Am. J. Hum. Genet. 70:1038–1043, 2002

Report

Evolutionary Fate of an Unstable Human Minisatellite Deduced from Sperm-Mutation Spectra of Individual Alleles

Jérôme Buard, Charles Brenner,* and Alec J. Jeffreys

Department of Genetics, University of Leicester, Leicester, United Kingdom

Although mutation processes at some human minisatellites have been extensively characterized, the evolutionary fate of these unstable loci is unknown. Minisatellite instability is largely germline specific, with mutation rates up to several percent and with expansion events predominating over contractions. Using allele-specific small-pool polymerase chain reaction, we have determined sperm-mutation spectra of individual alleles of the highly unstable human minisatellite CEB1 (i.e., D2S90). We show that, as allele size increases, the proportion of contractions rises from <5% to 50%, with the average size of deletion increasing and eventually exceeding the average size of expansion. The expected net effect of these trends after many generations is an equilibrium distribution of allele sizes, and allele-frequency data suggest that this equilibrium state has been reached in some contemporary human populations.

Mutation is a major driving force in genome evolution. Tandem repeats are ubiquitous elements that account for a substantial proportion of nuclear DNA and include some of the most unstable elements in eukaryotic genomes. Microsatellites (short tandem repeats) and some AT-rich minisatellites are thought to gain and lose repeat units via replication-slippage errors that escape correction by the mismatch-repair system, particularly when hairpinlike secondary structures could form (Levinson and Gutman 1987; Desmarais et al. 1993; Strand et al. 1993; Gacy et al. 1995; Yu et al. 1997; Schlotterer 2000). In contrast, all GC-rich minisatellites characterized to date mutate via a complex meiotic recombination pathway (reviewed in Buard and Jeffreys 1997 and Jeffreys et al. 1997). However, little is known about the evolutionary fate of these elements. A general picture of microsatellite evolution has recently emerged from analyses of several hundred mutations detected among hundreds of dinucleotide- and tetranucleotide-repeat loci, in the course of large-scale genome mapping and legal paternity testing (Ellegren 2000; Xu et al. 2000). On average, short alleles tend to increase, and longer alleles tend to decrease, resulting in an upper limit on allele length and an apparently stable distribution of microsatellite allele sizes. However, these analyses could be confounded by interlocus variation in the direction of mutation, and the same conclusions could be drawn by pooling data from microsatellites with opposite mutational trends.

The locus-specific analysis of mutation parameters requires the scoring of large numbers of mutants for a given tandem repeat. This is not feasible for microsatellites but is possible for tandem repeats, such as unstable GC-rich human minisatellites (Jeffreys et al. 1988; Vergnaud et al. 1991). Although minisatellite mutation is generally biased toward expansion, little information exists on the net repeat-copy–number change per generation. It is this net change that dictates the evolutionary fate of a locus, whether toward infinite growth, extinction by progressive shortening, or equilibrium. We have

Received July 23, 2001; accepted for publication January 10, 2001; electronically published February 21, 2002.

Address for correspondence and reprints: Dr. Jérôme Buard, Institut de Génétique Humaine, CNRS UPR 1142, 141 rue de la Cardonille, 34296 Montpellier cedex 5, France. E-mail: Jerome.Buard@igh.cnrs.fr

* Present affiliation: Department of Public Health, University of California, Berkeley.

0002-9297/2002/7004-0024$15.00

therefore analyzed allelic variation in sperm-mutation parameters at human minisatellite CEB1, which mutates almost exclusively in the male germline, at rates up to 21% for some alleles and with most alleles short enough to be analyzed for mutation by small-pool PCR of sperm DNA (Vergnaud et al. 1991; Buard and Vergnaud 1994; Buard et al. 1998). Although minisatellite CEB1 is one of the most unstable minisatellites characterized, most features of its instability (including numerical bias toward expansion, paternal bias for the origin of mutation, and a complex mutation process including intra- and interallelic rearrangements [reviewed in Jeffreys et al. 1997]) are found among other GC-rich minisatellites. Since the vast majority of minisatellite rearrangements are not associated with flanking marker exchange (Wolff et al. 1988, 1989; Vergnaud et al. 1991; Jeffreys et al. 1994), it is possible to use single-nucleotide polymorphisms (SNPs) that flank minisatellite arrays to selectively amplify and score, from diluted aliquots of sperm DNA, new length molecules that arise from one allele (Monckton et al. 1994). Allele-specific small-pool PCR (ASSP-PCR) therefore provides information on the mutation rate and the spectrum of individual alleles, allowing analysis of allele-specific variation in mutation rate and expansion bias. These parameters have, in turn, enabled us to simulate the evolution of this minisatellite over many generations and to compare predicted allele-size distributions with those observed in contemporary human populations.

Sperm-mutation spectra of 22 CEB1 alleles were determined by size fractionation followed by ASSP-PCR, for the two largest alleles (88 and 98 repeats), and by direct ASSP-PCR, for 20 alleles ranging in size from 5 to 69 repeats. Mutation rates of only 12 of these 20 alleles have been described elsewhere (Buard et al. 1998). The number of amplifiable molecules being screened for mutation varies from 20,000, for the shortest (5 repeats) allele tested, to 440, for the longest (>20 repeats) alleles. Although small-pool PCR does not produce PCR artifacts at significant rates (Jeffreys et al. 1994; May et al. 1996; Jeffreys and Neumann 1997; Buard et al. 1998; Tamaki et al. 1999; Buard et al. 2000*a*, 2000*b*), pilot experiments on alleles >80 repeats in length produced some very large deletions with weak signal intensities, suggesting that they may not correspond to authentic sperm mutants. We therefore used size fractionation of sperm DNA (Jeffreys and Neumann 1997; Buard et al. 2000*b*) to validate large deletions for an 88-repeat allele and a 98-repeat allele. Two sperm donors were analyzed; each carried long and short CEB1 alleles (i.e., one donor carried CEB1 alleles of 88 and 20 repeats, and the other donor carried CEB1 alleles of 98 and 11 repeats) and was heterozygous at a flanking SNP located 3′ to the minisatellite. Sperm DNA was digested with restriction endonuclease *Bgl*I and was separated by gel electrophoresis, and a size fraction was collected that contained the short allele plus any large deletions that were derived from the longer allele. DNA recovery was estimated by ASSP-PCR amplification of the short progenitor allele from limiting dilutions of the fractionated DNA and by Poisson analysis of the resulting proportion of positive reactions (Jeffreys et al. 1994). By use of the alternative allele-specific primer, large deletions derived from the longer allele and of the correct size range for the DNA fraction were recovered by ASSP-PCR. The internal structures of mutants derived from the 98-repeat allele, determined by minisatellite variant repeat PCR (Jeffreys et al. 1991), showed that most (15/17) had resulted unambiguously from large deletion events of the long progenitor allele, rather than from exchanges with the shorter allele (data not shown). The mutation spectra of each of the two longest alleles (one of which is shown in fig. 1) were thus constructed in part from ASSP-PCR analysis for expansions and relatively small deletions and in part from size-fractionation analysis of larger

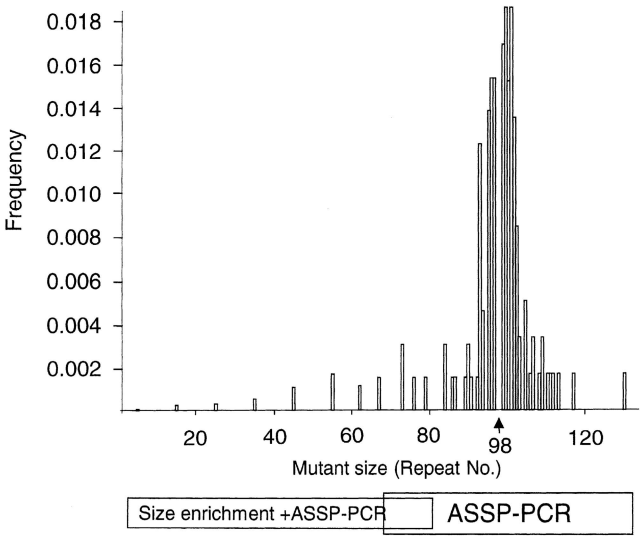

Figure 1 Sperm-mutation spectrum of a 98-repeat CEB1 allele. After PCR amplification of 44 diluted aliquots of sperm DNA (each containing ~12 CEB1 molecules of the 98-repeat allele) by use of universal 5′-flanking primer P9 and allele-specific 3′-flanking primer 384G (Buard et al. 1998), PCR products were electrophoresed in a 40-cm-long 0.7% agarose (Seakem HGT) gel in 0.5× Tris-borate EDTA buffer, were transferred to a nylon membrane, and were hybridized with a CEB1 probe. One hundred ten mutant molecules >2.8 kb (70 repeats) were detected. Sizes were converted to the number of CEB1 repeats (1 consensus repeat = 40 bp). Large deletions with only 5–80 repeats remaining were detected by size-enrichment ASSP-PCR (see text). The frequencies of mutants 70–80 repeats long (scored in both ASSP-PCR and size-enrichment ASSP-PCR) were similar between the two experiments, suggesting that PCR artifacts do not contribute significantly to losses of ≤28 repeats for this allele. The mutation frequency for each repeat copy number was deduced and merged from these two experiments.

deletions. Similar frequencies have been found for mutant molecules lying within the size range explored by the two approaches (−20 to −30 repeats), suggesting that no bias has been introduced by pooling together these two kinds of data for constructing the whole mutation spectrum.

A total of 777 sperm mutants were detected from these 22 alleles. As described elsewhere, the CEB1 sperm-mutation rate increases with array size (Buard et al. 1998), from <0.05%, for the shortest (5 repeats) allele tested, to 21%, for the longest (98 repeats) allele tested (fig. 2*a*).

The mutant molecules consist of 478 expansion events and 299 deletions. This is consistent with the 2:1 gain:loss ratio previously estimated from pedigree mutants (Vergnaud et al. 1991). However, the ASSP-PCR data showed that there was considerable variation in the gain bias among alleles. The proportion of deletion events increases asymptotically with array size, from <5%, for alleles with <10 repeats, to 50%, for arrays with >70 repeats (fig. 2*b*). For the shortest alleles, the average size of a deletion event is smaller than the average size of expansion. In contrast, the longest alleles show a mean deletion size that exceeds the mean expansion size. The mean deletion size increases with array length approximately four times more rapidly than does the mean expansion size; the two values achieve equality at ~62 repeats (fig. 2*c*).

The following findings resolve the apparent paradox between an overall excess of gain mutation events and the relative rarity of long CEB1 alleles:

1. Short alleles show a strong bias toward expansion events, whereas large alleles show equal rates of expansion and contraction.
2. The net repeat-copy–number change per mutation event is positive for short alleles, null for a critical array length, and negative for large alleles.

Hence, short CEB1 alleles tend to increase, and long CEB1 alleles tend to decrease.

These mutation parameters also enabled us to investigate the evolutionary fate of this minisatellite. For a given allele size, the allele frequency after one generation is given by the frequency of nonmutated alleles plus the frequency of mutant alleles reaching this size. Using the equations derived from the variable parameters of the mutation spectra (see fig. 2 legend), we built a transition matrix to estimate allele frequencies at generation $i + 1$ from allele frequencies at generation i. Irrespective of the starting allele size, the Markov process specified by the transition matrix produced the same stable and virtually unchanging allele-size distribution after 1,000 generations (~20,000 years) in an infinite population. The existence of such a robust equilibrium size distribution is to be expected, because of the Perron-Frobenius theorem (Moran 1976; Kingman 1977). At equilibrium, the allele-size distribution is unimodal, with maximum at 49 repeats, and is skewed toward longer alleles, with an average allele size of 63.5 repeats (fig. 3*a*). The robustness of this mathematical model of CEB1 evolution was further tested by experimentation with various alternative curve-fitting choices (for details, see Appendix A, available online only). The alteration of the choice of

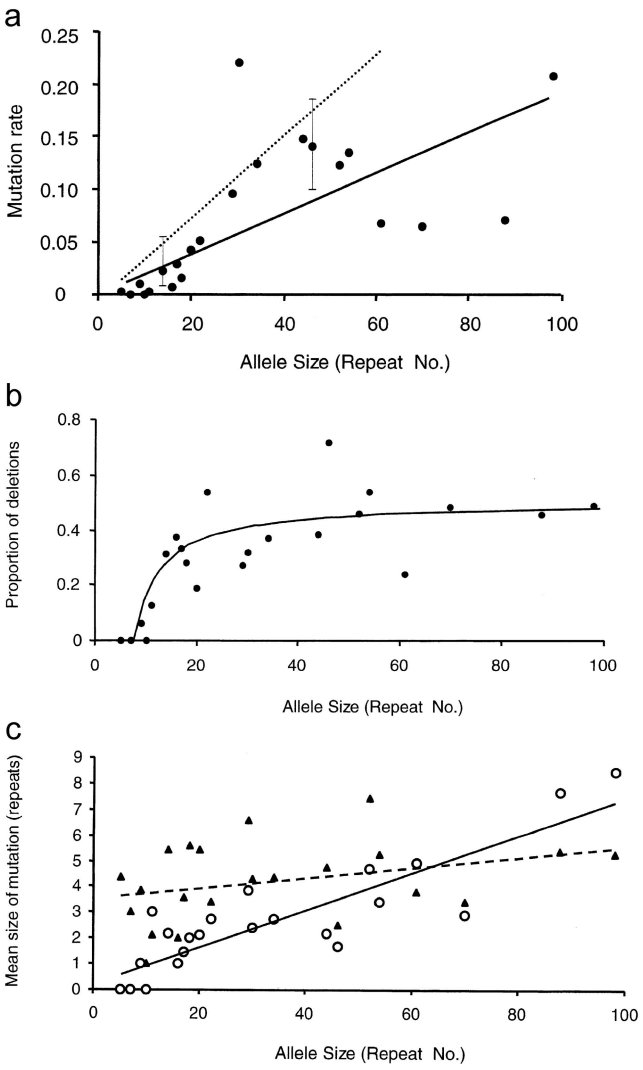

Figure 2 Influence that array length has on the rate and extent of CEB1 expansions and deletions in sperm. *a,* Mutation rate versus array length. Sperm-mutation rates of 22 CEB1 alleles ranging from 5 to 98 repeats long were determined from small-pool PCR data. Mutation rate (m) increases with array size (s) as $m = s/521$ ($R^2 = 0.4$, $t = 3.63$, $P < .01$). Different models for mutation rate as a function of array size have been chosen for simulating CEB1 evolution, such as $m = s/260$ (*dashed line*) for instance. *b,* Proportion of deletions versus array length. The proportion of deletions d increases with array length (s) as $d = 0.5 - (3.3/s)$ ($R^2 = 0.63$, $t = 5.8$, $P < .001$). *c,* Mean size of expansion (*triangles*) or deletion (*circles*) per mutation event. Fit for mean size of expansion (*dashed line*) $E = (s/50) + 3.54$ ($R^2 = 0.11$, $t = 1.6$, $P = .12$). Fit for mean size of deletion (*solid line*) $D = s/13$ ($R^2 = 0.67$, $t = 5.8$, $P < .001$).

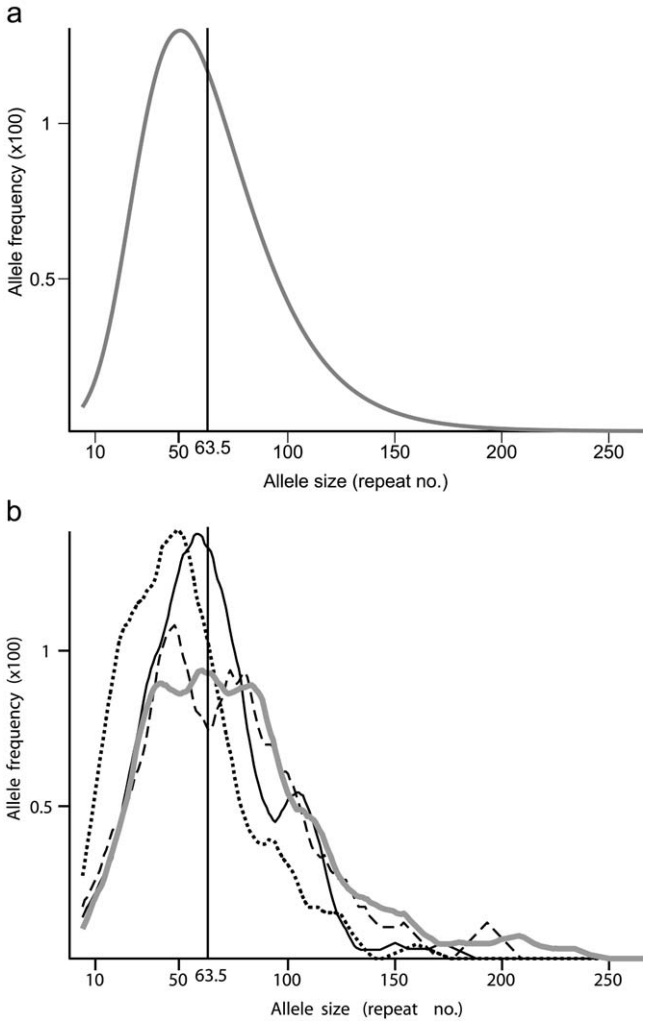

Figure 3 Allele-size distributions at CEB1. *a,* Equilibrium distribution simulated using CEB1 sperm-mutation parameters. The model assumes an infinite population of alleles, with arbitrary initial distribution of sizes and with new alleles generated from old by a modified step model of mutation. A transition matrix was built by use of CEB1 rates and sizes of expansions and deletions relative to array size and was used to estimate allele frequencies at generation $i + 1$ from allele frequencies at generation i in an infinite population (for details, see Appendix A, available online only). The equilibrium distribution (mean allele size 63.5 repeats), achieved within 1,000 generations, is shown. *b,* CEB1 allele-size distributions in contemporary European (*thick gray line*), Indian (*dashed line*), African (*dotted line*), and Japanese (*thin black line*) human populations, determined by Southern blot sizing of 250, 48, 120, and 110 alleles, respectively. Mean allele sizes are 78, 51, 74, and 80 repeats, respectively, and can be compared with mean allele size of the simulated equilibrium distribution (63.5 repeats). European, Japanese, and, in particular, African allele-size distributions are distinct (Kolmogorov-Smirnoff test; $P < .005$ for all pairwise comparisons). The smoothed plot of the population distributions is obtained by convolution with a triangular-shaped kernel, which approximately imitates gel-measurement uncertainty.

model for the rate of mutation as a function of array size (fig. 2*a*) affects the time to equilibrium but has a negligible effect on the equilibrium distribution. Also, regression analysis showed that there was little to choose between the best linear curve fit for the average size of expansion (E) relative to array length ($R^2 = 0.11$; fig. 2*c*), and a constant mean size of expansion across all alleles ($E = 4.2$). The latter model reduces the predicted incidence of longer (>80 repeats) alleles, thus modestly reducing the average allele size (by 6 repeats) and the variance at equilibrium. A surprisingly critical aspect of the model, however, is the distribution of deletion amounts about its stipulated mean, which we model with a slower-than-exponential decay. If exponential decay is instead used (as it is for expansions), then nearly all alleles of <20 repeats disappear from the equilibrium population. This suggests that there may be two modes of contraction: one that favors small changes, whose relative proportion may conform to a rule of exponential decay; and another that is a cataclysmic mode that produces small alleles more or less independently of the progenitor size and is responsible for populating the short-allele end of the spectrum.

To see whether this predicted equilibrium distribution is seen in humans, we determined CEB1 allele-size distributions in European (British), African (Zimbabwean), Japanese, and Indian populations (fig. 3*b*). All populations are similar to the simulated size distribution at equilibrium (fig. 3*a*) with respect to both mean and variance and in having positive skew. This suggests that mathematical simulation by use of sperm-mutation spectra parameters of CEB1 alleles provides an adequate model of minisatellite evolution. It also implies that these four populations are near mutation equilibrium and that their allele distributions will remain stable in the indefinite future, despite the numerical bias toward mutational expansion at CEB1. By Kolmogorov-Smirnoff tests, most of the populations are found to be distinct both from one another and from the simulated equilibrium (6/10 P values <.01), but this may merely signify that the sample sizes are large enough to reveal the genetic drift that inevitably has occurred, which separated the populations to some extent. However, the Zimbabweans are particularly distinct, with a notable tendency toward shorter allele lengths. Further refinement of the evolutionary model and collection of additional population data are required to see whether deviations from the predicted equilibrium distribution signal the existence of alleles with atypical mutational spectra or, instead, whether such abnormal distributions arise from departures from equilibrium in finite populations—for example, by a recent partial fixation of short alleles via a bottleneck that could lead to a temporary reduction in mean allele size.

These balancing effects of minisatellite expansions and

deletions are strikingly reminiscent of the general mutational behavior of microsatellites, with the net repeat-copy–number change being positive, for short alleles, and negative, for long alleles (Ellegren 2000; Xu et al. 2000). This similarity is puzzling, because microsatellites and minisatellites are thought to mutate via very different pathways. However, evidence that microsatellites mutate in the human germline by replication slippage remains indirect and inconclusive and is countered both by difficulties in the explanation of how slippage can result in changes in expansion bias with array size (Xu et al. 2000) and by arguments that microsatellite instability may be influenced by relative array sizes in heterozygotes (Amos et al. 1996). In contrast, there is a high affinity of bacterial RecA protein for CA repeats (Dutreix 1997), in addition to a significant correlation between chromosomal distribution of CA repeats and recombination hotspots along human chromosome 22 (Majewski and Ott 2000). It therefore remains possible that recombination may play a role in microsatellite instability in the germline. Both minisatellite and microsatellite contractions could result from replication-based repair of germline-specific double-strand breaks, with minisatellites also gaining repeats via recombination-based repair of the same initiating lesion.

Acknowledgments

We thank Jane Blower, for supplying semen samples from anonymous donors, and Kathryn Lilley and Stuart Bayliss, for oligonucleotide synthesis. We are grateful to colleagues for helpful discussions. A.J.J. was supported in part by an International Research Scholars Award from the Howard Hughes Medical Institute and grants from the Wellcome Trust, Medical Research Council, and Royal Society. J.B. is a Medical Research Council fellow.

References

Amos W, Sawcer SJ, Feakes RW, Rubinsztein DC (1996) Microsatellites show mutational bias and heterozygote instability. Nat Genet 13:390–391

Buard J, Bourdet A, Yardley J, Dubrova Y, Jeffreys AJ (1998) Influences of array size and homogeneity on minisatellite mutation. EMBO J 17:3495–3502

Buard J, Collick A, Brown J, Jeffreys AJ (2000*a*) Somatic versus germline mutation processes at minisatellite CEB1 (D2S90) in humans and transgenic mice. Genomics 65:95–103

Buard J, Jeffreys AJ (1997) Big, bad minisatellites. Nat Genet 15:327–328

Buard J, Shone AC, Jeffreys AJ (2000*b*) Meiotic recombination and flanking marker exchange at the highly unstable human minisatellite CEB1 (D2S90). Am J Hum Genet 67:333–344

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

Desmarais E, Vigneron S, Buresi C, Cambien F, Cambou JP, Roizes G (1993) Variant mapping of the Apo(B) AT rich minisatellite. Dependence on nucleotide sequence of the copy number variations. Instability of the non-canonical alleles. Nucleic Acids Res 21:2179–2184

Dutreix M (1997) $(GT)_n$ repetitive tracts affect several stages of RecA-promoted recombination. J Mol Biol 273:105–113

Ellegren H (2000) Heterogeneous mutation processes in human microsatellite DNA sequences. Nat Genet 24:400–402

Gacy AM, Goellner G, Juranic N, Macura S, McMurray CT (1995) Trinucleotide repeats that expand in human disease form hairpin structures in vitro. Cell 81:533–540

Jeffreys AJ, Bois P, Buard J, Collick A, Dubrova Y, Hollies CR, May CA, et al (1997) Spontaneous and induced minisatellite instability. Electrophoresis 18:1501–1511

Jeffreys AJ, MacLeod A, Tamaki K, Neil DL, Monckton DG (1991) Minisatellite repeat coding as a digital approach to DNA typing. Nature 354:204–209

Jeffreys AJ, Neumann R (1997) Somatic mutation processes at a human minisatellite. Hum Mol Genet 6:129–136

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

Jeffreys AJ, Tamaki K, MacLeod A, Monckton DG, Neil DL, Armour JAL (1994) Complex gene conversion events in germline mutation at human minisatellites. Nat Genet 6: 136–145

Kingman JF (1977) On the properties of bilinear models for the balance between genetic mutation and selection. Math Proc Camb Philos Soc 81:443–453

Levinson G, Gutman GA (1987) High frequency of short frameshifts in poly-CA/TG tandem repeats borne by bacteriophage M13 in *Escherichia coli* K-12. Nucleic Acids Res 15:5323–5338

Majewski J, Ott J (2000) GT repeats are associated with recombination on human chromosome 22. Genome Res 10: 1108–1114

May CA, Jeffreys AJ, Armour JAL (1996) Mutation rate heterogeneity and the generation of allele diversity at the human minisatellite MS205 (D16S309). Hum Mol Genet 5:1823–1833

Monckton DG, Neumann R, Guram T, Fretwell N, Tamaki K, MacLeod A, Jeffreys AJ (1994) Minisatellite mutation rate variation associated with a flanking DNA sequence polymorphism. Nat Genet 8:162–170

Moran PAP (1976) Global stability of genetic systems governed by mutation and selection. Math Proc Camb Philos Soc 80:331–336

Schlotterer C (2000) Evolutionary dynamics of microsatellite DNA. Chromosoma 109:365–371

Strand M, Prolla TA, Liskay RM, Petes TD (1993) Destabilization of tracts of simple repetitive DNA in yeast by mutations affecting DNA mismatch repair. Nature 365:274–276

Tamaki K, May CA, Dubrova YE, Jeffreys AJ (1999) Extremely complex repeat shuffling during germline mutation at human minisatellite B6.7. Hum Mol Genet 8: 879–888

Vergnaud G, Mariat D, Apiou F, Aurias A, Lathrop M, Lau-

thier V (1991) The use of synthetic tandem repeats to isolate new VNTR loci: cloning of a human hypermutable sequence. Genomics 11:135–144

Wolff RK, Nakamura Y, White R (1988) Molecular characterization of a spontaneously generated new allele at a VNTR locus: no exchange of flanking DNA sequence. Genomics 3:347–351

Wolff RK, Plaetke R, Jeffreys AJ, White R (1989) Unequal crossingover between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. Genomics 5:382–384

Xu X, Peng M, Fang Z (2000) The direction of microsatellite mutations is dependent upon allele length. Nat Genet 24: 396–399

Yu S, Mangelsdorf M, Hewet D, Hobson L, Baker E, Eyre HJ, Lapsys N, et al (1997) Human chromosomal fragile site FRA16B is an amplified AT-rich minisatellite repeat. Cell 88:367–374