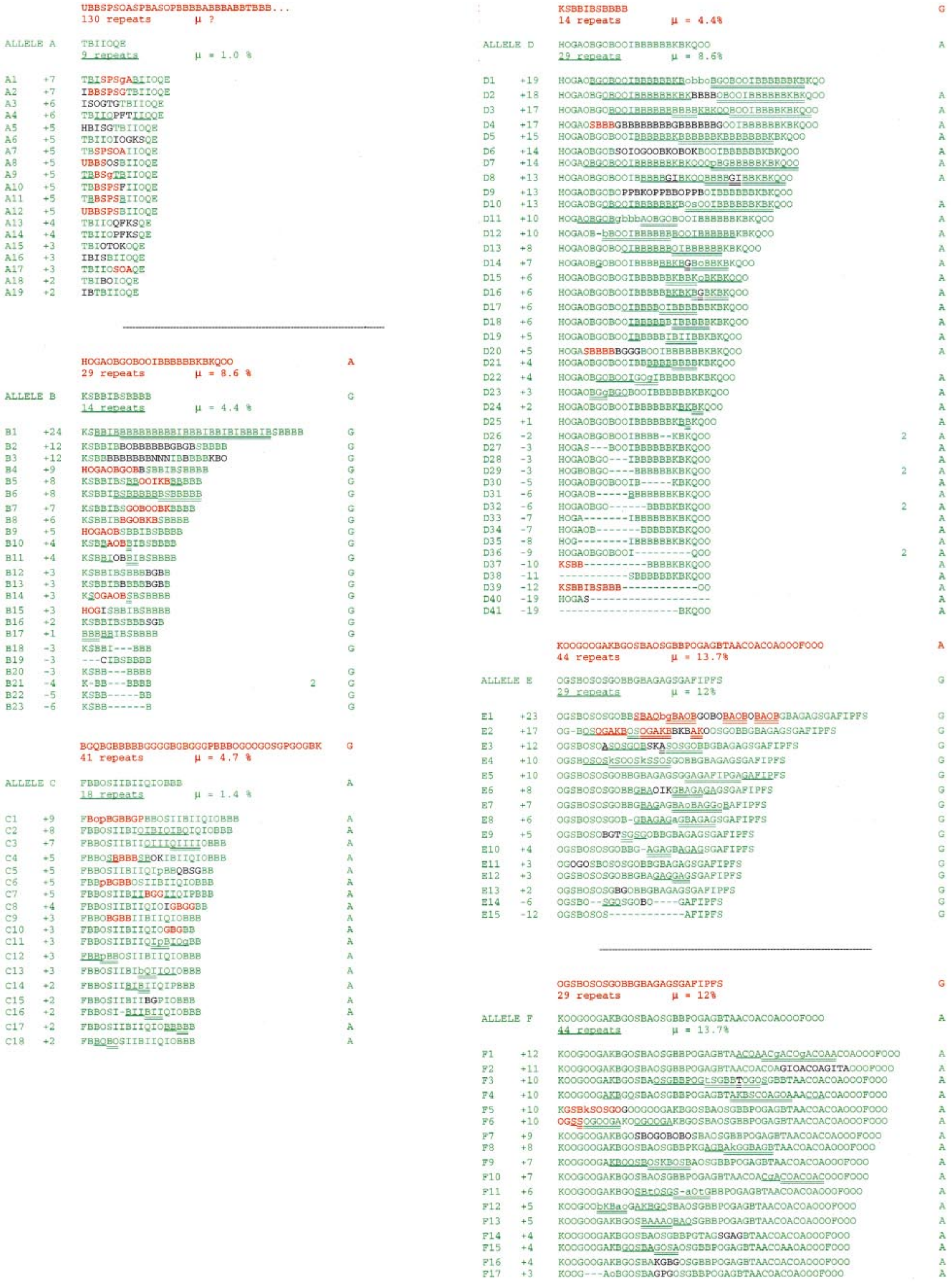
In contrast to the huge variation in rates of intra-allelic duplication/deletion (>100 fold), there is no significant difference between frequencies of inter-allelic transfer over the six alleles analysed. For instance, allele A shows a 1% rate of expansion and MVR maps indicate that all interpreted gains (nine out of nine) are inter-allelic insertions, implying an overall frequency of inter-allelic insertions of 1%. Allele D, with its 5% rate of gains, displays four inter-allelic insertions out of 25 expansions, suggesting an overall frequency of inter-allelic transfers of ~0.8%. Inter-allelic transfers can be extremely complex, with imperfect duplications of the inserted portion (e.g. mutant E1) or deletions within the insertion (e.g. mutant C1). The 32 inter-allelic events show seven cases of duplication of a group of repeats derived from the recipient allele at each side of the inserted group (e.g. mutant A1) and eight cases of deletion of repeats at the insertion site (e.g. mutant C9). For eight transfers, six of which are associated with a loss of repeats of the recipient allele at the junction, the beginning of the donor allele is fused with the end of the recipient allele.

Mild polarity of CEB1 inter-allelic transfers

The insertion site of the group of repeats from the donor allele in the recipient allele has been determined for 32 inter-allelic transfers (Figure 6B). The distribution of insertion sites at CEB1 is not random, with more than three quarters of the insertions occuring within the first half of the tandem array [$\chi^2 = 8.4$ (1 d.f.), $P = 0.004$]. However, this degree of polarity is significantly lower than that observed for minisatellite MS32 (Jeffreys *et al.*, 1991, 1994). (Kolmogorov–Smirnoff test, $P < 10^{-4}$), for which >90% of insertion sites are clustered within the first quarter of the tandem array (Figure 6B).

Discussion

Direct analysis of mutation events at minisatellite CEB1 in sperm shows that the male germline mutation rate varies by three orders of magnitude between alleles, and that structural features of the repeat array such as size and homogeneity play a major role in this variation. The sperm mutation rate range estimated for Caucasian CEB1 alleles



UBBSPSOASPBASOPBBBBABBBABBTBBB...
130 repeats μ ?
ALLELE A TBIIOQE
9 repeats μ = 1.0 %
A1 +7 TBISPSgABIIOQE
A2 +7 IBBSPSGTBIIOQE
A3 +6 ISOGTGTBIIOQE
A4 +6 TBIIOPFTIIOQE
A5 +5 HBISGTBIIOQE
A6 +5 TBIIOIOGKSQE
A7 +5 TBSPSOAIIOQE
A8 +5 UBBSOSBIIOQE
A9 +5 TBBSgTBIIOQE
A10 +5 TBBSPSFIIOQE
A11 +5 TBBSPSBIIOQE
A12 +5 UBBSPSBIIOQE
A13 +4 TBIIOQFKSQE
A14 +4 TBIIOPFKSQE
A15 +3 TBIOTOKOQE
A16 +3 IBISBIIOQE
A17 +3 TBIIOSOAQE
A18 +2 TBIBOIOQE
A19 +2 IBTBIIOQE
HOGAOBGOBOOIBBBBBBKBKQOO A
29 repeats μ = 8.6 %
ALLELE B KSBBIBSBBBB G
14 repeats μ = 4.4 %
B1 +24 KSBBIBBBBBBBBBIBBBIBBIBIBBBIBSBBBB G
B2 +12 KSBBIBBOBBBBBBGBGBSBBBB G
B3 +12 KSBBBBBBBBBNNNIBBBBBKBO G
B4 +9 HOGAOBGOBBSBBIBSBBBB G
B5 +8 KSBBIBSBBOOIKBBBBBB G
B6 +8 KSBBIBSBBBBBBSBBBBB G
B7 +7 KSBBIBSGOBOOBKBBBB G
B8 +6 KSBBIBBGOBKBSBBBB G
B9 +5 HOGAOBSBBIBSBBBB G
B10 +4 KSBBAOBBIBSBBBB G
B11 +4 KSBBIOBBIBSBBBB G
B12 +3 KSBBIBSBBBBGBB G
B13 +3 KSBBIBBBBBBGBB G
B14 +3 KSOGAOBSBSBBBB G
B15 +3 HOGISBBIBSBBBB G
B16 +2 KSBBIBSBBBSGB G
B17 +1 BBBBBIBSBBBB G
B18 -3 KSBBI---BBB G
B19 -3 ---CIBSBBBB
B20 -3 KSBB---BBBB G
B21 -4 K-BB---BBBB 2 G
B22 -5 KSBB-----BB G
B23 -6 KSBB------B G
BGQBGBBBBBGGGGBGBGGGPBBBOGOOGOSGPGOGBK G
41 repeats μ = 4.7 %
ALLELE C FBBOSIIBIIQIOBBB A
18 repeats μ = 1.4 %
C1 +9 FBopBGBBGPBBOSIIBIIQIOBBB A
C2 +8 FBBOSIIBIQIBIQIBQIQIOBBB A
C3 +7 FBBOSIIBIIQIIIQIIIIOBBB A
C4 +5 FBBOSBBBBSBOKIBIIQIOBBB A
C5 +5 FBBOSIIBIIQIpBBQBSGBB A
C6 +5 FBBpBGBBOSIIBIIQIOBBB A
C7 +5 FBBOSIIBIIBGGIIQIPBBB A
C8 +4 FBBOSIIBIIQIOIGBGGBB A
C9 +3 FBBOBGBBIIBIIQIOBBB A
C10 +3 FBBOSIIBIIQIOGBGBB A
C11 +3 FBBOSIIBIIQIpBIOgBB A
C12 +3 FBBpBBOSIIBIIQIOBBB A
C13 +3 FBBOSIIBIbQIIQIOBBB A
C14 +2 FBBOSIIBIBIIQIPBBB A
C15 +2 FBBOSIIBIIBGPIOBBB A
C16 +2 FBBOSI-BIIBIIQIOBBB A
C17 +2 FBBOSIIBIIQIOBBBBB A
C18 +2 FBBQBOSIIBIIQIOBBB A
KSBBIBSBBBB G
14 repeats μ = 4.4%
ALLELE D HOGAOBGOBOOIBBBBBBKBKQOO A
29 repeats μ = 8.6%
D1 +19 HOGAOBGOBOOIBBBBBBKBobboBGOBOOIBBBBBBKBKQO
D2 +18 HOGAOBGOBOOIBBBBBBKBKBBBBOBOOIBBBBBBKBKQOO A
D3 +17 HOGAOBGOBOOIBBBBBBBBBBKBKQOBOOIBBBBKBKQOO A
D4 +17 HOGAOSBBBGBBBBBBBBGBBBBBBGOOIBBBBBBKBKQOO A
D5 +15 HOGAOBGOBOOIBBBBBBKBBBBBBKBBBBBBBKBKQOO A
D6 +14 HOGAOBGOBSOIOGOOBKOBOKBOOIBBBBBBKBKQOO A
D7 +14 HOGAOBGOBOOIBBBBBBKBKQOOpBGBBBBBKBKQOO A
D8 +13 HOGAOBGOBOOIBBBBBGIBKQOBBBBGIBBKBKQOO A
D9 +13 HOGAOBGOBOPPBKOPPBBOPPBOIBBBBBBKBKQOO
D10 +13 HOGAOBGOBOOIBBBBBBKBOsOOIBBBBBBKBKQOO A
D11 +10 HOGAOBGOBgbbbAOBGOBOOIBBBBBBKBKQOO A
D12 +10 HOGAOB-bBOOIBBBBBBBOOIBBBBBBKBKQOO A
D13 +8 HOGAOBGOBOOIBBBBBBOIBBBBBBKBKQOO A
D14 +7 HOGAOBGOBOOIBBBBBBKBGBoBBKBKQOO A
D15 +6 HOGAOBGOBOGIBBBBBBKBBKoBKBKQOO A
D16 +6 HOGAOBGOBOOIBBBBBBKBKBGBKBKQOO A
D17 +6 HOGAOBGOBOOIBBBBOIBBBBBBKBKQOO A
D18 +6 HOGAOBGOBOOIBBBBBBIBBBBBKBKQOO A
D19 +5 HOGAOBGOBOOIBBBBBIBIIBBKBKQOO A
D20 +5 HOGASBBBBBGGGBOOIBBBBBBKBKQOO A
D21 +4 HOGAOBGOBOOIBBBBBBBBBBKBKQOO
D22 +4 HOGAOBGOBOOIGOgIBBBBBBKBKQOO A
D23 +3 HOGAOBGgBGOBOOIBBBBBBKBKQOO A
D24 +2 HOGAOBGOBOOIBBBBBBKBKBKQOO A
D25 +1 HOGAOBGOBOOIBBBBBBKBBKQOO A
D26 -2 HOGAOBGOBOOIBBBB--KBKQOO 2 A
D27 -3 HOGAS---BOOIBBBBBBKBKQOO A
D28 -3 HOGAOBGO---IBBBBBBKBKQOO A
D29 -3 HOGBOBGO----BBBBBBKBKQOO 2 A
D30 -5 HOGAOBGOBOOIB-----KBKQOO A
D31 -6 HOGAOB-----BBBBBBBKBKQOO A
D32 -6 HOGAOBGO------BBBBKBKQOO 2 A
D33 -7 HOGA-------IBBBBBBKBKQOO A
D34 -7 HOGAOB-------BBBBBKBKQOO A
D35 -8 HOG--------IBBBBBBKBKQOO A
D36 -9 HOGAOBGOBOOI---------QOO 2 A
D37 -10 KSBB----------BBBBKBKQOO A
D38 -11 -----------SBBBBBBKBKQOO A
D39 -12 KSBBIBSBBB-----------OO A
D40 -19 HOGAS------------------- A
D41 -19 -------------------BKQOO A
KOOGOOGAKBGOSBAOSGBBPOGAGBTAACOACOAOOOFOOO A
44 repeats μ = 13.7%
ALLELE E OGSBOSOSGOBBGBAGAGSGAFIPFS G
29 repeats μ = 12%
E1 +23 OGSBOSOSGOBBSBAObgBAOBGOBOBAOBOBAOBGBAGAGSGAFIPFS G
E2 +17 OG-BOSOGAKBOSOGAKBBKBAKOOSGOBBGBAGAGSGAFIPFS G
E3 +12 OGSBOSOASOSGOBSKASOSGOBBGBAGAGSGAFIPFS G
E4 +10 OGSBOSOSkSOOSkSSOSGOBBGBAGAGSGAFIPFS G
E5 +10 OGSBOSOSGOBBGBAGAGSGGAGAFIPGAGAFIPFS G
E6 +8 OGSBOSOSGOBBGBAOIKGBAGAGAGSGAFIPFS G
E7 +7 OGSBOSOSGOBBGBAGAGBAoBAGGoBAFIPFS G
E8 +6 OGSBOSOSGOB-GBAGAGaGBAGAGSGAFIPFS G
E9 +5 OGSBOSOBGTSGSGOBBGBAGAGSGAFIPFS G
E10 +4 OGSBOSOSGOBBG-AGAGBAGAGSGAFIPFS G
E11 +3 OGOGOSBOSOSGOBBGBAGAGSGAFIPFS G
E12 +3 OGSBOSOSGOBBGBAGAGGAGSGAFIPFS G
E13 +2 OGSBOSOSGBGOBBGBAGAGSGAFIPFS G
E14 -6 OGSBO--SGOSGOBO----GAFIPFS G
E15 -12 OGSBOSOS------------AFIPFS G
OGSBOSOSGOBBGBAGAGSGAFIPFS G
29 repeats μ = 12%
ALLELE F KOOGOOGAKBGOSBAOSGBBPOGAGBTAACOACOAOOOFOOO A
44 repeats μ = 13.7%
F1 +12 KOOGOOGAKBGOSBAOSGBBPOGAGBTAACOAACgACOgACOAACOAOOOFOOO A
F2 +11 KOOGOOGAKBGOSBAOSGBBPOGAGBTAACOACOAGIOACOAGITAOOOFOOO A
F3 +10 KOOGOOGAKBGOSBAOSGBBPOGtSGBBTOGOSGBBTAACOACOAOOOFOOO A
F4 +10 KOOGOOGAKBGOSBAOSGBBPOGAGBTAKBSCOAGOAAACOACOAOOOFOOO A
F5 +10 KGSBkSOSGOGOOGOOGAKBGOSBAOSGBBPOGAGBTAACOACOAOOOFOOO A
F6 +10 OGSSOGOOGAKOOGOOGAKBGOSBAOSGBBPOGAGBTAACOACOAOOOFOOO A
F7 +9 KOOGOOGAKBGOSBOGOBOBOSBAOSGBBPOGAGBTAACOACOAOOOFOOO A
F8 +8 KOOGOOGAKBGOSBAOSGBBPKGAGBAkGGBAGBTAACOACOAOOOFOOO A
F9 +7 KOOGOOGAKBOOSBOSKBOSBAOSGBBPOGAGBTAACOACOAOOOFOOO A
F10 +7 KOOGOOGAKBGOSBAOSGBBPOGAGBTAACOACgACOACOACOOOFOOO A
F11 +6 KOOGOOGAKBGOSBtOSGS-aOtGBBPOGAGBTAACOACOAOOOFOOO A
F12 +5 KOOGOObKBaoGAKBGOSBAOSGBBPOGAGBTAACOACOAOOOFOOO A
F13 +5 KOOGOOGAKBGOSBAAAOBAOSGBBPOGAGBTAACOACOAOOOFOOO A
F14 +4 KOOGOOGAKBGOSBAOSGBBPGTAGSGAGBTAACOACOAOOOFOOO A
F15 +4 KOOGOOGAKBGOSBAGOSAOSGBBPOGAGBTAACOAAOAOOOFOOO A
F16 +4 KOOGOOGAKBGOSBAKGBGOSGBBPOGAGBTAACOACOAOOOFOOO A
F17 +3 KOOG---AoBGOSBAGPGOSGBBPOGAGBTAACOACOAOOOFOOO A

(0.05–25%, average 9.3%) is in reasonable agreement, albeit slightly lower, with the 13% average male mutation rate estimated previously from CEPH pedigrees (Vergnaud *et al.*, 1991). This slight underestimation might be due to the over-representation of short alleles in this study, and also to the technical difficulty of analysing the largest CEB1 alleles by SP-PCR; mutations at such alleles can be detected in pedigrees and might have the highest mutation rates if the general trend of higher instability for larger alleles observed (Figure 2B) is extrapolated.

The SP-PCR derived mutation rate in blood DNA is <0.04% (~2000 progenitor molecules were analysed for each of four different alleles and no mutants were observed; data not shown). This indicates that CEB1 instability is germline-specific and further that potential SP-PCR artefacts do not interfere significantly with the detection of mutants in sperm.

CEB1 instability increases steadily with the size of the tandem array up to 60 repeats. Above this size, instability does not appear to increase but the observed plateau could be due to the technical difficulty of amplifying and scoring mutants derived from large alleles. Such a relationship between size and instability has also been observed at minisatellite B6.7 (Jeffreys *et al.*, 1997; Tamaki,K. and Jeffreys,A.J., in preparation). Minisatellite MS32 mutates by a predominantly inter-allelic gene conversion-like process and shows no trend towards higher mutation rates of longer alleles (Jeffreys *et al.*, 1994); this is consistent with the fairly constant rate of inter-allelic transfers at CEB1 alleles of different lengths. In contrast to MS32, CEB1 alleles do show a trend towards higher mutation rates of longer alleles, caused largely if not completely by an increase in the frequency of intra-allelic rearrangements. The observation that such mutations tend to cluster at relatively homogeneous regions of the array, with this clustering increasing possibly with the degree of homogeneity of such regions, is reminiscent of the requirement of homogeneity for instability of trinucleotide repeat arrays implicated in several neurodegenerative inherited diseases and fragile sites (Eichler *et al.*, 1994; Andrew *et al.*, 1997).

However, trinucleotide repeat and minisatellite instability may have different mechanistic bases. Several indirect lines of evidence support this hypothesis (Buard and Jeffreys, 1997). Somatic instability is significant for trinucleotide repeats (Monckton *et al.*, 1995), whereas minisatellites mutate almost exclusively in the germline (Jeffreys and Neuman, 1997); this germline-specificity includes intra-allelic duplications and deletions at CEB1. Secondly, the strand asymmetry which characterizes most GC-rich minisatellites formally excludes the formation of hairpin-like structures, formed by $(CNG)_n$ single strands and thought to play a key role in trinucleotide expansions (Gacy *et al.*, 1995; Mitas, 1997). Thirdly, although CEB1 array homogeneity can promote instability, a highly heterogeneous CEB1 allele can show a rate of intra-allelic rearrangement above 5% (Figure 5, alleles E and F) which contrasts with the strong requirement for perfect homogeneity for trinucleotide repeat instability (Chung *et al.*, 1993; Eichler *et al.*, 1994; Warren, 1996; Andrew *et al.*, 1997). Taken together, these observations strongly suggest that intra-allelic minisatellite mutation is not driven by a replication-based mechanism as proposed for

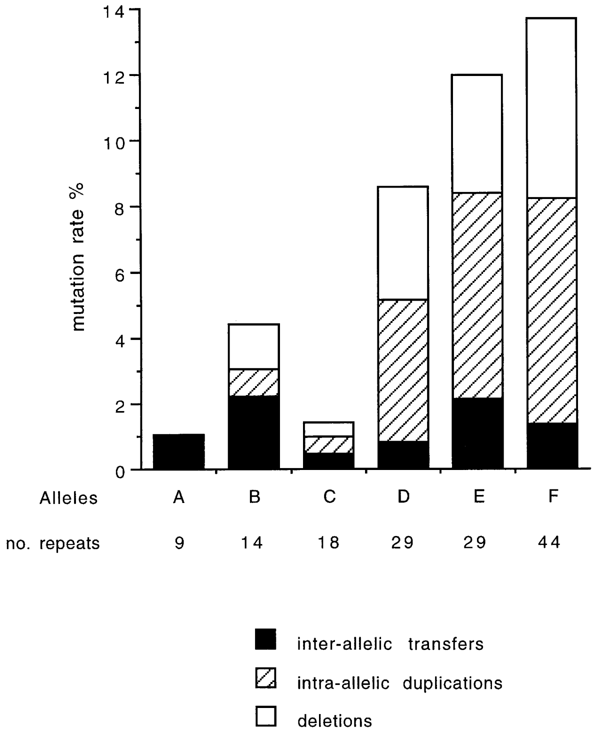


Fig. 5. Rate of inter-allelic transfers and intra-allelic duplication/deletion for six CEB1 alleles. Mutation rates estimated by SP-PCR of the six alleles analysed by MVR-PCR have been further subdivided into deletion rate, intra-allelic duplication rate and inter-allelic transfer rates. Test for homogeneity of rearrangement rates derived from the Poisson distribution: inter-allelic transfers, $\chi^2 = 2.52$ (5 d.f.), $P = 0.7735$; intra-allelic duplications, $\chi^2 = 86.25$ (5 d.f.), $P << 0.0001$; deletions, $\chi^2 = 111.48$ (5 d.f.), $P << 0.0001$.

Fig. 4. Structure of CEB1 mutant alleles determined by MVR analysis. 138 mutant molecules derived from six alleles (A–F) were mapped by MVR-PCR. For each semen donor, the MVR maps of the two progenitor alleles are shown together with their total number of repeats and their mutation rates (μ), and are aligned with the MVR maps of the mutant molecules. Each mutant is shown together with its identification number, the number of repeats gained or lost (e.g. gain of seven repeats for mutant A1), the status of one 3′ flanking nucleotide variation (variation +384 A/G for allele B and variation +256 A/G for alleles C–F, no heterozygous site was available for allele A) and the number of independent mutant molecules displaying the same MVR map (e.g. structure B21 has been found for two different mutant molecules); all maps without a number are unique. Supernumerary blocks of repeats identical to a nearby progenitor block are interpreted as intra-allelic duplication events and are underlined (single and double). Repeats lost in deletion events are represented by dashes. Supernumerary blocks of repeats identical to blocks present in the partner allele and interpreted as inter-allelic transfers are represented in red (e.g. mutant A1). Lowercase is used to represent a repeat which is likely to derive, through a single change (among the three variations used for CEB1 MVR mapping), from the corresponding repeat within the original red or green block (e.g. mutant A1). A rearrangement has been interpreted as an inter-allelic event only if the supernumerary block contains a block of at least three contiguous repeats identical to a three-repeat block within the donor (red) allele. A second criterion, which is a strong rule for all unstable minisatellites analysed, is that the transfer tends to occur in-register between donor and recipient alleles (Buard and Vergnaud, 1994; Jeffreys *et al.*, 1994; May *et al.*, 1996). For example, the red group in mutant A1 is inserted after the third repeat of allele A and corresponds to a block beginning at repeat number 4 in the donor allele. A number of rearrangements remain completely or partially uninterpreted and the corresponding repeats are indicated in black.

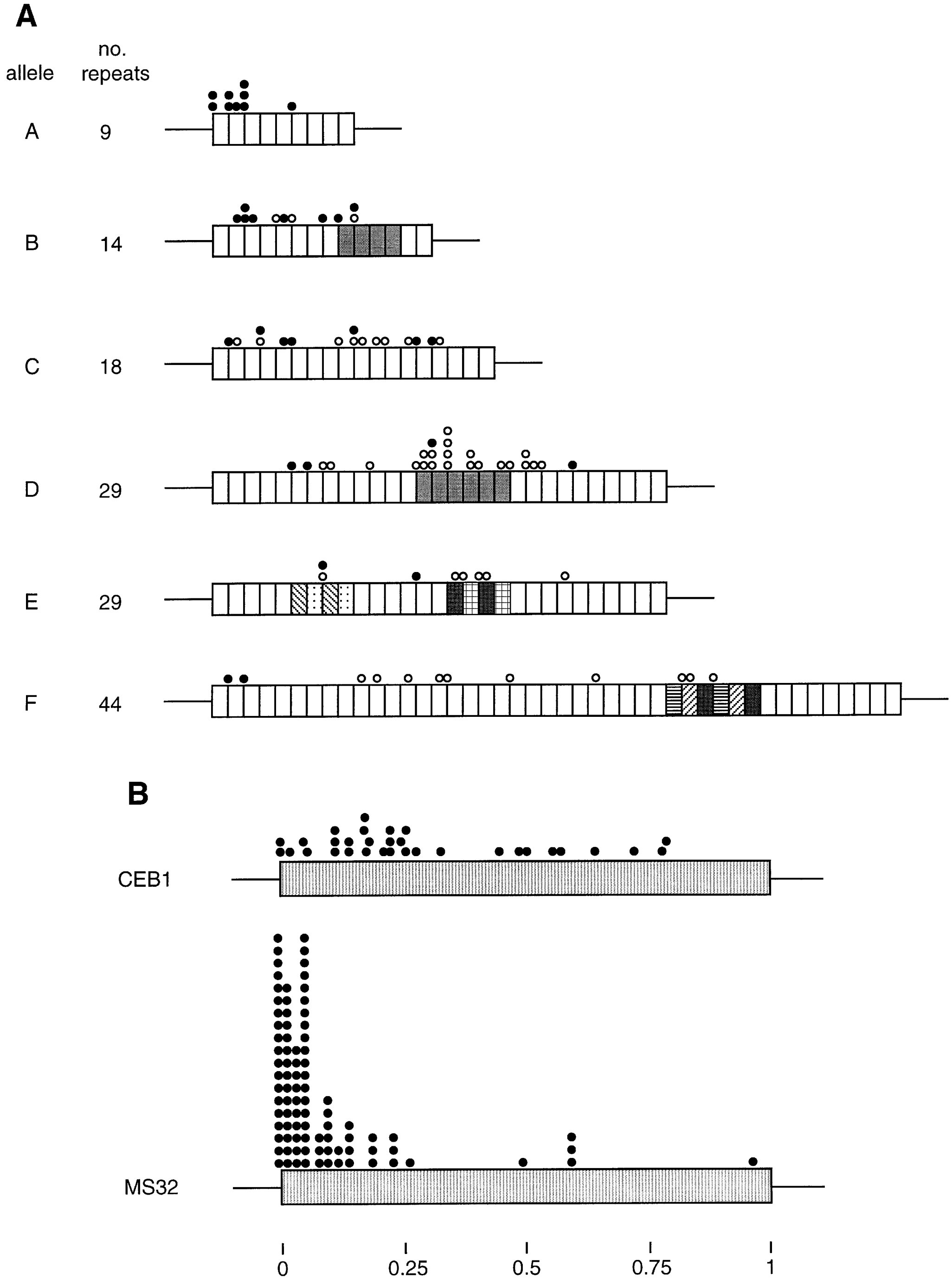


Fig. 6. Distributions of mutation events along repeat arrays. (**A**) Distribution of gain events for six CEB1 alleles. Each allele analysed by MVR-PCR is represented by a series of boxes representing the repeats. Black circles represent the insertion sites of inter-allelic transfers (or the middle of the duplicated block in case of target duplications of recipient allele repeats at the insertion site). White circles represent the middle of the original block of repeats which have been duplicated in intra-allelic duplication events. Homogeneous stretches at least four repeats long are represented by patterns. For example, allele D contains an homogeneous string of six identical repeats and allele F contains a string of six repeats made up of two identical copies of three contiguous repeats. Alleles D and E show a clustering of intra-allelic duplication events within the homogeneous portion of the allele. (**B**) Distributions of inter-allelic events at CEB1 and MS32. The 32 inter-allelic insertion sites observed for the six CEB1 alleles, and 90 inter-allelic insertion sites from four different MS32 alleles (Jeffreys *et al.*, 1994) are represented on a single imaginary tandem array according to their relative location.

trinucleotide repeats (Eichler *et al.*, 1994; Warren, 1996). Instead, most intra-allelic duplications are complex and involve secondary rearrangements very similar to those observed for complex inter-allelic transfers; this similarity suggests that inter- and intra-allelic mutations may proceed, to some extent, via a common recombination-based pathway.

Inter-allelic transfers frequently show target site duplications or deletions at the insertion site, consistent with the previously proposed model (Buard and Vergnaud, 1994) in which staggered nicks initiate the formation of a DSB in the tandem array. A protruding single strand of the broken allele could use the intact partner allele as a template for repair, leading to inter-allelic transfer. It could also more simply pair with the complementary single-strand end exposed by the staggered nicks. The shorter these single-strand ends, the more frequently pairing would tend to occur in register to restore the initial allele structure. This hypothesis predicts that DSBs are formed in stable short alleles but that they are repaired without addition or subtraction of repeats. For larger alleles generating longer single-strand ends more frequently, this pairing process could occur more often out of register, particularly if several contiguous repeats at the ends of the complementary single strands are identical, leading both to higher rates of intra-allelic duplication and to the clustering of these events in homogeneous stretches of repeats.

Each inter-allelic transfer and intra-allelic duplication event mapped in this study is unique, consistent with a meiotic mutation process for expansions. In contrast, five instances of duplicate mutants were found among deletions, suggesting that these losses can occur pre-meiotically during replication of proliferative germ cells. Four of these deletion events involve the simple loss of a block of contiguous repeat units, suggesting a distinct stem cell mutation process and consistent with the simple mode of somatic instability defined at minisatellite MS32 (Jeffreys and Neuman, 1997).

As with other human minisatellites, CEB1 shows polarity for inter-allelic transfers of repeats, albeit relatively diffuse, consistent with evidence that suggests that array instability could be influenced by flanking DNA (Monckton *et al.*, 1994). Intra-allelic rearrangements, though apparently mechanistically related to inter-allelic mutation, are not polar and imply different processing of recombination intermediates according to their location within the repeat array. As yet there is no clear evidence for the flanking haplotype influencing CEB1 array instability, though the remarkable difference in inter-allelic rate between a 10 and a nine-repeat allele (0.06 and 1.0%, respectively) suggests that factors other than array length and internal structure might influence the frequency of conversion. Finally, several mutants consist of the beginning of one allele fused to the end of the other. It remains to be seen whether these represent true recombinant alleles arising at meiosis and, more generally, whether minisatellite instability is related to true recombination at meiosis.

Materials and methods

Small-pool PCR

All manipulations of genomic DNAs, sperm and blood were performed in a laminar flow hood to minimize the risk of contamination. Sperm DNAs were extracted as described elsewhere (Jeffreys *et al.*, 1994). *Mbo*I- or *Bgl*I-digested sperm DNAs were typically diluted to 240 pg/µl and 480 pg/µl. Eight 1µl aliquots of each dilution were amplified in 7 µl total volume using the PCR buffer described elsewhere (Jeffreys *et al.*, 1990) and 0.2 µM of each primer. Universal CEB1 primers P9 and P14 were used for simultaneous amplification of both alleles from *Mbo*I digests, and one allele-specific primer plus a universal primer were used for specific amplification of only one of the two alleles from *Bgl*I digests in individuals heterozygous for at least one of four flanking substitutional polymorphisms identified by sequencing. Amplifications were performed in a Perkin–Elmer GeneAmp PCR system 9600 at 96° for 45 s followed by 28 cycles at 96° for 30 s, 68° for 45 s and 70° for 4 min. Agarose gel electrophoresis, transfer to nylon membranes and hybridization of PCR products with the CEB1 probe were performed as described elsewhere (Jeffreys *et al*., 1994).

Statistical analysis

Most statistical procedures used are described by Sokal and Rohlf (1995). For each allele and each SP-PCR experiment, the number of amplifiable progenitor molecules per SP-PCR was estimated by Poisson analysis of data from a parallel PCR with limiting dilutions of sperm DNA as described elsewhere (Jeffreys *et al.*, 1994). The 95% confidence intervals for the mutation rates take into account both errors associated with this Poisson estimation (Sachs, 1982) and sampling errors in counting mutants.

Recovery of CEB1 mutant molecules and structural analysis

CEB1 mutant molecules were recovered either from SP-PCRs or, for the majority of MVR-mapped mutants, from SESP-PCR reactions (Jeffreys and Neuman, 1997). Size enrichment of mutants was performed in most cases because a high level of contamination by the progenitor allele was often observed after re-amplification of CEB1 mutant molecules gel-purified from SP-PCRs. Typically, a mutant detected by SESP-PCR was re-amplified for six cycles from 1 µl of the initial PCR using the same primers, resolved by electrophoresis and recovered by gel-purification through glass-wool columns (Heery *et al.*, 1990), then reamplified for 30 cycles using 1 µl of the purified DNA and nested primers. The resulting mutant PCR product was detected by staining with ethidium bromide and gel-purified as above. MVR-PCR was performed as described elsewhere (Buard and Vergnaud, 1994), using an appropriate flanking primer (universal or allele-specific) and 1 µl of a 200-fold dilution of the purified PCR product. The MVR coding system used previously at CEB1 (Buard and Vergnaud, 1994), and which reflects the actual sequence at each of three polymorphic sites between repeats (each repeat being read as a triplet), has been replaced by a more synthetic code (each repeat being read as one alphabetical letter). The correspondence between the two codes is: ACC = A; GCC = B; oCC = C; AoC = D; GoC = E; ooC = F; oTC = G; ACT = H; GCT = I; oCT = J; oTT = K; AoT = L; GoT = M; ooT = N; oTo = O; ooo = P; Goo = Q; Aoo = R; GCo = S; ACo = T; oCo = U; and GTC = V, where o represents unknown variants that block priming by MVR primers.

Sequences of primers

5′ flanking sequence primers of CEB1 are as follows. Universal primer: P9, 5′-CGG AGC TCT GCT GAG TCA GAG-3′. Allele-specific primers: –4C, 5′-GGC AGG AGC TCT GCT GAG TCC-3′; –4A, 5′-GGC AGG AGC TCT GCT GAG TCA-3′; –72G, 5′-CGG ACC CCA GTG TAA TGG GG-3′; –72A, 5′-CGG ACC CCA GTG TAA TGG GA-3′.

3′ flanking sequence primers of CEB1 are as follows. Universal primer: P14, 5′-GGA TCC TCT CCT GTG CCT TTC CT-3′. Allele-specific primers: +384G, 5′-GAG GAA GAT CTT CAG GAC CAG-3′; +384A, 5′-GAG GAA GAT CTT CAG GAC CAA-3′; +256G, 5′-TAA TCT GGA GTT GGT CTG GCG-3′; +256A, 5′-TAA TCT GGA GTT GGT CTG GCA-3′.

Sequences of variant repeat specific primers used for MVR-PCR are described elsewhere (Buard and Vergnaud, 1994).

Acknowledgements

We thank J.Blower for sample collection and semen donors for their contribution. We are also grateful to colleagues for helpful discussions and in particular to Celia May for very helpful comments on the manuscript. J.B. is an EMBO Research Fellow. A.J.J. is supported in part by an International Research Scholars Award from the Howard Hughes Medical Institute and in part by grants from the Medical Research Council, the Wellcome Trust and the Royal Society.

References

Andreassen,R., Egeland,T. and Olaisen,B. (1996) Mutation rate in the hypervariable VNTR g3 (D7S22) is affected by allele length and a flanking DNA sequence polymorphism near the repeat array. *Am. J. Hum. Genet.*, **59**, 360–367.

Andrew,S.E., Goldberg,Y.P. and Hayden,M.R. (1997) Rethinking genotype and phenotype correlations in polyglutamine expansion disorders. *Hum. Mol. Genet.*, **6**, 2005–2010.

Buard,J. and Vergnaud,G. (1994) Complex recombination events at the hypermutable minisatellite CEB1 (D2S90). *EMBO J.*, **13**, 3203–3210.

Buard,J. and Jeffreys,A.J. (1997) Big, bad minisatellites. *Nature Genet.*, **15**, 327–328.

Chung,M.-Y., Ranum,L.P.W., Duvick,L.A., Servadio,A., Zoghbi,H.Y. and Orr,H.T. (1993) Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nature Genet.*, **5**, 254–258.

Eichler,E.E., Holden,J.J.A., Popovich,B.W., Reiss,A.L., Snow,K., Thibodeau,S.N., Richards,C.S., Ward,P.A. and Nelson,D.L. (1994) Length of uninterrupted CGG repeats determines instability in the FMR1 gene. *Nature Genet.*, **8**, 88–94.

Gacy,A.M., Goellner,G., Juranic,N., Macura,S. and McMurray,C.T. (1995) Trinucleotide repeats that expand in human disease form hairpin structures *in vitro*. *Cell*, **81**, 533–540.

Heery,D.M., Gannon,F. and Powell,R. (1990) A simple method for subcloning DNA fragments from gel slices. *Trends Genet.*, **6**, 173.

Jeffreys,A.J. and Neuman,R. (1997) Somatic mutation processes at a human minisatellite. *Hum. Mol. Genet.*, **6**, 129–136.

Jeffreys,A.J., Royle,N.J., Wilson,V. and Wong,Z. (1988) Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature*, **332**, 278–281.

Jeffreys,A.J., Neumann,R. and Wilson,V. (1990) Repeat unit sequence variation in minisatellites: a novel source of DNA polymorphism for studying variation and mutation by single molecule analysis. *Cell*, **60**, 473–485.

Jeffreys,A.J., MacLeod,A., Tamaki,K., Neil,D.L. and Monckton,D.G. (1991) Minisatellite repeat coding as a digital approach to DNA typing. *Nature*, **354**, 204–209.

Jeffreys,A.J., Tamaki,K., MacLeod,A., Monckton,D.G., Neil,D.L. and Armour,J.A.L. (1994) Complex gene conversion events in germline mutation at human minisatellites. *Nature Genet.*, **6**, 136–145.

Jeffreys,A.J. *et al.* (1997) Spontaneous and induced minisatellite instability. *Electrophoresis*, **18**, 1501–1511.

May,C.A., Jeffreys,A.J. and Armour,J.A.L. (1996) Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). *Hum. Mol. Genet.*, **5**, 1823–1833.

Mitas,M. (1997) Trinucleotide repeats associated with human disease. *Nucleic Acids Res.*, **25**, 2245–2253.

Monckton,D.G., Tamaki,K., MacLeod,A., Neil,D.L. and Jeffreys,A.J. (1993) Allele-specific MVR-PCR analysis at minisatellite D1S8. *Hum. Mol. Genet.*, **2**, 513–519.

Monckton,D.G., Neumann,R., Guram,T., Fretwell,N., Tamaki,K., MacLeod,A. and Jeffreys,A.J. (1994) Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. *Nature Genet.*, **8**, 162–170.

Monckton,D.G., Wong,L.-J.C., Ashizawa,T. and Caskey,C.T. (1995) Somatic mosaicism, germline expansions, germline reversions and intergenerational reductions in myotonic dystrophy males: small pool PCR analyses. *Hum. Mol. Genet.*, **4**, 1–8.

Sachs,L. (1982) *Applied Statistics*. Springer-Verlag, New York, NY.

Sokal,R.R. and Rohlf,F.J. (1995) *Biometry.* Freeman, New York, NY.

Vergnaud,G., Mariat,D., Apiou,F., Aurias,A., Lathrop,M. and Lauthier,V. (1991) The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence. *Genomics*, **11**, 135–144.

Warren,S.T. (1996) The expanding world of trinucleotide repeats. *Science*, **271**, 1374–1375.

Received March 4, 1998; revised April 8, 1998; accepted April 20, 1998