Evolution of the cryptic *FMR1* CGG repeat

Evan E. Eichler[1], Catherine B. Kunst[2], Kellie A. Lugenbeel[1], Oliver A. Ryder[3], Daniel Davison[4], Stephen T. Warren[2] & David L. Nelson[1]

We have sequenced the 5' untranslated region of the orthologous *FMR1* gene from 44 species of mammals. The CGG repeat is present in each species, suggesting conservation of the repeat over 150 million years of mammalian radiation. Most mammals possess small contiguous repeats (mean number of repeats = 8.0 +/– 0.8), but, in primates, the repeats are larger (mean= 20.0 +/– 2.3) and more highly interrupted. Parsimony analysis predicts that enlargement of the *FMR1* CGG repeat beyond 20 triplets has occurred in three different primate lineages. In man and gorilla, AGG interruptions occur with higher-order periodicity, suggesting that historical enlargement has involved incremental and vectorial addition of larger arrays demarcated by an interruption. Our data suggest that replication slippage and unequal crossing over have been operative during the evolution of this repeat.

[1]*Department of Molecular and Human Genetics, Human Genome Center, Baylor College of Medicine, Houston, Texas 77030, USA*
[2]*Howard Hughes Medical Institute, Departments of Biochemistry and Pediatrics, Emory University, School of Medicine, Atlanta, Georgia 30322, USA*
[3]*Center for Reproduction of Endangered Species, Zoological Society of San Diego, San Diego, California 92112, USA*
[4]*Department of Biochemical and Biophysical Sciences and Computer Science, University of Houston, 4800 Calhoun, Houston, Texas 77024-5934, USA*

Correspondence should be addressed to D.L.N.

Five unstable CGG trinucleotide repeat loci have been identified in the human genome, each of which is capable of hyperexpansion to generate a folate-sensitive fragile site[1–5]. So far, the *FRAXA* and *FRA11B* CGG triplet repeats are the only fragile sites clearly associated with a gene and a human disease[1, 6–8]. The molecular basis of the fragile X syndrome (*FRAXA*) is well established: the CGG repeat is located in the 5' UTR of the *FMR1* gene and is polymorphic, ranging from 5 to 50 repeat units in the human population[5, 9–11]. Once the length of the repeat increases beyond 60 repeats, the triplet, upon maternal transmission, is capable of hyperexpansion and subsequent methylation with the concomitant abolition of *FMR1* gene expression [12]. The majority of human *FMR1* CGG repeats possess a cryptic substructure punctuated by AGG interspersions[10, 13–15]. While the majority of CGG repeat alleles possess two AGG interspersions occurring once every 9/10 CGG repeat units, almost all alleles predisposed to hyperexpansion (premutation alleles) have a single or no AGG interruptions[10, 14]. Similarly, spinal cerebellar ataxia CAG repeats predisposed to expansion demonstrate the loss of interrupting CAT trinucleotides[16]. Furthermore, unstable *FMR1* alleles in the normal population possess fewer interruptions when compared to alleles of similar size in which only stable intergenerational transmissions have been demonstrated[14]. Thus, the reduction in the number of AGG interruptions may be a critical event in predisposing *FMR1* CGG repeat alleles to instability and eventual hyperexpansion[10, 14, 15].

Hyperexpansion of the *FMR1* CGG repeat is selectively disadvantageous suggesting an evolutionary pressure for loss of the triplet. Surprisingly, no human alleles have been identified with fewer than five repeats[5, 17]. This observation, along with the fact that *FMR1* CGG repeats greater than 5 units in length are capable of binding specific proteins[18] and the observation that brain mRNA appears particularly enriched for genes with CGG repeats located upstream of the translational initiation signal[19], implies that the position of the repeat in the 5' UTR of these genes is of some functional importance. One approach to evaluate the functional significance of microsatellites has been to assess their level of conservation by comparing sequence and position from a variety of distantly related organisms[20]. In this study, we have analyzed the evolution of the *FMR1* CGG repeat by analysis of 44 mammalian species.

FMR1 CGG repeats among non-primates

The *FMR1* CGG repeat locus was sequenced from 24 non-primate species representing 7 orders of mammals: 6 species of Carnivora, 5 species of Chiroptera, 7 species of Rodentia, 3 species of Artiodactyla, and single species representatives of the orders Cetacea, Edentata and Monotremata (Table 1 and Methods). The majority of mammalian orders have short uninterrupted CGG repeats (mean=8.0 +/– 0.8 repeats) at an orthologous position in the 5' UTR of the gene (Fig. 1). With the exception of phyllostamid bats (*Artibeus jamaicensis* and *Artibeus obscura*), repeat length was strikingly similar among the different non-primate species, ranging from 4–12 units (Table 1). Only one third (8/24) of the species examined were found to have interspersions disrupting the continuity of the

Table 1 *FMR1* CGG triplet repeats among non-primate mammals

Common name	Species	Repeat length	Sequence of triplet repeat
Cat	*Felis doesticus*	9	$(CGG)_9$
Lion	*Panthera leo*	8	$(CGG)_8$
Dog	*Canis familiaris*	11	$(CGG)_{11}$
Polar bear	*Euarctos americanus*	9	$(CGG)_9$
Mongoose	*Herpestes ichneumon*	6	(CGG)**CAG**$(CGG)_9$**CGA**
Seal	*Zalophus californianus*	9	$(CGG)_6$**AGG**$(CGG)_2$
Bat	*Myotis yumaneni*	8	$(CGG)_8$
Bat	*Carollia perspicillata*	9	$(CGG)_9$
Bat	*Micronycteris hirsuta*	12	$(CGG)_5$**TGG**$(CGG)_6$
Bat	*Artibeus jamaicensis*	19	$(CGG)_8$**CAG**$(CGG)_{10}$
Bat	*Artibeus obscura*	18	$(CGG)_7$**CAGCAG**$(CGG)_9$
Mouse	*Mus spretus*	10	$(CGG)_{10}$
Mouse	*Mus caroli*	9	$(CGG)_9$
Mouse	*Mus musculus*	9	$(CGG)_6$**CGA**$(CGG)_2$
Hamster	*Mesocricetus auratus*	5	$(CGG)_5$
Vole	*Microtus agrestis*	4	$(CGG)_4$
Rat	*Rattus norvegicus*	6	$(CGG)_4$**CGA**CGG
Squirrel	*Ammospermophilius harrisi*	11	$(CGG)_{11}$
Cow	*Bos taurus*	4	$(CGG)_4$
Sheep	*Ovis aries*	4	$(CGG)_4$
Pig	*Sus scrofa*	12	$(CGG)_{12}$
Dolphin	*Stenella plagiodon*	5	$(CGG)_2$**CAG**$(CGG)_2$
Hairy armadillo	*Chaetophractus villosus*	4	$(CGG)_4$
Platypus	*Ornithorhynchus anatinus*	5	$(CGG)_5$

The CGG repeat from exon 1 of *FMR1* was subcloned and sequenced from 24 species comprising 7 orders of mammals. The position of the CGG repeat was determined by alignment of flanking sequences (see Fig. 1). Most repeats were small (8.0 +/– 0.8) and without interruption.

repeat. When interspersions were observed (predominantly CAG and CGA), they were usually few in number (Table 1).

Primate *FMR1* CGG repeats

The *FMR1* CGG repeat was sequenced from 20 different species of primates. Seven families were represented: the single species of Hominidea (man), four species of Pongidae (orangutan, chimpanzee, bonobo and gorilla), two species of Hylobatidae (gibbon and siamang), eight species of Cercopithecoidea (Allen's monkey, rhesus monkey, baboon, drill, proboscis monkey, colobus monkey, golden monkey, and Douc's langur), two species of Cebidae (marmoset and squirrel monkey), one species of Lorisdae (slow loris) and two species of Lemuridae (brown lemur and black and white ruffled lemur) (Table 2 and Methods).

In contrast to non-primate mammals, the mean length of the repeat among primate species (20.1 +/– 2.3, n=19) is significantly greater (P<0.0001, t test) than that of non-primate species (8.0 +/–0.8) (Tables 1 and 2). In certain primate lineages, the increase in CGG repeat length is accompanied by an increase in the number of interruptions. The type of interruptions, furthermore, appears to be lineage-specific (AGG among hominoids, CGGG among cercopithecoids and CGA among hylobatids, Table 2).

Evolutionary analysis

Evolutionary genetic analysis was performed using PAUP (phylogenetic analysis using parsimony) v. 3.1.1 and MacClade v. 3.1 software packages. Considering the cercopithecoid and hominoid sequences (n=17), a total of 450–462 equally parsimonious trees were generated after both branch-and-bound and heursitic tree searches . Encoding the data (see Methods) did not significantly simplify parsimony. The majority-rule (>50%) consensus tree generated from the sequences closely approximates the generally accepted phylogeny of these primates (Fig. 2). The most likely ancestral states were determined at each branchpoint using the character change reconstruction option within PAUP and are displayed within a framework of known timepoints of evolutionary change within the catarrhine phylogeny (Fig. 2). The addition of allele variants for species with high heterozygosity (*P. troglodytes* and *P. paniscus*; see below)

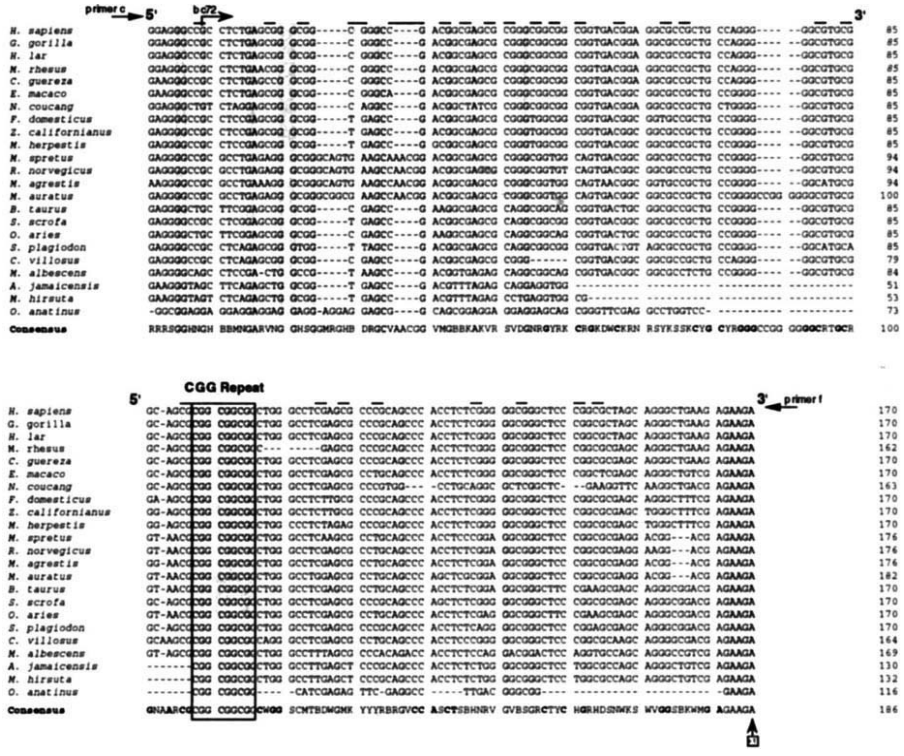

Fig 1 Alignment of sequence flanking the *FMR1* CGG repeat. Representative species from the different mammalian orders and families were chosen and alignment of the 5' UTR of the *FMR1* gene is depicted. Absolutely conserved nucleotides are shaded, while those which are conserved with the exception of being deleted in certain species are found in bold in the consensus sequence. For simplicity, the variable CGG repeat number was reduced to three triplets and is boxed in the alignment (for the length and content of this repeat for different species see Tables 1, 2 and 3). The position of the 25 CpG dinucleotides flanking the human repeat are denoted by bars above the sequence. The beginning of human cDNA clone bc72 is indicated with a fishhook horizontal arrow. The positions of primer c, which is located 38 bp downstream from the beginning of the human transcript and of primer f, which is contained entirely within the coding region of the *FMR1* gene [5, 9] are symbolised with horizontal arrows. The conserved adenine nucleotide of the ATG translational initiation codon is denoted by +1. In the consensus sequence: B=C/G/T, M=A/C, N=A/C/G/T, R=A/G, S=C/G, V=A/C/G, W=A/T, and Y=C/T.

Table 2 Primate *FMR1* CGG triplet repeats.

Common Name	Species	Repeat Length	Sequence of triplet repeat
Human	*Homo sapiens*	29	$(CGG)_9AGG(CGG)_9AGG(CGG)_9$*
Chimpanzee	*Pan troglodytes*	38	$(CGG)_9AGGCGGAGG(CGG)_9AGG(CGG)_{16}$
		36	$(CGG)_9AGGCGGAGG(CGG)_4AGGCGGAGGCGGAGG(CGG)_6AGG(CGG)_8$
		35	$(CGG)_9AGGCGGAGGCGGAGG(CGG)_9AGG(CGG)_{11}$
		34	$(CGG)_9AGGCGGAGG(CGG)_{22}$
		30	$(CGG)_9AGGCGGAGG(CGG)_{16}AGGCGG$
		30	$(CGG)_9AGG(CGG)_2AGG(CGG)_{17}$
		24	$(CGG)_8AGGCGGAGG(CGG)_2AGG(CGG)_2AGG(CGG)_7$
		27	$(CGG)_{10}AGG(CGG)_{16}$
		30	$(CGG)_{11}AGG(CGG)_{18}$
		20	$(CGG)_{20}$
Bonobo	*Pan paniscus*	39	$(CGG)_2CAG(CGG)_{12}AGG(CGG)_7AGG(CGG)_{15}$
		37	$(CGG)_2CAG(CGG)_{12}AGG(CGG)_7AGG(CGG)_{13}$
		37	$(CGG)_9AGG(CGG)_{13}AGG(CGG)_{13}$
		37	$(CGG)_{23}AGG(CGG)_{13}$
		37	$(CGG)_9AGG(CGG)_{27}$
Gorilla	*G. gorilla gorilla* *G. gorilla beringei* *G. gorilla graueri*	26	$(CGG)_8AGGCGGAGG(CGG)_3AGG(CGG)_3AGG(CGG)_3AGG(CGG)_3$
		22	$(CGG)_8AGGCGGAGG(CGG)_3AGG(CGG)_3AGG(CGG)_3$
Orangutan	*Pongo pygmaeus*	26	$(CGG)_9AGGCGGAGG(CGG)_2AGG(CGG)_2AGG(CGG)_8$
		26	$(CGG)_8AGGCGGAGG(CGG)_2AGG(CGG)_3AGG(CGG)_8$
Gibbon	*Hylobates lar*	24	$CGGCGACGGCGA(CGG)_2CGACGGCGA(CGG)_3AGG(CGG)_{11}$
		22	$CGGCGACGGCGA(CGG)_2CGA(CGG)_5AGG(CGG)_9$
Siamang	*Hylobates syndactylus*	24	$CGGCGACGGCGA(CGG)_2CGA(CGG)_5AGG(CGG)_3AGG(CGG)_7$
Baboon	*Papio*	32**	$(CGG)_{12}CGGG(CGG)_{19}$
Rhesus Monkey	*Macaca rhesus*	30**	$(CGG)_{13}CGGGCGGCGGG(CGG)_{14}$
Drill	*Mandrillus leucophaeus*	24**	$(CGG)_9CGGG(CGG)_{14}$
Allen's Monkey	*Allenopithecus nigroviridis*	20	$(CGG)_{20}$
Proboscis Monkey	*Nasalis larvatus*	7	$(CGG)_7$
Golden Monkey	*Rhinopithecus roxellana*	7	$(CGG)_7$
Douc's Langur	*Pygathrix nemaeus*	10	$(CGG)_{10}$
Colobus Monkey	*Colobus guereza*	8	$(CGG)_8$
Squirrel Monkey	*Saimiri sciuresus*	25	$CAG(CGG)_5CAG(CGG)_6CGA(CGG)_{12}$
Marmoset	*Lentopithecus saguinus*	9	$CGGCGA(CGG)_7$
Slow Loris	*Nycticebus coucang*	18	$(CGG)_{11}CAGCAG(CGG)_5$
		17	$(CGG)_{11}CAGCAG(CGG)_4$
Lemur	*Lemur fulvianus*	14	$(CGG)_{14}$
Ruffed Lemur	*Varecia variegata*	7	$(CGG)_7$

The *FMR1* CGG triplet repeat was sequenced from 19 different species of primates representing 8 different families within this order. Repeat length indicates the total number of CGG and interrupting triplets. The variability in the human *FMR1* CGG repeat sequence has been reported previously[10, 13-15]. **Among the cercopithecid, tribe Papionae, repeat length includes the CGGG tetranucleotide interspersion.

in subsequent parsimony analyses resulted in the placement of sympatric sequences in different clades. The cladistic disruption did not, however, extend beyond the different primate families (data not shown). A second parsimony analysis was performed using the derived ancestral catarrhine primate sequences and *M. spretus* and *O. anatinus* as outgroup sequences. These results predict that the ancestral primate sequence was short (7 CGG repeats), with the possibility of a polymorphic interspersion at the distal end of the repeat (Fig. 2). Parsimony analysis using branch-and-bound and heuristic searches was also performed with only the non-primate mammalian sequences. A reconstruction of the majority-rule ancestral consensus sequence from the various taxa predicts that the mammalian ancestral sequence was also short (ranging from 4 to 9 CGG repeats) and without interruption.

Population surveys of the *FMR1* CGG repeat

Five diverse mammalian populations (*Pan troglodytes, Gorilla gorilla, Ornithorhynchus anatinus, Artibeus jamaicensis* and *Mus musculus*) were examined in detail to assess the intraspecific variability of this locus within different mammalian species. Table 3 indicates that variation of the *FMR1* repeat locus is clearly not limited to humans. Platypi, artibean bats and chimpanzees all possess polymorphic *FMR1* CGG repeat loci. In contrast, the orthologous locus in gorilla and mouse appears to be virtually static in terms of polymorphic potential. Comparisons of intraspecific variability with the substructure of the repeat suggest a model in which both the length of the longest tract of uninterrupted CGG repeats as well as the position of the interruptions within the repeat are important factors in determining stability at this locus

Discussion

Evolution of the primate *FMR1* CGG repeat. Our survey of the *FMR1* CGG repeat among non-primate mammals reveals that the majority of mammalian orders have short uninterrupted CGG repeats at an orthologous position in the 5' UTR of the gene. Parsimony analysis of the various non-primate mammalian *FMR1* CGG repeat sequences, in conjunction with the observation of repeat length and content among other mammalian orders, predicts that the mammalian ancestral state of the repeat was short and without interruption — $(CGG)_{4-9}$. In contrast, there has been a tendency among primate mammals to increase the overall length of the *FMR1* CGG repeat (Table 2). Parsimony analysis within the Infraorder Catarrhina[21] suggests that increases in CGG repeat length (beyond ~20 repeats) have occurred at least three times independently during the course of primate evolution. A specific type of interruption is associated with each expansion. Among hylobatids (the lesser apes), pongids/hominids (man and the greater apes), and cercopithecids (old world monkeys represented by baboons and rhesus monkey) repeat lengths increased with the simultaneous addition of specific interspersions; namely AGG, CGA and CGGG, respectively (Fig. 2).

At the time of divergence between the hominoids and cercopithecoids, postulated to have occurred some 25–31 million years ago (mya)[21], there likely already existed a difference between the sequences in these two lineages. The presence of at least a single AGG interspersion among all hylobatids, hominids and pongid species, tested, supports a reconstruction of ancestral states by parsimony which predicts that a single AGG trinucleotide was present in the early hominoid ancestor (Fig.2). Within the pongid/hominid clade, the acquisition of additional AGG interruptions appears to have occurred relatively early, generating the AGGCGGAGG motif which is still found among three extant species (Table 2 and Fig. 2). In man and the bonobo, this sequence appears to have been lost, perhaps due to a dramatic increase in instability at this locus or to genetic drift of rare variants in the founding populations of these species. The chimpanzee alleles may be at a point of transition, with respect to this sequence, as only half of all alleles have retained the ancestral AGGCGGAGG motif (Table 2). While AGG interruptions have been used by both hominids and pongids to disrupt the continuity of the *FMR1* CGG repeat, mem-

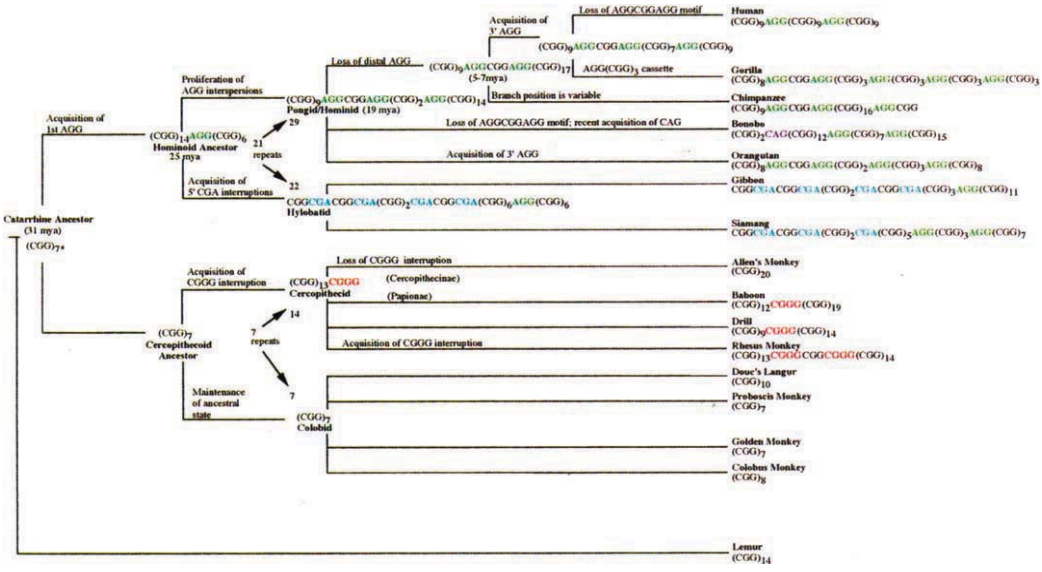

Fig 2 Hypothetical catarrhine evolution of the *FMR1* CGG repeat. Majority rule (>50%) consensus phylogenetic tree of catarrhine *FMR1* CGG repeat and flanking sequences was determined by parsimony analysis with the lemur (*E. macaco*) as the specified outgroup. Both branch-and-branch bound and heuristic searches were performed. The most likely ancestral sequence, as determined by the character change reconstruction PAUP option, is depicted at each branchpoint in the cladogram. Due to the extreme length variability of compared sequences, ambiguous character states were not considered in the reconstruction of ancestral sequences. Divergence times are shown in brackets and are based on a recent comparative determinant analysis[21]. *Although the consensus ancestral sequence for the catarrhine ancestor is short (7 CGG repeats) and without interruption, parsimony also predicts a short repeat with an interruption at the distal end of the repeat, which is polymorphic with respect to interspersion type. Similar results were obtained using derived ancestral sequences for catarrhine primate families and *M. domesticus* and *O. anatinus* as specified

bers of the Hylobatidae appear to have evolved a different strategy. CGA interruptions proliferated 5' of the AGG interspersion to generate large and highly interrupted alleles (Table 2 and Fig. 2, siamang and gibbon). The acquisition of multiple CGA interruptions is predicted to have occurred concurrently with the proliferation of the hominid/pongid AGG interspersions (~20 mya, Fig. 2).

In contrast to the hominoid line of evolution, repeats in the cercopithecoid branch appear to be much less complex. Only one lineage, members of the cercopithecid tribe Papionae (Fig. 2, baboon, drill and rhesus monkey) acquired CGGG interruptions in association with an enlargement of repeat size. While the cercopithecids show a dramatic increase in repeat size, the colobine branch demonstrates simple and short *FMR1* CGG repeats, presumably representing the maintenance of the ancestral repeat state (Fig. 2).

Proliferation of interspersions: a role for recombination-based mechanisms? Once interruptions disrupted the homogeneity of the repeat, it appears that in hominoids, and to a lesser extent among the cercopithecids, the interspersions proliferated rapidly. Furthermore, the acquisition of interspersions shows directional bias. Among the chimps, orangutan and gorilla, for example, parsimony analysis predicts that AGG interruptions were added distal to the first three primary AGG interspersions, resulting in a 5' tract of pure CGG repeats which is largely invariant (ranging from 8–11 triplets among the different species, Table 2 and Fig. 2). These observations are consistent with recent findings in humans where mutational change is biased toward the 3' end of the repeat[10, 14, 15]. Hylobatids (gibbon and siamang), in contrast, appear to have proliferated their CGA interspersions at the 5' end of the repeat suggesting an opposite polarity in the variation of the *FMR1* CGG repeat locus in these species.

It is highly unlikely that the multiplicity of interspersions in hominoid *FMR1* CGG repeats arose by a series of independent mutational events. The rapid proliferation of interspersions over a short evolutionary period and the homogeneity of the type of interspersion are incompatible with known rates of mutation and random mutation theory[22]. More likely is that proliferation of interspersions has occurred primarily by a mechanism of recombination such as unequal crossing over or by gene conversion as has been suggested for VNTR evolution[23–25]. Intergenerational transmission studies of repeat length in human pedigrees in which substantial contractions and enlargements of normal alleles have been documented[26, 27] further support such a model. The implication that unequal crossing-over or gene conversion has played a part in the evolution of the *FMR1* CGG repeats is unexpected as most previous models promote polymerase slippage as the primary force of microsatellite evolution[25, 28–31].

The existence of higher-order periodicity of AGG interspersions in the gorilla *FMR1* CGG repeat sequence and, as has been suggested for the human *FMR1* repeat[13], may be taken as strong evidence that there has been a relatively high rate of recombination at this locus[32], for replication slippage alone would not be capable of generating such a highly organized structure within a short tandem repeat[29, 32]. Larger alleles, at least in gorilla and man, historically may have been

Table 3 Variability of the *FMR1* CGG repeat among different mammalian populations

Species	Total # of alleles	%Hetero-zygosity	Mean length of longest CGG	Sequence content of CGG repeat	# of alleles	Repeat length
Mus musculus	15	0	6.0	$(CGG)_6CGA(CGG)_2$ *	15	9
Orinthorhyncus anatinus	10	CGG=66.0 AGG=74.0	6.0 +/– 0.4	$(AGG)_6/(CGG)_5$ **	1	5
				$(AGG)_7/(CGG)_5$	2	5
				$(AGG)_9/(CGG)_5$	1	5
				$(AGG)_6/(CGG)_6$	1	6
				$(AGG)_8/(CGG)_6$	3	6
				$(AGG)_9/(CGG)_7$	1	7
				$(AGG)_7/(CGG)_8$	1	8
Gorilla gorilla ***	16	30.2	8.0	$(CGG)_8AGGCGGAGG(CGG)_3$-$AGG(CGG)_3AGG(CGG)_3$	13	22
				$(CGG)_8AGGCGGAGG(CGG)_3$-$AGG(CGG)_3AGG(CGG)_3AGG(CGG)_3$	3	26
Artibeus jamaicensis	9	99.9	8.4 +/– 0.4	$(CGG)_8CAG(CGG)_{10}$	1	19
				$(CGG)_7CAG(CGG)_7$	1	15
				$(CGG)_9CAG(CGG)_8$	1	15
				$(CGG)_8CAG(CGG)_{11}$	1	20
				$(CGG)_8CAG(CGG)_7$	1	16
				$(CGG)_8CAG(CGG)_8$	1	17
				$(CGG)_6CAG(CGG)_7$	1	14
				$(CGG)_6CAG(CGG)_8$	1	15
				$(CGG)_5CAG(CGG)_8$	1	14
Pan troglodytes	10	99.9	15.2 +/– 4.8	see Table 2	11(1)	20–38

Polymorphic variability both in sequence content and length is shown for five diverse mammalian species at the *FMR1* CGG repeat tract. (For ascertainment of individuals within each population see Methods.) Repeat length indicates the total number of triplets observed, including CGG repeats and interspersions. Mean length of longest CGG repeat defines the triplet repeat length of the longest tract of uninterrupted CGG's. *Only five individuals (3 wild-type derived strains and 2 inbred strains) were sequenced in their entirety, the remaining 10 alleles were assessed by length variation only. **The platypus contains an AGG repeat 12 bp upstream from its CGG repeat, heterozygosity was assessed for both triplets independently. ***Within the gorilla population, all three subspecies were represented: 11 alleles from *G. g. gorilla*, 4 alleles from *G. g. beringei* and 1 allele from *G. g. graueri*.

constructed in a stepwise fashion, with the position of AGG interruptions demarcating the cassette unit of recombination. It is interesting that in gorillas the core sequence $AGG(CGG)_3$ is different from the repetitive core sequence found in humans, $AGG(CGG)_9$. These differences in substructure may suggest simple stochastic differences in the position and occurrence of interspersions within the ancestral lineages or could reflect subtle differences in the mechanics of recombination between species such as differences in match length requirements[32, 33]. An alternative explanation, excluding the involvement of recombination-based processes, may be that the unit of slipped-strand mispairing extends beyond the triplet repeat to include a larger array (ie. $AGG(CGG)_3$ or $AGG(CGG)_9$). Such a model, we believe, is less likely since it would require two different match length requirements for slippage at the same locus within the same familiy of species[33], the larger being less thermodynamically stable[34].

Recent investigations into the substructure of the human *FMR1* CGG repeat[10, 13–15] suggest a dichotomy in the type of variation found at this locus. The vast majority of human variation involves differences of a single CGG repeat occuring at the 3' end of the repeat[14, 15]. A second class of variation, accounting for less than 5% of all differences in *FMR1* CGG repeat structure (E.E.E. *et al.*, submitted), involve reiterations of the $AGG(CGG)_9$ core sequence[10, 13–15]. Recombination processes such as unequal chromosome exchange or gene conversion have been invoked to account for these changes. In humans, then, it would appear that intraspecific variation of AGG interruptions can result from recombination, replication slippage or a combination of these two events. Among other species, such as bonobo and chimpanzee, no higher-order structure for the *FMR1* CGG repeat has been observed (Table 2). Populations which lack alleles with higher-order substructure may be the result of both relatively high rates of recombination and replication slippage, as has been suggested for other short tandem repeats with a complicated cryptic substructure[35]. In such a scenario, both forces could be envisioned to act antagonistically. Recombination could reduce the length of the longest tract required for slipped-strand mispairing by the fortuitous interjection of an interspersion after an unequal exchange, while replication slippage might disrupt the ability of interspersions to align during meiotic exchange, forcing out-of-register recombination events. Such an antagonism would clearly be dynamic, allowing for the generation of a diverse set of alleles ranging from no interspersions to many different interruptions without higher-order periodicity. This may explain the surprisingly high level of heterozygosity of chimpanzee alleles when compared to the human *FMR1* CGG repeat locus which retains some higher-order periodicity (see below).

Intraspecific variability of the *FMR1* CGG repeat. Our survey of *FMR1* CGG intraspecific variability within five diverse mammalian populations indicates that polymorphism as measured by percentage heterozygosity can not be predicted based solely on the total length of the repeat (Table 3). Previously, it was shown that the polymorphic stability of dinucleotide and trinucleotide repeats correlates well with the longest tract of pure repeats[14, 19, 36]. Some cross-species sequence comparisons support this model. Chimpanzee *FMR1* CGG repeat alleles, for example, appear

more heterozygous than human alleles (99.9% vs. 67.7%). Although humans and chimps both have on average similar numbers of interspersions per allele (1.8 +/– 0.4 and 2.6 +/– 1.8 AGG interruptions, respectively $P = 0.20$), the position of interspersions differs radically between these two species. In humans, AGG interspersions are symmetrically located within the tract occurring once every 9/10 CGG repeats, while among chimpanzees this distribution is without periodicity. The net effect is that the mean length of the longest tract of pure CGG repeats among chimps (15.2 +/– 4.8) is greater (t test, $P<0.05$) than the longest tract in humans (11.3 +/– 3.6). This longer tract of repeats in chimpanzee may be more apt to undergo replication slippage to generate new alleles, accounting for the high level of heteozygosity in this species.

Examination of polymorphic stability of the *FMR1* CGG repeat in other species, however, suggests that this model may be too simplistic. For example, platypi demonstrate modest polymorphic variability, with their CGG repeat number ranging from 5–8 (mean 6.0 +/– 0.4 repeats; Table 3). In contrast, both wild-type and inbred strains of *Mus musculus* showed no polymorphic variability, although the mean length of the longest tract of pure repeats (6.0 +/– 0.0) was similar to the platypus (Table 3). In addition, a survey of the *FMR1* CGG repeat locus in distantly related gorilla subspecies revealed no variation in the longest tract of pure repeats which was, in this case, 8.0 CGG repeat units. Once again, this observation contrasts with the extensive polymorphic variability observed within the phyllostomid bat species, *Artibeus jamaicensis*, in which the average length of the longest tract of uninterrupted repeats was 8.4 +/– 0.4 (Table 3). Thus, in these species a reliable prediction of polymorphic potential could not be made based solely on the longest tract of pure repeats. This lack of correlation may reflect differences in gene flow and heterogeneity within founder populations. However, based on mitochondrial D-loop investigations, it is highly unlikely that the different subspecies of gorilla or that the different strains and species of mouse, considered in our analysis, have been subjected to similar evolutionary genetic bottlenecks and founder effects (OAR unpublished data)[37]. It is interesting that the polymorphic and non-polymorphic *FMR1* CGG repeat species are more apt to be distinguished by the position of the longest tract of pure repeats relative to the interspersions, rather than by length alone. Among the gorilla and mouse, both relatively static in terms of length variation, the longest tract of pure repeats is located 5' to the position of the first interruption. In contrast, among artibean bats and humans, the longest tract of CGG repeats appears to be located 3' to the CAG and AGG interruptions respectively. Both species, similarly, are highly polymorphic. In man, bias in the polarity of mutational change has already been documented, with most differences occurring at the 3' end of the repeat[14,15]. It appears that polarized variability, like the CGG repeats themselves, has been conserved throughout mammalian evolution.

Implications. The conservation of the CGG repeat for over 150 million years of evolution provides strong evidence of its functionality in the 5' UTR of *FMR1*. Based on comparative sequencing and parsimony analysis of this locus from 44 mammalian species, we propose a model for how the repeat has evolved. The mammalian ancestral state was short (approximately 4–9 repeats) and without interruption. In primates, there has been a tendency to expand the length of the repeat beyond 20 repeats. Accompanying primate enlargement of the *FMR1* CGG repeat, there has been a polarized proliferation of lineage-specific interspersions. In some species, such as man and gorilla this event has involved the vectorial and incremental addition of smaller arrays demarcated by an interruption. Our data support a model in which both recombination-based mechanisms, such as unequal chromatid exchange or gene conversion, and replication slippage have been operative during the evolutionary construction of larger repeats. These results, furthermore, confirm the need to consider both sequence content and length variation in any attempt to model intraspecific and interspecific variability at triplet repeat loci.

The propensity for primates to enlarge repeats has been reported for other triplet repeat loci[28,38]. Does this generalized phenomenon of primate expansion serve a functional role or does it represent a deficiency in some mechanism that limits the length of triplet repeats? If length control is a molecular shortcoming among primates, then the recent proliferation of interspersions in lineages, which have increased their overall repeat length, may be seen as compensatory, from an evolutionary perspective. These studies of intraspecific variability of this locus within diverse mammalian populations clearly confirm the stabilizing effect of interruptions. Since most mammalian species posess short and/or highly interrupted CGG repeats, it is likely that CGG repeat hyperexpansion and its associated genetic disease, the fragile X syndrome, is a phenomenon restricted to man with a possible extension to our closest relatives (bonobo and chimpanzee).

Methods

DNA samples. DNA samples were prepared or obtained from several sources. DNA from *Panthera leo* (gir lion), *Herpestes ichneumon* (mongoose), *Microtus agrestis* (vole), *Ammospermophilus harrisi* (ground squirrel), *Carrolia perspicillata* (short-tailed fruit bat), *Artibeus jamaicensis* (Jamaican fruit-eating bat), *Myotis yumaneni* (Indian bat), *Euarctos americanus* (polar bear), *Lemur fulvianus* (brown lemur) and *Zalophus californianus* (seal) were obtained from lymphoblast cell lines obtained from Dr. T.C. Hsu, Houston, Texas. Similarly, DNA from *Stenella plagiodon* (dolphin) was obtained from a kidney fibroblast cell line, Sp1K, from the American Type Culture Collection. DNA was prepared from cell lines using a standard lysis protocol. Briefly, 2 x 10^5 cells were transferred to a microcentrifuge tube, washed with 500 µl of Hank's balanced salt solution, and resuspended in 40 µl of PCR lysis buffer (1X exo(–) *Pfu* Buffer {20 mM Tris-Cl pH 8.75, 10 mM KCl, 10 mM $(NH_4)SO_4$, 2mM $MgCl_2$, 0.1% Triton X-100, 100 µg ml^{-1} BSA}, 0.45% Tween-20, 0.45% Nonidet P-40) with 2 µl of 1 mg ml^{-1} proteinase K. Samples were lysed for 1 h at 56 °C and then heated to 94 °C for 5 min to inactivate the proteinase K. Cellular debris was pelleted in a microcentrifuge (14,000 rpm for 30 s) and samples were stored at –20 °C for future use.

The remaining DNA samples were obtained directly from tissue samples. DNA from *Bos taurus* (cow), *Felis domesticus* (domestic cat), *Ovis aries* (sheep), *Canis familiaris* (domestic dog), *Mesocricetus auratus* (hamster), *Sus scrofa* (pig), *Rattus norvegicus* (rat), *Callithrix* (marmoset) and 3 individuals of *Pongo pygmaeus* (orangutan) was commercially available from

BIOS laboratories. DNA from *Chaetophractus villosus* (South American hairy armadillo) was obtained from Dr. Mark Springer, University of California, Riverside. Similarly, primate DNA samples for *Hylobates lar* (gibbon), *Macaca rhesus* (rhesus monkey), *Saimiri sciureus* (squirrel monkey), and *Papio* (baboon) were kindly provided by Dr. Xiangwei Wu, University of Princeton. Primate DNA samples for *Varecia variegata* (black and white ruffled lemur), *Colobus guereza* (colobus monkey), *Rhinopithecus roxellana* (golden monkey), *Nasalis larvatus* (proboscis monkey), *Allenopithecus nigroviridis* (Allen's monkey), *Pygathrix nemaeus* (Douc's langur), *Hylobates syndactylus* (siamang), *Mandrillus leucophaeus* (drill), 2 unrelated samples of *Nycticebus coucang* (slow loris) and 4 unrelated individuals of *Pan paniscus* (bonobo) were obtained from the collection of O.A.R.

Within the five populations chosen for assessment of *FMR1* CGG repeat polymorphic variability, individuals were selected, wherever possible, to maximize genetic diversity. A collection of DNA samples from 9 gorillas (4 females and 5 males) representing all three subspecies (*Gorilla gorilla gorilla, Gorilla g. beringei* and *Gorilla g. graueri*) was assembled in which the individuals were known to be distantly related based on mitochondrial D loop data (O.A.R. *et al.*, in preparation). DNA for the platypus (*Ornithorhynchus anatinus*) was obtained from individuals that originated from diverse river localities in Australia and were unlikely to be closely related (2 samples from the Brisbane River, South Queensland; 4 samples from Shoothaven River, New South Wales; 1 sample from Victoria; 1 sample from the Thedbo River, New South Wales; and 1 sample from the DuckMaloi River, New South Wales). Mouse DNA samples (*Mus caroli, Mus musculus* and *Mus spretus*) were all purchased from the Jackson Laboratory, Bar Harbor. Survey of variability in *Mus musculus* was based on 15 different stocks of mice (3 wild-type derived strains {*Mus musculus poschiavinus* (Tirano), *Mus musculus poschiavinus* (Zalende), and *Mus musculus musculus* (Czech II)} and 12 inbred strains (AKR/J, SJL/WtBm, SWR/J, BalbC/CJ, C3H/HeJ, C57BL/10J, C57BL/6By, C57BL/6J, C57BR/cdJ, CBA/J, DBA/1J and DBA/2J)[37]. Thirteen Jamaican fruit-eating bats (*Artibeus jamaicensis*) were tested for *FMR1* CGG repeat variability. Samples were collected by Dr. J. T. Baker, Lubbock, from bat colonies located in diverse geographic locations. DNA samples from *Artibeus obscura* and *Micronycteris hirsuta* were also obtained from Dr. Baker's collection. The chimpanzee (*Pan troglodytes*) samples were derived from a random collection of DNAs from both BIOS laboratories as well as the Yerkes Primate Center. Other samples from Yerkes included one gibbon (*Hylobates lar*), one gorilla (*Gorilla gorilla*), one orangutan (*Pongo pygmaeus*) and one bonobo (*Pan paniscus*). Finally, for a comparison with humans, 60 DNA samples were obtained from random blood donors in Houston, Texas in which the ethnic background was known (15 Caucasian, 15 African-American, 15 Hispanic, and 15 Asian).

PCR analysis. Amplification of the FMR1 CGG repeat from the various species was performed as previously described[39], using a modified protocol which replaces *Taq* (Roche) with exo(–) *Pfu* (Stratagene)[40]. Typically, PCR reactions were performed as previously reported[40] with the exception that the annealing temperature was reduced from 65 °C to 55 °C to allow for the generation of lower stringency PCR products from the different species. Size estimates of products were made on a 5% denaturing sequencing gel relative to an M13 sequencing ladder and/or by comparison of digested PCR products to a 10-bp ladder (Stratagene) on a 3.0% Metaphor™ agarose gel (FMC Bioproducts)[5,40].

Sequencing analysis. PCR products were subcloned both into a blunt ended cloning vector, pCRScript KS+ (Stratagene), and into a TA cloning vector, pCR II (Invitrogen), using manufacturers' suggested protocols. Ligation products were transformed into XL1-Blue supercompetent cells (Stratagene), and transformants were screened by PCR to identify clones which contained inserts of correct length. Positive clones were sequenced with M13 forward primers and fluorescently labelled dideoxy terminators from a single-strand template using an automated DNA sequencer (ABI 373). Multiple clones from independent ligations were analysed to determine the identity of the CGG repeats and their flanking sequence. To confirm the sequence of longer clones, particularly among the hominoids, direct sequencing of PCR products was performed as described[15].

Evolutionary genetic analysis. PAUP (v. 3.1.1, Illinois Natural History Survey) and MacClade (v. 3.1, Sinauer Associates) were employed to derive relationships among the various primate sequences. The sequences were aligned at the nucleotide level using GeneWorks software, version 2.3.1(Intelligenetics) and parsimony analysis was performed. During alignment, parameters were chosen (low gap penalty assignment and an encoded data set) which allowed for the occurrence of slippage events (changes in length of one or two triplets). The non-default phylogenetic option was used to assume hard polytomies (mutliple speciation events may be allowed). As the differential repeat lengths are problematic for alignment of the sequences, a strategy of encoding the data (by triplet repeat length and content) was employed in addition to use of the unencoded data. The results were identical with and without defining the interspersion types as equate macros. Uninformative sites were excluded from the analysis. Branch-and-bound and heuristic searches were performed. Strict and majority-rule consensus methods were used to construct a phylogenetic tree of the *FMR1* CGG repeat sequences. Initially, cercopithecoid and hominoid sequences were considered, using *E. macaco* and *N. coucang* as outgroups (n=17). The most likely ancestral sequences were determined at the branch points in the cladogram using the character change reconstruction option. Using the derived ancestral states at each branchpoint, a second round of parsimony analysis was performed using *M. spretus* and *O. anatinus* as outgroup sequences (n=6). Due to the extreme sequence heterzygosity of some species (*P. troglodytes* and *P. paniscus*), different allele variants were added in subsequent parsimony analyses to determine their effect on branching points of the phylogenetic tree. Similar parameters and ancestral-state determination methods were employed in the parsimony analysis of the non-primate mammalian *FMR1* CGG repeat sequences.

Acknowledgements

We are grateful to J.T. Baker, T.C. Hsu, N. Gemmell and U. Arnason for kindly providing DNA from different species for sequence analysis. We would like to thank M.Y. Eichler and the BCM Human Genome Center Sequencing Core (HG00210) for technical assistance. This work was supported, in part, by NIH grant (HD29256) to D.L.N. S.T.W. is an investigator of the Howard Hughes Medical Insitute.

Received 12 June; accepted 28 August 1995.

1. Jones, C. *et al.* Association of a chromosome deletion syndrome with a fragile site within proto-oncogene CBL2. *Nature* **376**, 145–149 (1995).
2. Parrish, J.E. *et al.* Isolation of a GCC repeat showing expansion in FRAXF, a fragile site distal to FRAXA and FRAXE. *Nature Genet.* **8**, 229–235 (1994).
3. Nancarrow, J.K. *et al.* Implications of FRA16A Structure for the mechanism of chromosomal fragile site genesis. *Science* **264**, 1938–1941 (1994).
4. Knight, S.J. *et al.* Trinucleotide repeat amplification and hypermethylation of a CpG island in FRAXE mental retardation. *Cell* **74**, 127–134 (1993).
5. Fu, Y.H. *et al.* Variation of the CGG repeat at the fragile X site results in genetic instability: resolution of the Sherman paradox. *Cell* **67**, 1047–1058 (1991).
6. Verkerk, A.J.M.H. *et al.* Identification of a gene (*FMR-1*) containing a CGG repeat coincident with a breakpoint cluster region exhibiting length variation in fragile X syndrome. *Cell* **65**, 905–914 (1991).
7. Yu, S. *et al.* Fragile X genotype characterized by an unstable region of DNA. *Science* **252**, 1179–81 (1991).
8. Oberle, I. *et al.* Instability of a 550-base pair DNA segment and abnormal methylation in fragile X syndrome. *Science* **252**, 1097–102 (1991).
9. Ashley, C.T. *et al.* Human and murine *FMR-1*: evidence for alternative splicing and translational initiation downstream of the CGG-repeat. *Nature Genet.* **4**, 244–251 (1993).
10. Snow, K., Tester, D.J., Kruckenberg, K.E., Schaid, D.J. & Thibodeau, S.N. Sequence analysis of the fragile X trinucleotide repeat: implications for the origin of the fragile X mutation. *Hum. molec. Genet.* **3**, 1543–1551 (1994).
11. Richards, R.I. *et al.* Fragile X syndrome: genetic localisation by linkage mapping of two microsatellite repeats FRAXAC1 and FRAXAC2 which immediately flank the fragile site. *J. Med. Genet.* **28**, 818–23 (1991).
12. Pieretti, M. *et al.* Absence of expression of the *FMR-1* gene in fragile X syndrome. *Cell* **66**, 817–822 (1991).
13. Hirst, M.C., Grewal, P.K. & Davies, K.E. Precursor arrays for triplet repeat expansion at the fragile X locus. *Hum. molec. Genet.* **3**, 1553–1560 (1994).
14. Eichler, E.E. *et al.* Length of uniterrupted CGG repeats determines stability in the *FMR1* gene. *Nature Genet.* **8**, 88–94 (1994).
15. Kunst, C.B. & Warren, S.T. Cryptic and polar variation of the fragile X repeat could result in predisposing normal alleles. *Cell* **77**, 853–861 (1994).
16. Chung, M.-Y. *et al.* Evidence for a mechanism predisposing to intergenerational CAG repeat instability in spinocerebellar ataxia type I. *Nature Genet.* **5**, 254–258 (1993).
17. Snow, K. *et al.* Analysis of a CGG sequence at the FMR-1 locus in the fragile X families and in the general population. *Am. J. hum. Genet.* **53**, 1217–1228 (1993).
18. Richards, R.I., Holman, K., Yu, S. & Sutherland, G.R. Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins. *Hum. molec. Genet.* **2**, 1429–1435 (1993).
19. Riggins, G.J. *et al.* Human genes containing polymorphic trinucleotide repeats. *Nature Genet.* **2**, 186–191 (1992).
20. Perrin-Pecontal, P., Gouy, M., Nigon, V.-M. & Trabuchet, G. Evolution of the primate beta-globin gene region: nucleotide sequence of the delta-beta-globin intergenic region of gorilla and phylogenetic relationships between African apes and man. *J. mol. Evol.* **34**, 17–30 (1992).
21. Bauer, K. Primate phylogeny studied by comparative determinant analysis. *Exp. Clin. Immunogenet.* **10**, 56–60 (1993).
22. Miyamoto, M., Slightom, J. & Goodman, M. Phylogenetic relations of humans and apes from DNA sequences in the psi eta-globin region. *Science* **238**, 369–373 (1987).
23. Jeffreys, A.J. *et al.* Complex gene conversion events in germline mutation at human minisatellites. *Nature Genet.* **6**, 136–145 (1994).
24. Wolff, R.K., Plaetke, R., Jeffreys, A.J. & White, R. Unequal crossingover between homologous chromosomes is not the major mechanism involved in the generation of new alleles at VNTR loci. *Genomics* **5**, 382–384 (1989).
25. Levinson, G. & Gutman, G.A. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Molec. Biol. Evol.* **4**, 203–221 (1986).
26. Arinami, T., Asano, M., Kobayashi, K., Yanagi, H. & Hamaguchi, H. Data on the CGG repeat at the fragile X site in the non-retarded Japanese population and family suggest the presence of a subgroup of normal alleles predisposing to mutate. *Hum. Genet.* **92**, 431–436 (1993).
27. Macpherson, J. *et al.* Unusual (CGG)n expansion and recombination in a family with fragile X and DiGeorge syndrome. *J. med. Genet.* **32**, 236–239 (1995).
28. Rubinsztein, D.C. *et al.* Mutational bias provides a model for the evolution of Huntington's disease and predicts a general increase in disease prevalence. *Nature Genet.* **7**, 525–530 (1994).
29. Stephan, W. & Cho, S. Possible role of natural selection in the formation of tandem-repetitive noncoding DNA. *Genetics.* **136**, 333–341 (1994).
30. Tachida, H. & Iizuka, M. Persistence of repeated sequences that evolve by replication slippage. *Genetics.* **131**, 471–478 (1992).
31. Walsh, J.B. Persistence of tandem arrays: implications for satellite and simple-sequence DNAs. *Genetics.* **115**, 553–567 (1987).
32. Smith, G.P. Evolution of repeated DNA sequences by unequal crossover. *Science* **191**, 528–535 (1976).
33. Stephan, W. Tandem-repetetive noncoding DNA: forms and forces. *Molec. Biol. Evol.* **6**, 198–212 (1989).
34. Gacy, A., Goellner, G., Juranic, N., Macura, S. & McMurray, C. Trinucleotide repeats that expand in human disease form hairpin structures in vitro. *Cell* **81**, 533–540 (1995).
35. Jacobson, D.P., Schmelling, P. & Sommer, S.S. Characterization of the patterns of polymorphism in a "cryptic repeat" reveals a novel type of hypervariable sequence. *Am. J. hum. Genet.* **53**, 443–450 (1993).
36. Weber, J.L. Informativeness of human $(dC\text{-}dA)_n.(cG\text{-}dT)_n$ polymorphisms. *Genomics* **7**, 524–530 (1990).
37. Tucker, P.K., Lee, B.K., Lundrigan, B.L. & Eicher, E.M. Geographic origin of the Y chromosomes in "old" inbred strains of mice. *Mammalian genome* **3**, 254–261 (1992).
38. Rubinsztein, D.C., Leggo, J., Amos, W., Barton, D.E. & Ferguson-Smith, M.A. Myotonic dystrophy CTG repeats and the associated insertion/deletion polymorphism in human and primate populations. *Hum. molec. Genet.* **3**, 2031–2035 (1994).
39. Deelen, W., Bakker, C., Halley, D.J.J. & Oostra, B.A. Conservation of CGG region in FMR1 gene in mammals. *Am. J. med. Genet.* **51**, 001–008 (1994).
40. Chong, S.S., Eichler, E.E., Hughes, M.R. & Nelson, D.L. Robust amplification of the fragile X syndrome CGG repeat using *Pfu* polymerase: ethidium bromide detection of normal and premutation alleles. *Am. J. med. Genet.* **51**, 522–526 (1994).