THE ROYAL SOCIETY

# Microsatellite and trinucleotide-repeat evolution: evidence for mutational bias and different rates of evolution in different lineages

**David C. Rubinsztein**[1*], **Bill Amos**[2] **and Gillian Cooper**[2]

[1]*Department of Medical Genetics, Cambridge Institute for Medical Research, Wellcome/MRC Building, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2XY, UK*
[2]*Department of Zoology, Downing Street, Cambridge CB2 3EJ, UK*

Microsatellites are stretches of repetitive DNA, where individual repeat units comprise one to six bases. These sequences are often highly polymorphic with respect to repeat number and include trinucleotide repeats, which are abnormally expanded in a number of diseases. It has been widely assumed that microsatellite loci are as likely to gain and lose repeats when they mutate. In this review, we present population genetic and empirical data arguing that microsatellites, including normal alleles at trinucleotide-repeat disease loci, are more likely to expand in length when they mutate. In addition, our experiments suggest that the rates of expansion of such sequences differ in related species.

**Keywords:** microsatellite; trinucleotide repeat; triplet repeat; Huntington's disease; mutation

## 1. INTRODUCTION

Microsatellites are stretches of repetitive DNA found in eukaryotic genomes. The repeat units at these loci comprise one to six bases and repeat number frequently varies at a given locus. This high polymorphism rate and the accessibility of these markers to PCR amplification has led to microsatellites being used as major tools for genetic mapping (Weissenbach *et al*. 1992), studies of human (Bowcock *et al*. 1994) and animal diversity (Bruford & Wayne 1993) and for forensic investigations (Jeffreys *et al*. 1992). An understanding of the mutational and evolutionary processes of such loci can aid data interpretation from many of these applications. In addition, trinucleotide repeats are microsatellites, and a study of their mutational features can help to explain the prevalences of different diseases in different populations and how these mutations originate.

At the time we initiated the work reviewed in this paper, most workers considered that microsatellites mutate symmetrically, with expansion and contraction mutations occurring at equal frequencies (Di Rienzo *et al*. 1994; Goldstein *et al*. 1995*a*; Kruglyak *et al*. 1998). Furthermore, it has been assumed that such loci show similar rates of evolution in related populations and species (Goldstein *et al*. 1995*b*; Pollock *et al*. 1998). These assumptions underlie many of the formulae that have been used to compute genetic distances (Goldstein *et al*. 1995*a*). In this review, we will consider evidence which challenges these two important assumptions.

* Author for correspondence (dcrl000@cus.cam.ac.uk).

## 2. WILD-TYPE HUNTINGTON'S DISEASE CAG REPEATS PROVIDE CLUES ABOUT MICROSATELLITE BEHAVIOUR

Huntington's disease (HD), like most other diseases caused by abnormal expansions of trinucleotide-repeat tracts, shows anticipation—the age at onset in affected individuals in a family tends to decrease in successive generations (Duyao *et al*. 1993). This can be explained by the correlation between increasing CAG repeat number on disease chromosomes with earlier age-at-onset of symptoms and the overall tendency for the CAG repeat mutation to increase in size in successive generations, which is particularly marked in male transmissions. The overall mutational bias in favour of expansions seen in HD alleles is also a feature of disease alleles at most trinucleotide-repeat loci (for a review, see Rubinsztein & Amos 1998).

The mutational processes of mutant trinucleotide-repeat disease alleles is comparatively easy to assess by studying disease pedigrees, since disease alleles have high mutation rates. However, we were interested to study wild-type alleles to understand the origins of trinucleotide-repeat mutations. Since these do not mutate frequently, family studies were impractical. Thus, we used population genetic approaches.

Huntington's disease varies in prevalence. For example, it is comparatively common in East Anglia (1/10 000) but rare in Japan ($< 1/1\,000\,000$). The range of normal alleles is from 8–35 CAG repeats and there are no data which suggest that these alleles are associated with variable genetic fitness. Accordingly, we expected that the distributions of normal alleles at the HD locus would

 

reflect mutational processes and random genetic drift. We typed normal HD gene alleles in a panel of human populations and in non-human primates. The human allele distributions showed three features.

(i) There was a relationship between the proportion of long normal alleles in a population and its frequency of HD. For instance, a significantly greater number of long normal alleles were seen in East Anglian people compared with Japanese. This relationship was confirmed in HD by the Hayden group and has also been described for myotonic dystrophy, dentatorubral–pallidoluysian atrophy, among other trinucleotide-repeat diseases (for a review, see Rubinsztein & Amos 1998). This suggests that the majority of new mutations at trinucleotide-repeat disease loci originate from the upper end of the normal allele length distribution.

(ii) The allele distributions among all human populations showed a positive skew in that there was an excess of alleles with longer repeat lengths than the modal length (Rubinsztein *et al*. 1994). Such a positive skew was one of the first clues that led us to consider the possibility that microsatellite mutations are not symmetrical with respect to length changes. Indeed, this type of skewed distribution has been found in the hypermutable minisatellites such as CEB1, and these have been shown empirically to have an excess of expansion versus contraction mutations (Jeffreys *et al*. 1994; Monckton *et al*. 1994; Vergnaud *et al*. 1991).

(iii) The HD CAG repeats in a large panel of non-human primates appeared to be shorter than those seen in human populations. Since all primates share a common ancestor, the most parsimonious explanation for this finding was that the repeats have expanded in length in the human lineage.

We performed a series of computer simulation experiments, in order to explore possible mechanisms leading to the positively skewed allele length distributions in all the populations we studied and to account for the expansion of allele lengths in the human lineage. Our empirical data were best explained by a mutational model that incorporated mutational bias in favour of expansions as a key component (Rubinsztein *et al*. 1994). This model predicts that the HD CAG repeats would continue to expand over an evolutionary time-scale. If not stopped, such a process would lead to ever-increasing disease prevalence, in the absence of selection.

## 3. MICROSATELLITES SHOW MUTATIONAL BIAS AND DIFFERENT EXPANSION RATES IN DIFFERENT LINEAGES

The conclusion that the HD triplet repeats have an inherent tendency to expand with time, potentially leading to ever-increasing disease incidence, is disturbing and has resulted in some controversy. Therefore, we decided to expand the above approach and consider the features of microsatellites in general, in order to see whether the triplet repeats at the HD locus were typical or exceptional. We examined allele frequency distributions of more than 300 microsatellite loci in humans and found that positively skewed length distributions were more common than all other distributions pooled together (J. Swinton and B. Amos, unpublished data). Similar findings have been reported by Farrall & Weeks (1998).

A panel of 44 polymorphic human microsatellite loci were studied in humans, chimpanzees, gorillas, orang-utans, baboons, macaques and marmosets (Rubinsztein *et al*. 1995). These included 'neutral' dinucleotide- and trinucleotide-repeat loci and some trinucleotide disease genes. Allele sizes in 33 of the human loci were significantly larger than those in the chimpanzee homologues; seven of the loci were significantly larger in chimpanzees compared with the humans (33:7 is significantly different from 1:1, $p < 0.0005$). Two loci were of similar size in both species and two did not amplify in chimpanzees. Similar significant trends in favour of larger human loci compared with their homologues in other primates were seen in gorillas, orang-utans, baboons and macaques. In marmosets, the number of loci amplified was too small to detect a significant difference. Similar significant trends were observed when the loci were confined to those which were 'neutral', by excluding trinucleotide-repeat disease loci and trinucleotide repeats associated with brain cDNAs from the microsatellite panel. For example, 21 neutral loci were longer in humans compared with five which were longer in chimpanzees ($p < 0.005$).

The greater length of human microsatellites can be accounted for by three possibilities. First, selection might favour longer repeats in humans or shorter repeats in primates. This explanation seems unlikely, since the trends were observed with diverse loci and remained even if the analyses were confined to neutral loci. Also, the very large number of independently segregating microsatellites means that the effect on any one locus must be small.

Second, if microsatellites are subject to a mutational bias in favour of expansion, then the greater length of human microsatellites could arise through a genome-wide increase in mutation rate in humans or a decrease in mutation rate in chimpanzees. A mutational bias in favour of expansions is supported by the following empirical observations. Weber & Wong (1993) considered the possibility of mutation bias when they observed an excess of expansion mutations in microsatellites typed in CEPH pedigrees. Unfortunately, only 22 out of the 62 mutations observed occurred in genomic DNA from untransformed cells. This left the possibility that the trends they observed were due to mutations occurring in transformed cells. Of the 22 mutations observed in genomic DNA, they detected 14 gains and eight losses in allele lengths ($p = 0.14$, exact binomial probability). However, we later supplemented Weber & Wong's (1993) data with microsatellite mutations observed by S. Sawcer and R. Feakes, who had been performing a linkage study on multiple sclerosis using genomic DNA from untransformed cells (Amos *et al*. 1996). This study yielded a further 15 mutations comprising seven gains, one loss and seven ambiguous mutations. Combined, the two data sets show a significant bias in favour of expansion: 21 gains and nine losses ($p = 0.021$, exact binomial probability).

The third possibility involves an observation bias associated with the fact that the microsatellites used in the human–non-human primate comparison were cloned

from human and were sufficiently polymorphic to be useful in linkage studies. Being selected for above average length, some argued that one would expect such markers to be longer than their homologues in related species and hence that the effects that we observed could be due entirely to an ascertainment bias (Ellegren *et al*. 1995, 1997).

This possibility was directly addressed by cloning a series of long polymorphic microsatellites from chimpanzees (Cooper *et al*. 1998). A total of 38 chimpanzee-derived loci were initially investigated, including 24 TG/CA repeats which we identified (maximum 23 repeats, minimum 12 repeats, mean, 17.3 repeats), one interrupted locus with 23 repeats and 13 chimpanzee loci described by Takenaka *et al*. (1993). Nine loci were excluded from further analysis as they did not amplify products of the expected size consistently and one locus did not amplify human DNA. The remaining 28 chimpanzee-derived loci were used for chimpanzee–human length comparisons. These chimpanzee loci had a similar mean repeat number compared with the human-derived loci, which were previously compared with their homologues in non-human primates above ($p = 0.13$).

Fourteen of these chimpanzee loci were significantly shorter than their human homologues, eight were significantly longer in chimpanzees and six loci were similar in length in the two species. If the 33:7 ratio of human-derived loci, which were significantly longer in humans compared with those longer in chimpanzees, was entirely due to an ascertainment bias, then one would expect a similar excess of loci longer in chimpanzees for chimpanzee-derived loci. However, 14:8 (chimpanzee-derived loci longer in humans) does not differ significantly from 33:7 (human-derived loci longer in humans) ($p = 0.18$, Fisher's exact test); while 14:8 does differ significantly from 7:33, the ratio expected if the effects we had observed with the human-derived loci were due entirely to ascertainment bias ($p = 0.0007$).

We also considered the possibility that our interpretation may have been confounded by an ascertainment bias resulting from the comparison of polymorphic loci in one species with monomorphic homologues in a related species. Monomorphic loci will tend to have lower mutation rates compared with polymorphic homologues. Thus, monomorphic loci will tend to be shorter than their polymorphic homologues in related species, if mutations are biased in favour of expansions. This source of bias was demonstrated by Crawford *et al*. (1998), when they analysed polymorphic cattle-derived microsatellites in sheep. When the homologues of these loci were polymorphic in sheep (308 loci), then 198 (64%) were longer in sheep. Conversely, if the analyses were confined to the 131 loci monomorphic in sheep, then 109 (83%) were longer in cattle.

We next confined our analyses to loci which were polymorphic in both species, or by further restricting such loci to dinucleotide repeats. In both scenarios, we observed similar trends showing no significant differences between the ratio of human loci longer than chimpanzee loci for human-derived versus chimpanzee-derived loci. In these scenarios we also observed highly significant differences between the ratios of human loci longer than chimpanzee loci for human-derived microsatellites, compared to the ratio expected from chimpanzee microsatellites, if the data from the human-derived loci were purely due to an ascertainment bias.

These data can be explained by a mutational bias in favour of expansions and a greater expansion rate in the human lineage compared with the chimpanzee lineage. A similar discrepancy in microsatellite allele lengths at homologous loci in related species was recently reported in reciprocal comparisons of sheep and cattle microsatellite loci (Crawford *et al*. 1998). Out of 20 polymorphic sheep-derived loci that were polymorphic in both, 16 were longer in sheep compared with their cattle homologues. However, 198 out of the 308 cattle-derived loci, which were polymorphic in both species, were significantly longer in sheep.

The molecular mechanisms underlying these observations are unclear, although the following non-mutually exclusive possibilities may help to explain why microsatellites expand faster in humans compared with chimpanzees.

(i) Human polymerases might be more error-prone compared with those in non-human primates.

(ii) Greater microsatellite mutation rates in humans might reflect the longer lag in humans between the onset of sexual maturity and reproduction. It appears that microsatellite mutation rates are higher in males than in females (Weber & Wong 1993; Amos *et al*. 1996; Primmer *et al*. 1998) probably reflecting larger number of cell divisions between zygote and sperm compared with between zygote and ovum. Later reproduction implies more cell divisions and hence, plausibly, greater mutation rates for microsatellites. Interestingly, HD disease alleles also show greater mutability in males (Duyao *et al*. 1993).

(iii) Microsatellite mutations may involve inter-chromosomal events. This possibility is supported by data on the trinucleotide repeat disease Machado–Joseph disease, where the mutation rate of the mutant chromosome depends partly in the haplotype of the normal chromosome (Takiyama *et al*. 1997). If inter-chromosomal processes are involved in the mutations of normal microsatellites, and if mutations are more likely in heterozygote individuals (Amos *et al*. 1996; Amos & Harwood 1998), then humans may show increased rates of microsatellite evolution because of their greater effective population size.

In conclusion, our data suggest that microsatellites show a mutational bias in favour of expansion mutations, a process reminiscent of the mutant alleles at most trinucleotide repeat disease loci. The rate of expansion of microsatellites in related species seems to differ, raising the possibility that these major deviations from molecular clock predictions may be a genome-wide phenomenon not specifically confined to microsatellite loci.

This discussion raises a number of speculative questions. First, are trinucleotide-repeat diseases found in man and not in other primates because human microsatellites appear to be generally longer than those in other primates? This question is difficult to address because the chance of ever finding a late-onset, rare ($< 1/10\,000$), neurodegenerative diseases in non-human primates is

remote. However, chimpanzees do appear to have shorter normal alleles than humans for several triplet diseases (Djian *et al*. 1996). Also, transgenic mice with expanded human trinculeotides from diseases genes exhibit relatively low mutation rates and smaller sized mutations than would be expected of the same alleles in humans, suggesting that the mutation process in humans may be unusual (reviewed in Rubinsztein & Amos 1998).

Second, how important are interchromosomal events in microsatellite mutations? The molecular mechanisms underlying microsatellite mutations are still unclear. If interchromsomal events are confirmed, then population size and structure may affect microsatellite mutations. This interesting (almost Lamarckian) concept has a possible precedent, which comes from studies of hybrid populations. Here, the meeting of dissimilar chromosomes usually causes an increase in heterozygosity. In hybrid zones, rare alleles not present in either population occur so frequently that they have been named 'hybrizymes' (Barton *et al*. 1983; Woodruff 1989). Limited DNA sequence data suggest that these are likely to be new mutations rather than the result of recombination events (Hoffman & Brown 1995). Such a pattern is certainly consistent with the notion that heterozygosity may act to modulate mutation rate.

Third, we should ask about the implications of these processes for human populations. If population mixing does act to accelerate expansion, we should find triplet repeat disease incidence to be highest and perhaps rising most rapidly in populations where historical events have brought together peoples with diverse origins. At the same time, over evolutionary time-scales, it is possible that we will see the emergence of new diseases associated with trinucleotide expansions which have yet to reach a disease threshold.

This work was funded by the Huntington's Disease Association, UK and the Leverhulme Trust. D.C.R. is a Glaxo Wellcome Research Fellow.

## REFERENCES

Amos, W. & Harwood, J. 1998 Factors affecting levels of genetic diversity in natural populations. *Phil. Trans. R. Soc. Lond.* B **353**, 177–186.

Amos, W., Sawcer, S. J., Feakes, R. & Rubinsztein, D. C. 1996 Microsatellites show mutational bias and heterozygote instability. *Nature Genet.* **13**, 390–391.

Barton, N. H., Halliday, R. B. & Hewitt, G. M. 1983 Rare electrophoretic variants in a hybrid zone. *Heredity* **50**, 139–146.

Bowcock, A. M., Ruiz Linares, A., Tomfohrde, J., Minch, E., Kidd, J. R. & Cavalli-Sforza, L. L. 1994 High resolution trees with polymorphic microsatellites. *Nature,* **368**, 455–457.

Bruford, M. W. & Wayne, R. K. 1993 Microsatellites and their application to population genetic studies. *Curr. Opin. Genet. Devel.* **3**, 939–943.

Cooper, G., Rubinsztein, D. C. & Amos, W. 1998 Ascertainment bias does not entirely account for human microsatellites being longer than their chimpanzee homologues. *Hum. Mol. Genet.* **7,** 1425–1429.

Crawford, A., Knappes, S. M., Paterson, K. A., deGotari, M. J., Dodds, K. G., Freking, R. T., Stone, R. T. & Beattie, C. W. 1998 Microsatellite evolution: testing the ascertainment bias hypothesis. *J. Mol. Evol.* **46**, 256–260.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M. & Slatkin, M. 1994 Mutational processes of simple sequence repeat loci in human populations. *Proc. Natl. Acad. Sci. USA* **91**, 3166–3170.

Djian, P., Hancock, J. M. & Chana, H. S. 1996 Codon repeats in genes associated with human diseases: fewer repeats in the genes of nonhuman primates and nucleotide substitutions concentrated at sites of reiteration. *Proc. Natl. Acad. Sci. USA* **93**, 417–421.

Duyao, M. (and 12 others) 1993 Trinucleotide repeat length instability and age of onset in Huntington's disease. *Nature Genet.* **4**, 387–392.

Ellegren, H., Primmer, C. R. & Sheldon, B. C. 1995 Microsatellite evolution: directionality or bias in locus selection. *Nature Genet.* **11**, 360–362.

Ellegren, H., Moore, S., Robinson, N., Byrne, K., Ward, W. & Sheldon, B. C. 1997 Microsatellite evolution—a reciprocal study of repeat lengths at homologous loci in cattle and sheep. *Mol. Biol. Evol.* **14**, 854–860.

Farrall, M. & Weeks, D. E. 1998 Mutational mechanisms for generating microsatellite allele frequency distributions: an analysis of 4,558 markers. *Am. J. Hum. Genet.* **62**, 1260–1262.

Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. 1995*a* An evaluation of genetic distances for use with microsatellite loci. *Genetics* **139**, 463–471.

Goldstein, D. B., Ruiz Linares, A., Cavalli-Sforza, L. L. & Feldman, M. W. 1995*b* Genetic absolute dating based on microsatellites and the origin of modern humans. *Proc. Natl. Acad. Sci. USA* **92**, 6723–6727.

Hoffman, S. M. G. & Brown, W. M. 1995 The molecular mechanism underlying the rare allele phenomenon in a subspecific hybrid zone of the California field-mouse, *Peromyscus californicus*. *J. Mol. Evol.* **41**, 1165–1169.

Jeffreys, A. J., Allen, M. J., Hagelberg, E. & Sonnberg, A. 1992 Identification of the skeletal remains of Josef Mengele by DNA analysis. *For. Sci. Int.* **56**, 65–76.

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

Kruglyak, S., Durrett, R. T., Schug, M. D. & Aquadro, C. F. 1998 Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. *Proc. Natl Acad. Sci. USA* **95**, 10 774–10 778.

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

Pollock, D. D., Bergman, A., Feldman, M. W. & Goldstein, D. B. 1998 Microsatellite behaviour with range constraints: parameter estimation and imporved distances for use in phylogenetic reconstruction. *Theor. Pop. Biol.* **53**, 256–271.

Primmer, C. R., Saino, N., Møller, A. P. & Ellegren, G. 1998 Unravelling the process of microsatellite evolution through analysis of germ line mutations in barn swallows *Hirundo rustica*. *Mol. Biol. Evol.* **15**, 1047–1054.

Rubinsztein, D. C. & Amos, W. 1998 Trinucleotide repeat mutation processes. In *Analysis of triplet repeat disorders* (ed. D. C. Rubinsztein & M. R. Hayden), pp. 257–268. Oxford: BIOS.

Rubinsztein, D. C., Amos, W., Leggo, J., Goodburn, S., Ramesar, R. S., Old, J., Bontrop, R., McMahon, R., Barton, D. E. & Ferguson-Smith, M. A. 1994 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.

Rubinsztein, D. C., Amos, W., Leggo, J., Goodburn, S., Jain, S., Li, S. H., Margolis, R. L., Ross, C. A. & Ferguson-Smith, M.

1995 Microsatellites are generally longer in humans compared to their homologues in non-human primates: evidence for directional evolution at microsatellite loci. *Nature Genet.* **10**, 337–343.

Takenaka, O., Takasaki, H., Kawamoto, S., Arakawa, M. & Takenaka, A. 1993 Polymorphic microsatellite DNA amplification customised for chimpanzee paternity testing. *Primates* **34**, 27–35.

Takiyama, Y. (and 10 others) 1997 Single sperm analysis of the CAG repeats in the gene for Machado–Joseph disease (MJD1): evidence for non-Mendelian transmission of the MJD1 gene and for the effect of the intragenic CGG/GGG polymorphism on the intergenerational instability. *Hum. Mol. Genet.* **7**, 525–530.

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

Weber, J. L. & Wong, C. 1993 Mutation of human short tandem repeats. *Hum. Mol. Genet.* **2**, 1123–1128.

Weissenbach, J., Gyapay, G., Dib, C., Vignal, A., Morissette, J., Millasseau, P., Vaysseix, G. & Lathrop, M. 1992 A second-generation linkage map of the human genome. *Nature* **359**, 794–801.

Woodruff, R. C. 1989 Genetic anomalies associated with *Cerion* hybrid zones: the origin and maintenance of new electrophoretic variants called hybrizymes. *Biol. J. Linn. Soc.* **36**, 281–294.