Heterogeneous nature and distribution of interruptions in dinucleotides may indicate the existence of biased substitutions underlying microsatellite evolution[*]

Miguel A. Varela, Roberto Sanmiguel, Ana González-Tizón, Andrés Martínez-Lage[†]

Departamento de Biología Celular y Molecular, Facultad de Ciencias, Universidade da Coruña. 15701 La Coruña, Spain

How to cite:

> Heterogeneous nature and distribution of interruptions in dinucleotides may indicate the existence of biased substitutions underlying microsatellite evolution Miguel A. Varela, Roberto Sanmiguel, Ana Gonzalez-Tizón, Andrés Martínez-Lage, *J Mol Evol.*, (2008), 66(6): 575–580. DOI 10.1007/s00239-008-9107-3.

Abstract

Some aspects of microsatellite evolution, such as the role of base substitutions, are far from being fully understood. To examine the significance of base substitutions underlying the evolution of microsatellites we explored the nature and the distribution of interruptions in dinucleotide repeats from the human genome. The frequencies that we inferred in the repetitive sequences were statistically different from the frequencies observed in other noncoding sequences. Additionally, we detected that the interruptions tended to be towards the ends of the microsatellites and 5′-3′ asymmetry. In all the estimates nucleotides forming the same repetitive motif seem to be affected by different base substitution rates in AC and AG. This tendency itself could generate patterning and similarity in flanking sequences and reconcile these phenomena with the high mutation rate found in flanking sequences without invoking convergent evolution. Nevertheless, our data suggest that there is a regional bias in the substitution pattern of microsatellites. The accumulation of random substitutions alone cannot explain the heterogeneity and the asymmetry of interruptions found in this study or the relative frequency of different compound microsatellites in the human genome. Therefore, we cannot rule out the possibility of a mutational bias leading to convergent or parallel evolution in flanking sequences.

Introduction

The nature and rate of base mutations are influenced by local sequence context (Goodman and Fygenson 1998; Marra and Schar 1999; Zhang and Gerstein 2003; Arndt et al. 2005), which is simple in form but displays a complex mutational behavior in microsatellites. Microsatellites are tandemly repeated DNA sequences consisting of 1–6 bp that typically have a high level of instability. Due to their elevated degree of polymorphism, microsatellite markers are widely used for mapping (Kong et al. 2002) and for studies of population and evolutionary genetics (Goldstein and Schlötterer 1999). In addition, the malign expansion of some of these repetitive sequences is involved in a number of human neurodegenerative diseases (Gatchel and Zoghbi 2005). Despite the effort to remove uncertainty around the complex mutational dynamics of microsatellites, some mutational processes occurring at microsatellite loci remain unclear. Replication slippage seems to play a protagonist role in microsatellite variability, and most mutations consist of an increase or decrease of only one repeat unit (Weber and Wong 1993; Ellegren 2000). In the human genome, the estimate of slippage mutation rate is about 10^{-3} per locus per gamete per generation (Weber and Wong 1993; Dib et al. 1996), which is several orders of magnitude higher than the point mutation rate, 10^{-8} per nucleotide per gamete per generation (Li 1997). However, base substitutions are the main source of new two-repeat loci (Zhu et al. 2000) and also essential in microsatellite degeneration, which is assumed to occur by acquiring multiple base substitutions that erode the tandem arrangement of the microsatellite sequence (Kruglyak et al. 1998; Taylor et al. 1999).

Of particular interest and complexity is the influence of base substitutions in microsatellite evolution, in particular, on the variability of sequences flanking microsatellites. Previous studies have described both high mutation rate (Stallings 1995) and similarity in microsatellite flanking sequences (Schlötterer et al. 1991). Vowles and Amos (2004) analyzed the flanking sequences of AC tracts and found a two-base periodicity and that sequences immediately flanking a microsatellite $(AC)_n$ of a given length show greater similarity to each other than to random sequences. Those authors reconciled similarity and patterning with elevated mutation rates by invoking convergence evolution determined by the influence of microsatellites on base substitutions in flanking sequences. On the other hand, Webster and Hagberg (2007) reproduced these phenomena using a simulation with random substitutions in sequences that previously formed part of the microsatellite, and they concluded that similarity and patterning could be explained without arguing that microsatellites promote biased substitutions in flanking sequences. In this context, we believe that a more detailed knowledge of the distribution of base substitutions along repetitive arrays is needed to more accurately describe the evolution of microsatellites and their flanking sequences.

The purpose of this study is to explore the sequence of repetitive arrays from the human genome in order to search for patterns in the distribution of mutations and a change in the spectrum of base substitutions. The slow rate of accumulation of base substitutions makes the inference of patterns difficult in these sequences, especially when using few loci. Therefore, we resorted to an indirect approach and analyzed the distribution of interruptions in dinucleotide repeats in the whole human genome. In this paper we show that although the erosion of the tandem arrangement could account for the periodicity and similarity observed by Vowles and Amos (2004) without invoking convergent evolution, some base substitutions seem to be biased in these repetitive sequences in comparison with other noncoding sequences.

Methods

Human chromosomes were downloaded from the National Center for Biotechnology Information (NCBI) database (v36 03/06; GenBank accession numbers NC_000001−NC_000024). Then we used the program Tandem Repeats Finder (TRF) 4.0 to search for microsatellites in each chromosome using the following parameters: match weight, 2; mismatch weight, −3; indel weight, −5; and minimum score parameter, 20. Because mismatches and indels give a penalty to the alignment score, TRF did not report any mismatch or indel at the two nucleotides situated at the beginning and the end of the microsatellite with the parameters

used in this study. In this approach, only microsatellites containing one interruption of the form $(AC)_nN(AC)_n$ were considered. Therefore, all interruptions are in a sequence context that only varies in the relative position of the interruption in the repetitive array to avoid the bias in the distribution of interruptions that the alignment algorithm would introduce if microsatellites with more than one interruption had been considered. This would entail detecting regions containing several substitutions in a row in the middle of the repetitive array but not close to the end of the microsatellite as a part of a microsatellite. Additionally, we developed a new program, Repeat Array Analyzer (RAA), in JAVA, available at http://www.udc.es/dep/bm/genetica/gibe/raa.htm. In this work we used RAA to process the data files obtained by TRF and count interruptions along repetitive arrays of a given size and repetitive motif. To analyze the distribution of interruptions, we combined the results of all chromosomes and only considered microsatellites with the same repetitive frame, i.e., $(AC)_n$ only consisted of microsatellites starting with AC at their 5′ end, and not CA, TG, or GT. Moreover, each size class was exclusive, i.e., a repetitive array containing 19 bp was only analyzed together with arrays of the same size without redundancy in other classes.

We tested whether there was a significant preference of interruptions to be at particular sites of the array, and whether substitution frequencies were statistically different by χ^2 tests using the program SPSS version 14 (SPSS Inc., Chicago, IL). The tendency of interruptions to occur toward the ends and in 5′-3′ asymmetry was also analyzed by χ^2 tests. To analyze the tendency of interruptions to occur toward ends, we compared the number of interruptions in the first and last scorable repeat units (in boldface), corresponding to sites +3 and +4 from the 5′ end of the array and sites −3 and −4 sites from 3′, against the number of interruptions found in central repeats, $(A^{+1}C^{+2})A^{+3}\ C^{+4}\ ACACACAC^{-4}\ A^{-3}\ (C^{-2}A^{-1})$. To test the 5′-3′ asymmetry in the distribution of interruptions, we compared the number of interruptions at the scorable nucleotide closest to the 5′ end against the scorable nucleotide closest to the 3′ end, interruptions involving the substitution of A, $(A^{+1}C^{+2})$ $A^{+3}\ C^{+4}\ldots C^{-4}\ A^{-3}\ (C^{-2}A^{-1})$, and idem for C, $(A^{+1}C^{+2})A^{+3}\ C^{\ +4}\ A\ldots AC^{-4}\ A^{-3}(C^{-2}A^{-1})$ (compared sites in boldface). Both the tendency of interruptions to occur toward ends and the 5′-3′ asymmetry were studied in microsatellites containing odd numbers of nucleotides to compare only symmetrical sites of the repetitive array.

Results

In this work, we explored the sequence of repetitive arrays from the human genome, searching for patterns in the distribution of interruptions. Table 1 reports the observed interruptions in dinucleotide repetitive sequences. If we assume that each interruption comes from a single base substitution and no back mutations, Table 1 infers a preference for base substitutions on some nucleotides of the repetitive units. For example, in $(AC)_{10}$, mutations involving the substitution of C should be about 3 times more frequent than those involving substation of A (Table 1). Then we investigated whether these preferences could be extrapolated to other dinucleotide repeats and confirmed that these results were significant by χ^2 tests in both $(AG)_{10}$ and $(AC)_{10}$ (Table 1). Afterward, we found that the frequencies that we inferred in the repetitive sequences were statistically different (χ^2 tests, $p < 0.001$) from the frequencies observed in nongenic, nonrepetitive sequence alignments of human and chimpanzee (Smith et al. 2002) and in human pseudogenes taking into account the immediately neighboring nucleotide of each substitution (Zhang and Gerstein 2003). Table 1 shows the expected values according to the frequency of substitutions calculated by Smith et al. (2002) and Zhang and Gerstein (2003). It can be observed that some of our estimates are overrepresented (A→T and T→A in AT dinucleotides, C→G and C→A in AC, and G→C in AG).

Table 1. Observed interruptions (**O**) at each site of dinucleotide repeat units and expected values according to Smith (2002) (**E₁**) and Zhang and Gerstein (2003) (**E₂**)

(AT)$_{10}$		O	E$_1$	E$_2$	(AC)$_{10}$		O	E$_1$	E$_2$
A	C	10	12.71	29.36	A	C	7	10.73	30.28
	G	62	90.09	85.45		G	24	66.12	40.97
	T	40[*]	12.71	16.03		T	12	10.73	21.37
	Total	112				Total	43[a]		
T	A	37[*]	12.71	13.21	C	G	32[*]	10.73	15.54
	C	71	90.09	71.54		T	72	85.98	71.31
	G	11	12.71	15.41		A	48[*]	10.73	15.54
	Total	119				Total	152[a]		
			118.71[b]	99.12[b]				202.31[b]	114.27[b]
			<0.001	<0.001				<0.001	<0.001

(AG)$_{10}$		O	E$_1$	E$_2$	(GC)$_{10}$		O	E$_1$	E$_2$
A	C	4	6.77	7.80	G	C	2	–	–
	G	15	41.71	25.99		A	7	–	–
	T	11	6.77	8.67		T	1	–	–
	Total	30[a]				Total	10		
G	C	29[*]	6.77	12.62	C	G	0	–	–
	A	58	54.23	58.22		T	6	–	–
	T	6	6.77	9.70		A	1	–	–
	Total	93[a]				Total	7		
			94.31[b]	29.80[b]				–	–
			<0.001	<0.001				–	–

Note. **O**, observed number of interruptions; **E₁**, expected number of interruptions considering a transition bias of 3.6 and a bias in mutations from G:C→A:T of 1.3 (Smith et al. 2002); **E₂**, expected number of interruptions taking into account the immediately neighboring nucleotide of each substitution (Zhang and Gerstein 2003)

[a] Statistical evidence for a difference in the number of interruptions at different sites of the repetitive unit, $p < 0.001$

[b] Statistical evidence for a difference between observed and expected values

[*] Interruptions overrepresented in comparison with both expected values that have a percentage deviation of more than 200%. (–) Not enough data to give estimates

We also found strong statistical evidence to support the tendency of interruptions to be toward the ends of the array and 5′-3′ asymmetry (Fig. 1). In Fig. 1A length dependency in the distribution of interruptions is observed. In arrays of 19 bp there are more interruptions at the center of the array at C sites and at the 3′ end at A sites. Among the sizes examined (19, 23, 27, 31, and 35), the preference for interruptions to be toward the ends was significant in the three longest classes (Fig. 1C–E), whereas at A sites the tendency of the interruptions was more frequent at the 3′ end and was significant for all size classes except for microsatellites of 23 bp (Fig. 1 A, C–E) (χ^2 tests, $p < 0.008$ after sequential Bonferroni correction). To assure that these tendencies were not due to a bias of the TRF algorithm to identify tandem repeats with interruptions close to the ends when repeat arrays are flanked by arrays of different motifs, we analyzed the flanking sequences of microsatellites of the form (AC)$_{16}$NCA (Fig. 1E). Of the 23 microsatellites we found 5 that were flanked by other arrays, but only three of the arrays were at the 3′ flanking side, and none of them could appear as an interruption close to the end of the core microsatellite. Subsequently, we carried out the same analysis using reverse sequences and the result was the mirror image of Fig. 1.
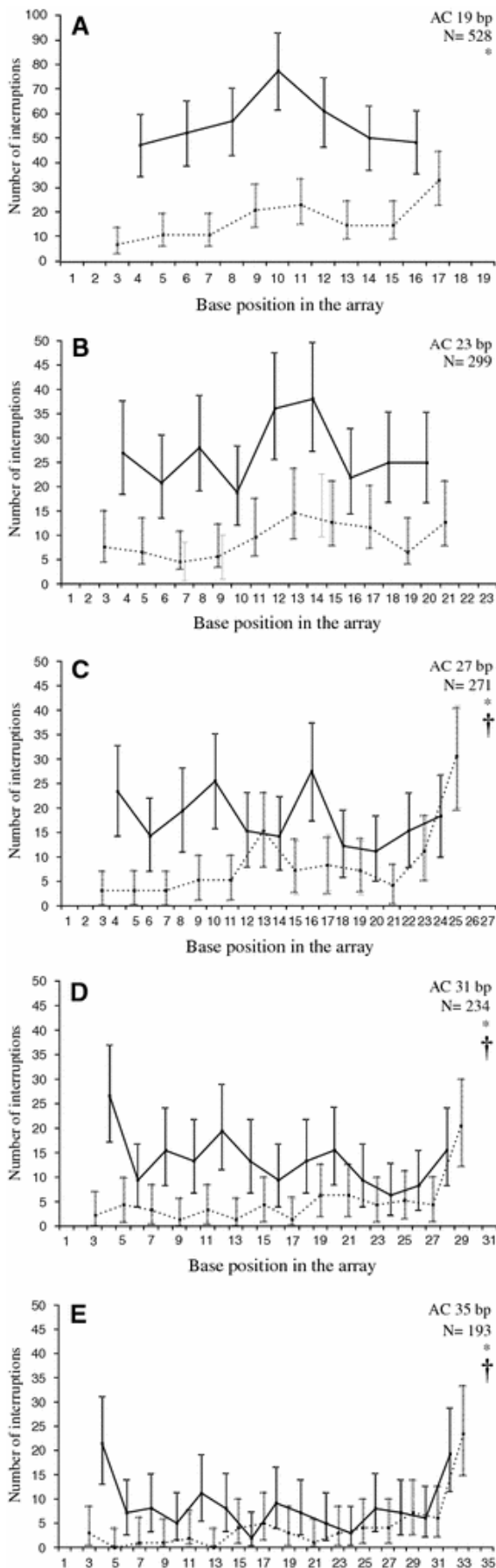
Figure 1. Distribution of interruptions along AC repetitive arrays (5′ -3′) of 19 bp (**A**), 23 bp (**B**), 27 bp (**C**), 31 bp (**D**), and 35 bp (**E**). Dotted lines indicate the number of interruptions at A sites. Solid lines indicate the number of interruptions at C sites. N, number of microsatellites considered. *Significant asymmetry of interruptions at A sites toward the 3′ end ($p < 0.001$). †Significant tendency of interruptions to be toward the ends ($p < 0.001$). Error bars represent 95% confidence intervals (binomial distributions)

Discussion

In this work, we analyzed the distribution of interruptions at dinucleotide repeats and estimated substitution rates that are significantly different from the results obtained in other noncoding sequences by Smith et al. (2002) and Zhang and Gerstein (2003). Our data suggest that microsatellites have substitution patterns that are not equivalent to those observed in other sequences. These patterns cannot be explained by the peculiar repetitive nature of these sequences taking into account the influence of the immediately neighboring nucleotide of each substitution as in Zhang and Gerstein (2003). These biases may intervene in the birth and early expansion of the microsatellite or be promoted by the microsatellite itself.

Base substitutions tended to occur toward the ends of the array, as reported previously in a study based on 22 bovine and ovine orthologous loci (Brohede and Ellegren 1999). On the other hand, 5′-3′ asymmetry of mutations was described in minisatellites (Armour et al. 1993; Jeffreys et al. 1994) and in triplet repeats involved in human neurodegenerative diseases (Eichler et al. 1995). Brohede and Ellegren (1999) proposed three mechanisms that could account for high mutation rates at the ends of the array, all involving inefficiency of mismatch repair systems, after replication slippage, associated with homologous recombination in microsatellite heterozygous regions, and after DNA damage. Additionally, we believe that 5′-3′ asymmetry in these mechanisms might be involved in the higher frequency of interruptions at A sites toward the 3′ end.

Here we formulate another hypothesis that could also introduce a bias for the interruptions to be toward the ends of long repetitive arrays. Microsatellite evolution is a process in which arrays suffer successive expansions and contractions at the same time they acquire base substitutions. These base substitutions stabilize the region were they arise due to creation of mismatches that reduce the slippage rate (Weber and Wong 1993; Jin et al. 1996). With enough time, this region tends to acquire more interruptions and degrade to a nonrepetitive sequence, whereas the unaffected region continues to expand normally, attracting the center point of the array. Thus, the dynamic mutational process of microsatellites could introduce a bias for the interruptions to appear in the ends of the microsatellite. Nevertheless, this explanation alone cannot satisfactorily explain the asymmetry that we found in some interruptions. In order to explain the asymmetry of interruptions without assuming the asymmetry of substitutions, it would be necessary that some ending sequences promoted microsatellite expansion more than others, and this would be the reason why these endings are overrepresented in long microsatellites. On the other hand, the assumption that an interruption is originated from a single mutation, and repetitive arrays without interruptions were not affected by base substitutions, could be violated by back mutations or the duplication or purification of interruptions by slippage (Harr et al. 2000). Consequently, these processes cannot be excluded, given that they can contribute to the heterogeneous distribution of interruptions.

The Origin of Patterning and Similarity in Flanking Sequences

Vowles and Amos (2004) found a remarkable level of similarity between different loci comprising 25–50 bases at either side of $(AC)_n$ and a two-base patterning, with bases being under- and overrepresented at adjacent sites. This patterning was significant even around just $(AC)_2$, reaching its maximal point around $(AC)_9$ and being broader but less pronounced on longer microsatellites. To explain similarity and patterning, those authors suggested that microsatellite formation could spark off mutational mechanisms in flanking sequences that lead to convergent evolution. Vowles and Amos (2004) proposed another two alternative explanations that could also account for the patterning and similarity of the flanking sequences. The first alternative hypothesis was that patterning could have arisen previously due to an unknown force and then microsatellites form or expand more efficiently in this context, where some base mutations are more common than others. The second hypothesis was the microsatellite erosion model, by which the patterning

arises by the occurrence of base substitutions in sequences that previously formed part of the microsatellite, and therefore these flanking sequences represent the remnants of longer microsatellites now partially degenerated by base substitutions. Webster and Hagberg (2007) supported this erosion model and showed using a simulation that similarity and patterning in flanking sequences could be explained by the accumulation of random substitutions in microsatellites using the rates calculated by Smith et al. (2002) in nonrepetitive sequences. Our data show that in AC and AG there is a preference for base substitutions on some nucleotides of the repetitive units that is significant by χ^2 tests in both $(AG)_{10}$ and $(AC)_{10}$. As a result, it also supports that the patterning of under- and overrepresentation of adjacent bases of flanking sequences in AC arrays could be generated by the partial erosion of the tandem arrangement of microsatellites.

With the current knowledge of the mutational processes, the microsatellite erosion model is the most parsimonious hypothesis. Another argument supporting the erosion model is that the influence of the microsatellites enhancing mutations at flanking sequences seems to be restricted to the bases closest to the repeat region. Brohede and Ellegren (1999) detected more substitutions within the five nucleotides immediately adjacent to the repeat region than in flanking sequences farther away. Similarly, Vowles and Amos (2004) found a significant overrepresentation of mutation rates in a region comprising −9 bases 5′ and +4 bases 3′ around long microsatellites, whereas around short microsatellites the mutation rate remained constant. However, they observed sequence similarity in a broader sequence comprising 25–50 bases at either side of AC tracts, and it was significant 10 bases around just $(AC)_2$. This model gains support if we take into account the long persistence of microsatellites predicted by theoretical studies of microsatellite evolution (Tachida and Iizuka 1992; Stephan and Kim 1998).

Microsatellites are often flanked by sequences called regions of cryptic simplicity that are made up of intermixed motives. Many of these sequences of cryptic simplicity originated by the erosion of the microsatellite are not detected as microsatellites by a search for pure tracts. However, in some cases they could still maintain clear signs of tandem arrangement, e.g., the AC tandem arrangement is degenerated preferentially by mutations on C (Fig. 1). This partial tandem arrangement could be responsible for the patterning of flanking sequences. Nevertheless, our study suggests that a regional bias in the substitution pattern of microsatellites could be affecting the evolution and the variability found in microsatellites and their flanking sequences.

Conclusion

This investigation examined the nature and the distribution of interruptions in numerous microsatellites from the human genome, and explored the origins of the patterning in flanking sequences (flanking sequences being the remnants of repetitive sequences partially eroded by base substitutions or convergent evolution of flanking sequences promoted by the influence of the microsatellite). On the basis of this examination we conclude that mutations in microsatellite DNA seem to have biases such as a tendency to occur toward the ends of the repetitive array, 5′-3′ asymmetry, and an unusual spectrum of base pair substitutions. Although the partial degeneration of the tandem arrangement could be responsible for the patterning found in sequences flanking microsatellites, our data suggest that there is a regional bias in the substitution pattern of microsatellites. The accumulation of random substitutions alone as simulated by Webster and Hagberg (2007) cannot explain the heterogeneity and the asymmetry of interruptions found in this study or the relative frequency of different compound microsatellites in the human genome (Bull et al. 1999). On this basis, we cannot exclude the existence of a mutational bias leading to convergent or parallel evolution in flanking sequences.

References

Armour JA, Harris PC, Jeffreys AJ (1993) Allelic diversity at minisatellite MS205 (D16S309): evidence for polarized variability. Hum Mol Genet 2:1137–1145

Arndt PF, Hwa T, Petrov DA (2005) Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. J Mol Evol 60:748–763

Brohede J, Ellegren H (1999) Microsatellite evolution: asymmetry of substitutions within repeats and neutrality of flanking sequences. Proc R Soc Lond B 266:825–833

Bull LN, Pabon-Pena CR, Freimer NB (1999) Compound microsatellite repeats: practical and theoretical features. Genome Res 9:830–838

Bull LN, Pabon-Pena CR, Freimer NB (1999) Compound microsatellite repeats: practical and theoretical features. Genome Res 9:830–838

Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. Nature 380:152–154

Eichler EE, Kunst CB, Lugenbeel KA, Ryder OA, Davison D, Warren ST (1995) Evolution of the cryptic FMR1 CGG repeat. Nat Genet 11:301–307

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

Gatchel JR, Zoghbi HY (2005) Diseases of unstable repeat expansion: mechanisms and common principles. Nat Genet 6:743–755

Goldstein DB, Schlötterer C (1998) Microsatellites: evolution and applications. Oxford University Press, New York

Goodman MF, Fygenson KD (1998) DNA polymerase fidelity: from genetics toward a biochemical understanding. Genetics 148:1475–1482

Harr B, Zangerl B, Schlotterer C (2000) Removal of microsatellite interruptions by DNA replication slippage: phylogenetic evidence from *Drosophila*. Mol Biol Evol 17:1001–1009

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

Jin L, Macaubas C, Hallmayer J, Kimura A, Mignot E (1996) Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. Proc Natl Acad Sci USA 93:15285–15288

Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. Nat Genet 31:241–247

Kruglyak S, Durrett RT, Schug MD, Aquadro CF (1998) Equilibrium distributions of microsatellite repeat length resulting from a balance between slippage events and point mutations. Proc Natl Acad Sci USA 95:10774–10778

Li WH (1997) Molecular evolution. Sinauer Associates, Sunderland, MA

Marra G, Shar P (1999) Recognition of DNA alterations by the mismatch repair system. Biochem J 338:1–13

Schlötterer C, Amos B, Tautz D (1991) Conservation of polymorphic simple sequence loci in cetacean species. Nature 354:63–65

Smith NG, Webster MT, Ellegren H (2002) Deterministic mutation rate variation in the human genome. Genome Res 12:1350–1356

Stallings RL (1995) Conservation and evolution of (Ct)(N)/(Ga)(N) microsatellite sequences at orthologous positions in diverse mammalian genomes. Genomics 25:107–113

Stephan W, Kim Y (1998) Persistence of microsatellite arrays in finite populations. Mol Biol Evol 15:1332–1336

Tachida H, Iizuka M (1992) Persistence of repeated sequences that evolve by replication slippage. Genetics 131:471–478

Taylor JS, Durkin JM, Breden F (1999) The death of a microsatellite: a phylogenetic perspective on microsatellite interruptions. Mol Biol Evol 16:567–572

Vowles EJ, Amos W (2004) Evidence for widespread convergent evolution around human microsatellites. PLoS Biol. 2:E199

Weber JL, Wong C (1993) Mutation of human short tandem repeats. Hum Mol Genet 2:1123–1128

Webster MT, Hagberg J (2007) Is there evidence for convergent evolution around human microsatellites? Mol Biol Evol 24:1097–1100

Zhang Z, Gerstein M (2003) Patterns of nucleotide substitution, insertion and deletion in the human genome inferred from pseudogenes. Nucleic Acids Res 31:5338–5348

Zhu Y, Strassmann JE, Queller DC (2000) Insertions, substitutions, and the origin of microsatellites. Genet Res 76:227–236

[†] andres@udc.es